A GLOBAL REPOSITORY FOR PLANET-SIZED EXPERIMENTS AND OBSERVATIONS

BY DEAN N. WILLIAMS, V. BALAJI, LUCA CINQUINI, SÉBASTIEN DENVIL, DANIEL DUFFY, BEN EVANS, ROBERT FERRARO, ROSE HANSEN, MICHAEL LAUTENSCHLAGER, AND CLAIRE TRENHAM

The Earth System Grid Federation (ESGF) represents a multinational effort to securely access, monitor, catalog, transport, and distribute petabytes of data for climate change research experiments and observations.

Climate scientists have taken advantage of the revolution in high-capacity, high-performance computing occurring over the past several decades by creating more accurate and detailed simulations (IPCC 2013). Today, coupled atmosphere–ocean general circulation models running virtual Earth laboratory experiments operate at much higher-resolution [25-km (¼°) grid spacing or even finer], compute more variables, and cover longer time spans than ever before (Fig. 1). With these and other tools at their fingertips, climate scientists are now able to better understand not only what tomorrow's climate might be like, but how today's climate came to be the way it is.

As model complexity has increased, so has the quantity of the data produced; model output has grown in the past 15 years from megabytes (millions of bytes) to terabytes (trillions of bytes) to petabytes (quadrillions of bytes) of information. The World Climate Research Programme (WCRP; see the appendix for a glossary of terms), the leading international body for coordinating and facilitating climate research, prescribes an assortment of repeated tests and trials at regular intervals to assess the state of climate models, to understand their behavior, and to project future climate change. Running even a single WCRP experiment on the world's fastest and most powerful supercomputers using today's advanced climate models can take many months and produce more data than any one researcher or research team can study in a reasonable time frame.

Further, to achieve breakthroughs, all of the data must be exploitable: scientists must be able to analyze

AFFILIATIONS: WILLIAMS AND HANSEN—Lawrence Livermore National Laboratory, Livermore, California; BALAJI—Princeton University, Princeton, New Jersey; CINQUINI AND FERRARO—Jet Propulsion Laboratory, Pasadena, California; DENVIL—Institut Pierre-Simon Laplace, Paris, France; DUFFY—Goddard Space Flight Center, Greenbelt, Maryland; EVANS AND TRENHAM—National Computational Infrastructure, Australian National University, Acton, Australian Capital Territory, Australia; LAUTENSCHLAGER—German Climate Computing Centre, Hamburg, Germany
CORRESPONDING AUTHOR: Dean N. Williams, Lawrence Livermore National Laboratory, Mail Stop L-103, 7000 East Ave., Livermore, CA 94550
E-mail: williams13@llnl.gov

The abstract for this article can be found in this issue, following the table of contents.
DOI:10.1175/BAMS-D-15-00132.1

In final form 29 July 2015

the data effectively and efficiently at multiple scales and easily communicate the results to different communities, such as other scientists, policy makers, and potential downstream consumers, such as farmers and teachers. The explosion in data complexity and scale makes these tasks difficult to achieve, particularly given that many climate experiments integrate different types of data, stored in a variety of formats. These include model output; observational data from National Aeronautics and Space Administration and European Space Agency satellites and instruments as well as in situ ground-based data such as surface and air temperatures; and reanalysis data, which include a mix of observational and model data. Because of the size and complexity of the datasets, the information technology infrastructure that enables the sharing of data has become a necessary and critical tool for the climate modeling community. Research groups tend to exchange models and model output, thereby enabling and encouraging different scientists to reanalyze results or even repeat experiments with different variables selected. The democratization of the data through open access and frequent reexamination of climate prediction helps to ensure both transparency and the scientific rigor of climate research.

The relentless growth in data and associated issues it produces has underscored the importance of developing a standardized and effective system for climate data management—an effort that requires an ongoing partnership between climate and data scientists. Climate scientists, oceanographers, meteorologists, geologists, ecologists, and biologists, the more visible faces of climate studies, are tasked with assessing the validity of climate predictions by identifying climate trends, the nature and duration of these trends, and what is causing them. They also analyze the potential effectiveness of proposed methods for mitigating or adapting to climate changes. Aided by modern computing tools, computational and data scientists assist climate scientists with organizing, manipulating, visualizing, processing, and transforming the data. Working together, these two groups are helping to accelerate the pace of climate science and circulate the best available information about the causes and effects of a changing climate.

EVOLUTION OF GLOBAL CLIMATE DATA MANAGEMENT. Understanding the behavior of climate models and gathering and sharing climate data are key efforts of the Coupled Model

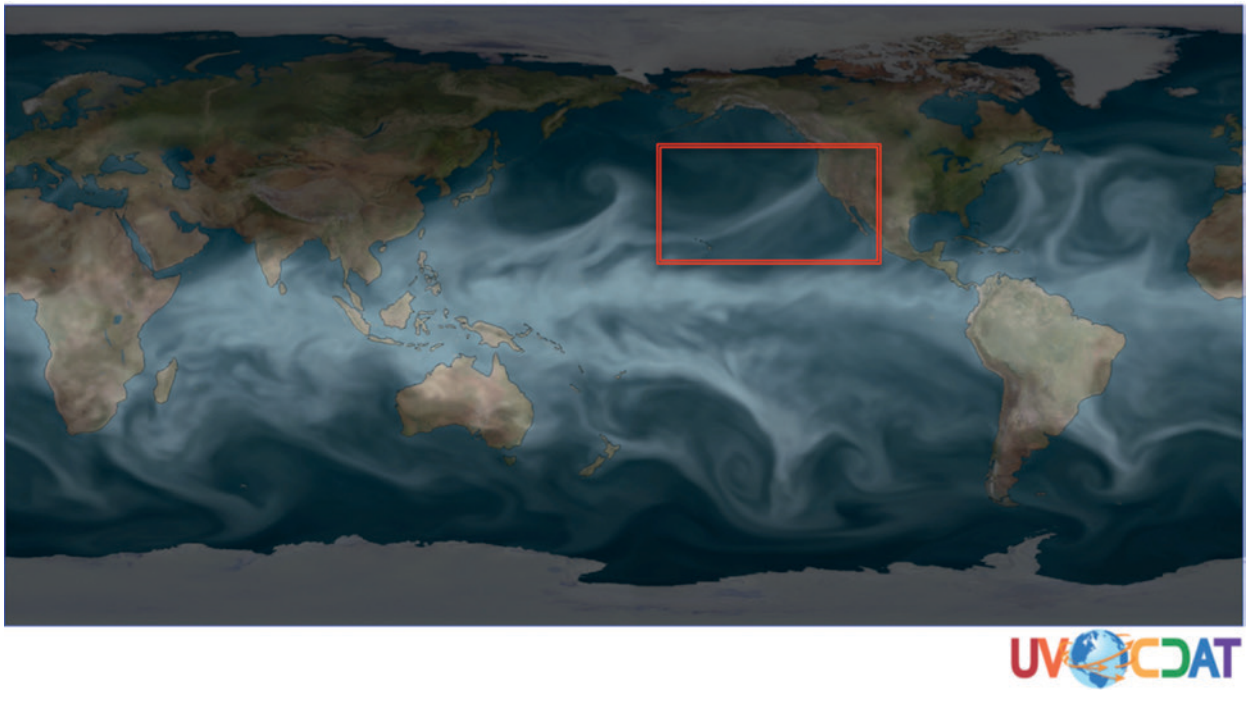

Fig. 1. Understanding today's climate requires increasingly higher spatial and temporal resolution. This visualization depicts atmospheric precipitable water (white) from the National Aeronautics and Space Administration (NASA) Modern Era Reanalysis for Research and Application (MERRA) model using 55-km grid spacing. Data were produced hourly for 35 yr. At this time resolution, short-term atmospheric phenomena can be seen in detail, such as the atmospheric river known as the "pineapple express" that brought rainfall to CA on 11 Dec 2014.

Intercomparison Project (CMIP), the worldwide standard experimental protocol for studying general circulation model output. Established in 1995 by the WCRP's Working Group on Coupled Modeling (WGCM), CMIP provides a community-based support structure for climate model diagnosis, evaluation, validation, comparison, enhancement, documentation, and data access (Meehl et al. 2014). Virtually the entire international climate modeling community has participated in this project since its inception.

Climate data management has come a long way in the past two decades, due in large part to CMIP. The project has both exposed the need for better data coordination and provided a timely opportunity for collaborative development. Initially, sharing massive datasets was difficult. To gather the data used for the third CMIP (CMIP3) in 2003, scientists shipped large data "bricks" around the globe. Climate researchers loaded their data and sent the bricks back to Lawrence Livermore National Laboratory (LLNL), where the entire 35 terabytes were then stored in a single centralized location.

The data were made accessible via a web portal called the Earth System Grid (ESG), but the process had limitations. Researchers faced a delay of weeks between data generation and its availability on ESG, and the lack of unified data protocols hindered the merger of output data from different modeling groups. Data errors were very difficult to correct, and file backups were onerous. Even worse, LLNL served as a single point of failure in the event of a system crash. Later the archive was replicated at the IPCC Data Distribution Centre at the German Climate Computing Centre (Deutsches Klimarechenzentrum; DKRZ) in Hamburg for long-term preservation and access. An updated version of the CMIP3 data archive (i.e., data updates after the release of IPCC's Fourth Assessment Report) can also be accessed at LLNL using the older ESG system (ftp://ftp-esg.ucllnl.org; to register for access, see https://esg.llnl.gov:8443/index.jsp).

In 2006, the Working Group on Coupled Modeling (WGCM) agreed to promote a new set of coordinated climate model experiments to explore, for instance, the ability of models to predict climate on decadal time scales. This, the fifth phase of CMIP (CMIP5), generated 50 times more data than CMIP3. To increase the accessibility and usefulness of the large volume of CMIP5 climate data, the international Earth System Grid Federation (ESGF) was formed (Cinquini et al. 2014). Collaborating partners in the

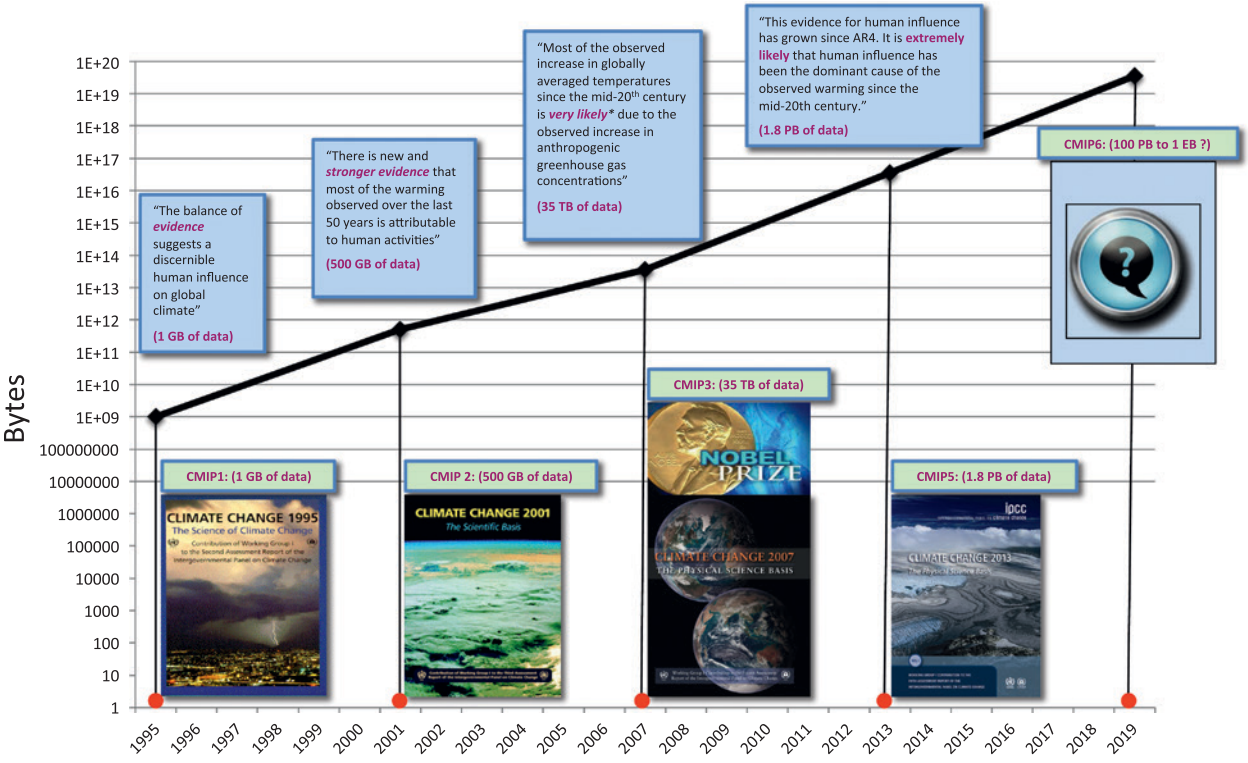

Fig. 2. ESG, now the ESGF, has supported data gathering, analysis, and dissemination for the past two phases of CMIP. CMIP simulation model runs are key components of periodic assessments by the IPCC. The IPCC—and the scientists whose research supported the IPCC, including those from the ESGF—shared the 2007 Nobel Peace Prize.

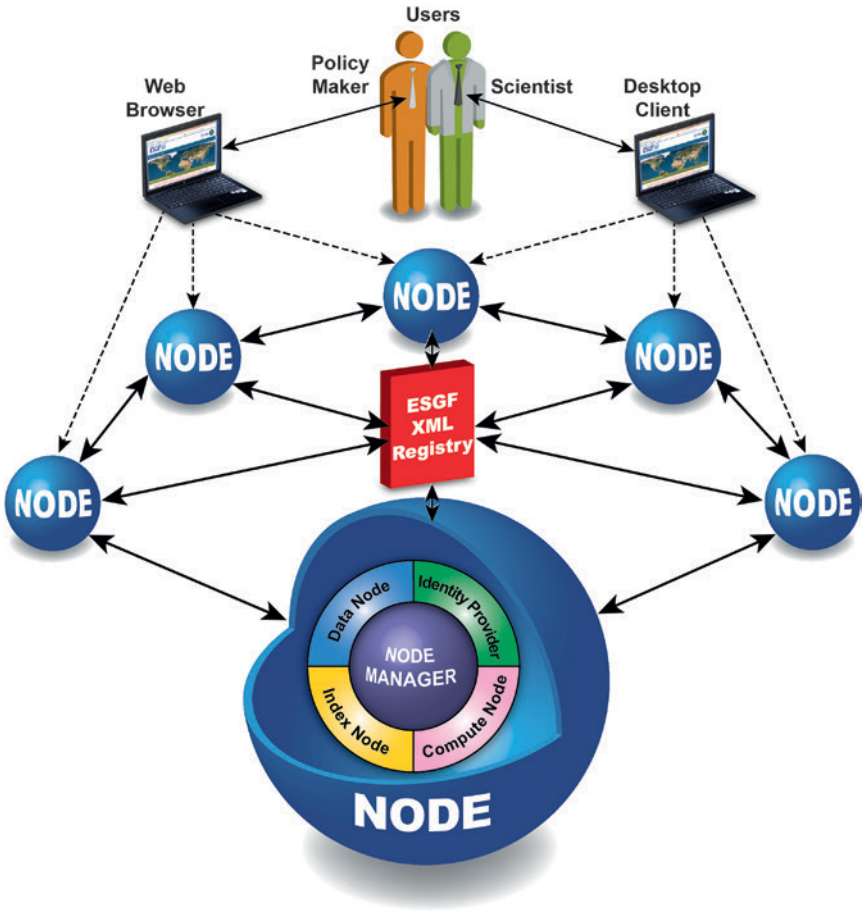

Fig. 3. ESGF's peer-to-peer architecture is based on the principles of modular components and standard protocols. Each system node, or software stack (shown in blue), can be configured to possess one or more "software types," each entailing a specific functionality: the "data node" publishes data and serves data through a variety of protocols; the "index node" harvests metadata and enables data discovery; the "identity provider" registers, authenticates, and authorizes users; and the "compute node" handles computational resources for data reduction, analytics, and visualization. All federated nodes interact as equals, so no single points of failure arise.

federation include LLNL and other national laboratories, as well as international organizations such as DKRZ, the British Atmosphere Data Centre (BADC), and the University of Tokyo Centre for Climate System Research, to name only a few.

A NEXT-GENERATION SYSTEM. Instead of consolidating all the data from numerous locations, as was done for CMIP3, the flexible, scalable ESGF framework links storage and open-source software assets distributed across many distinct networks and locations into a virtual system that is more powerful, robust, and capacious than any of its constituent networks or systems (Fig. 2). Participating modeling research centers download the ESGF software stack from the community-based repository and then install and use this software to immediately publish their data to the federation for collection and integration.

Although data and metadata are published, stored, and delivered by nodes around the globe and controlled by different authorities, they are searchable and accessible as if they resided in a single global archive (Fig. 3). Similar to online shopping, ESGF users can query the archive for what they need, add the data to a "shopping cart," and download it when they are ready. More than 20 web portals for registering and accessing ESGF data and services and 50 nodes for data and software provision are presently in use. The CMIP5 core data nodes (or data centers) are LLNL, DKRZ, and BADC. These sites back up one another's data for better user access and archival purposes. Major data centers such as the National Computational Infrastructure (NCI) in Australia and the Institut Pierre-Simon Laplace (IPSL) in France also joined the core data nodes to assist with data distribution and provide additional resources.

ESGF supports geospatial and temporal searches and includes a dashboard that shows system metrics, a user interface for notifications, and a rich set of analysis tools [such as the Climate Data Operators, Network Common Data Form (NetCDF) operators, Ultrascale Visualization Climate Data Analysis Tools (UV-CDAT), Ferret, and Grid Analysis and Display System (GrADS)] to help manipulate the data. For example, UV-CDAT (Williams 2014; see sidebar "Many data analysis tools in one") can automate many aspects of scientific analysis and visualization, making it easy for users to rerun analyses and to work together (Fig. 4). ESGF also enforces standard conventions for data transformation, quality control, and data validation across processes and projects. By ensuring that the way temperature, for example, is defined is the

MANY DATA ANALYSIS TOOLS IN ONE

A key component of the Earth System Grid Federation, the Ultrascale Visualization Climate Data Analysis Tools (UV-CDAT) assists climate researchers in solving complex data analysis and visualization challenges. The product of an award-winning partnership between four Department of Energy national laboratories, two universities, two information technology companies, the National Oceanic and Atmospheric Administration, and the National Aeronautics and Space Administration, UV-CDAT is an open-source, easy-to-use application that links more than 70 disparate software subsystems and packages to form an integrated environment for analysis and visualization. The tool set provides users access to more analysis and visualization products than any other single source.

Since none of these tools was specifically designed for climate research, climate scientists have built individual analyses around each tool, in the process often recreating code developed by other scientists. As part of integrating these tools into the UV-CDAT framework, its developers have customized them to process climate data, perform common analyses, and automate routine data operations, which free users to concentrate on scientific analysis rather than on the mundane chores of data movement and manipulation. Moreover, UV-CDAT tracks analysis workflows. This enables climate scientists to retrace the methods of fellow researchers and thereby supports scientific transparency.

same for all models, ESGF facilitates comparisons between datasets.

Over the past decade, ESGF has become an integral part of the way climate scientists carry out their work; today, roughly 27,000 users (researchers and nonresearchers) from 2,700 sites on six continents access data through ESGF. A key to ESGF's success has been its ability to produce, validate, and analyze research results collaboratively, so that new results generated by one team member are immediately accessible to the rest of the team, who can annotate, comment on, and otherwise interact with those results. The system also handles version control in a clever fashion; if the new results entail improvements to an existing computer model, for example, the system will automatically archive the results and notify individuals and data centers storing outdated output versions.

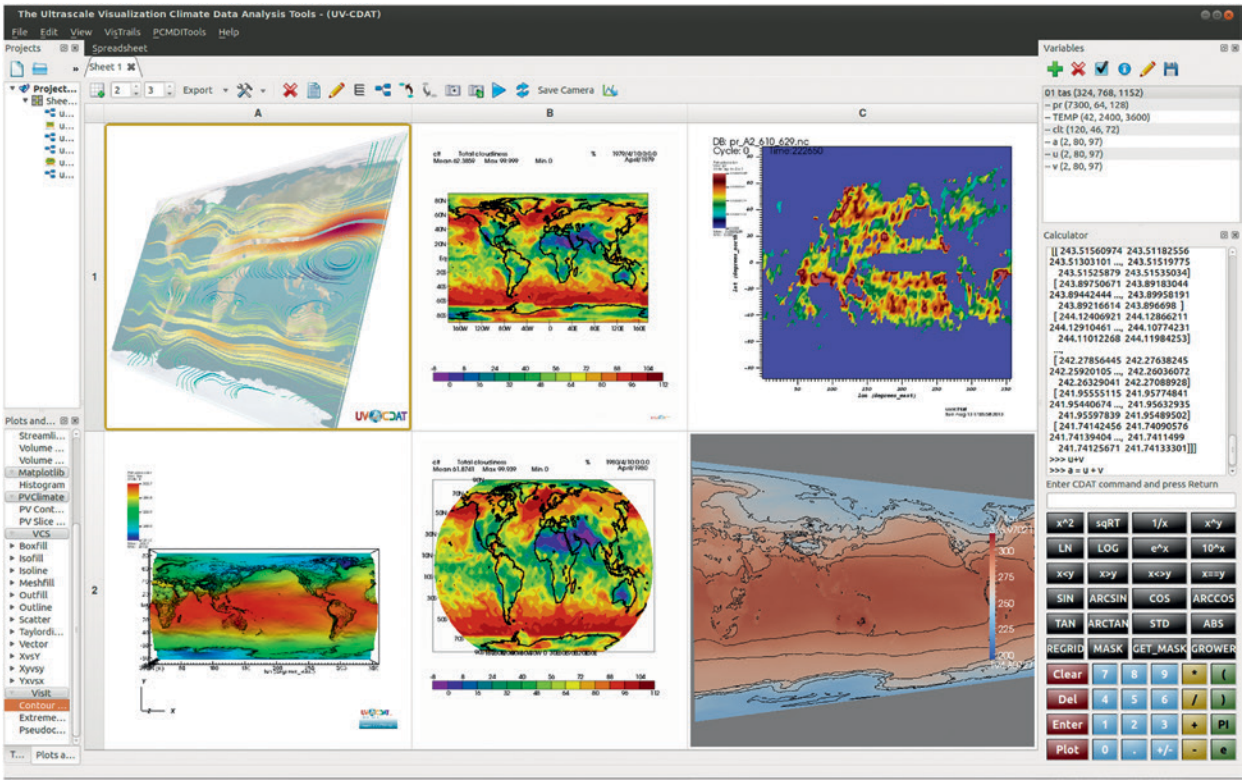

Fig. 4. Using intuitive drag-and-drop operations, scientists can create, modify, rearrange, and compare visualizations in UV-CDAT. Shown is output from four disparate analysis tools within the UV-CDAT framework: (top left) DV3D, (bottom left) R plot, (middle) CDAT plots, (top right) Visit, and (bottom right) ParaView.

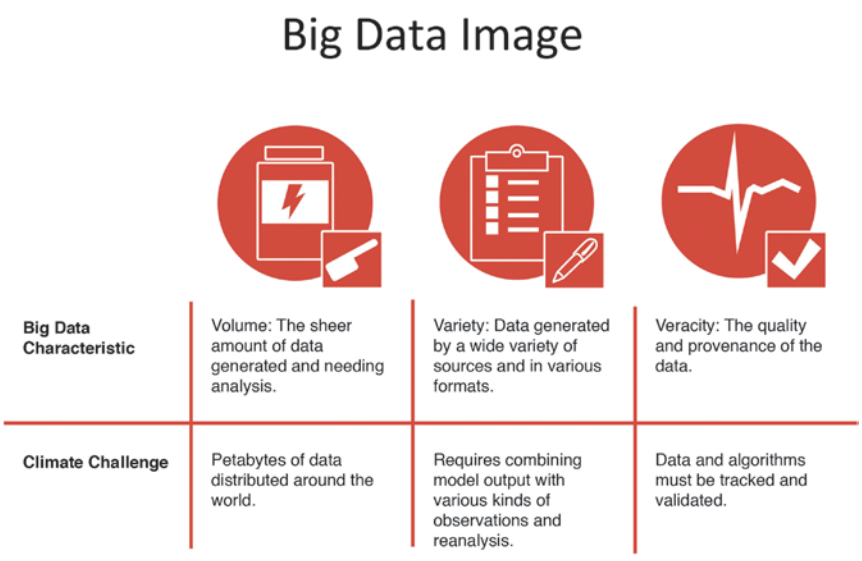

Fig. 5. Climate "big data" characteristics and challenges.

ONGOING CLIMATE FEDERATION CHALLENGES. ESGF and climate science are facing major data challenges now and in the future (Fig. 5). First, managing the ever-increasing data variety, complexity, velocity, and volume in a way that facilitates scientific research is a significant challenge. Second, ESGF must ensure that the fruits of climate research are accessible to and useful for a broad, growing, and interdisciplinary audience (Williams et al. 2014). Finally, an ongoing challenge is rooted in the federation's nature as an entity with limited resources and a set of needs and priorities as diverse as the science communities it supports.

By 2020, ESGF will likely house hundreds of petabytes (quadrillions of bytes) of environmental and scientific data, provided by observational satellites, instruments, model outputs, and more (Overpeck et al. 2011). An increasingly common and necessary activity in climate science is to work with datasets from multiple sources; this requires a robust search capability with power processes that understands diverse data formats, advanced management of distributed resources (e.g., networks and compute facilities), and a storage and management system scalable to billions of files. With tens of thousands of users already relying on access to the data, the federation also needs a system with high levels of reliability and performance. Since ESGF is not under the administrative control of a single organization, this is not a simple matter. Data availability is presently subject to the downtimes and outages of the various data nodes, most of which are run on a quasi-volunteer (nonfunded) basis. This liability in the federation can thus be characterized as a point of failure.

The interest in and need for climate data are growing, but since ESGF was designed primarily for climate scientists and assumes a fairly high level of technical understanding, the system can be daunting to navigate for individuals outside of the climate community, such as city planners, public health officials, resource managers, and more. ESGF needs to have products available for a wider range of users, both researchers and nonresearchers, and more intuitive methods for accessing those data for interested parties who are not "expert users."

Datasets have reached a point where they are taxing computing power, network bandwidth, and storage space at many research facilities, and the next generation of datasets will be even larger. Currently, datasets may consist of many files, with file sizes on average ranging from 460 MB to a terabyte or more (see sidebar "How does ESGF compare to other data systems?"). ESGF already allows users to access and perform some data analysis on files before

HOW DOES ESGF COMPARE TO OTHER DATA SYSTEMS?

The world is full of large-scale data management and retrieval enterprise systems, yet few employed by the climate community offer the features and flexibility of ESGF. The United States alone is home to half a dozen climate data systems, such as the Regional Climate Model Evaluation System and the NASA Distributed Active Archive Centers, but none boasts a distributed data model or interoperability among disparate datasets (such as simulations, observations, and reanalysis) as does ESGF. Even Google's data centers, the largest in the world, are not in the business of serving up large datasets for raw downloads as is the case for ESGF. Further, a search on ESGF from anywhere in the world will produce identical data, making the world a small, easily navigable community.

downloading them to their networks or computer systems, but developing methods to further server-side (i.e., remote) data manipulation, analysis, and visualization—to minimize or eliminate data movement—will be crucial for small- and medium-sized institutions to continue their participation in comparative climate study. Some of the analysis algorithms and uncertainty quantification (UQ) tasks that climate researchers would like to perform are extremely computationally intensive. UQ is an expanding field of science that focuses on quantifying the accuracy of simulated results—in particular, which of several possible outcomes are most likely to occur (Fig. 6). Executing rapid analysis and UQ on large data files with minimal data movement will require coupling local or remote high-performance computing capabilities to analysis tools such as UV-CDAT, among other tasks.

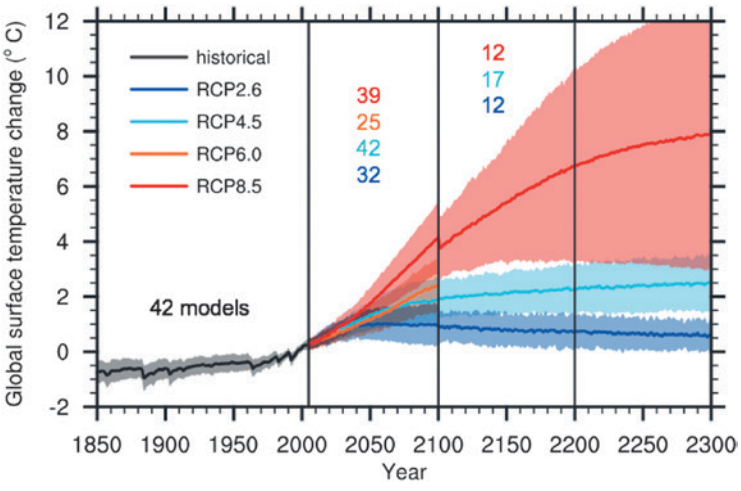

Fig. 6. Time series of global annual-mean surface air temperature anomalies (relative to 1986–2005) from CMIP5 concentration-driven experiments. Projections are shown for each representative concentration pathway (RCP) experiment for the multimodel mean (solid lines) and ±1.64 standard deviations (5%–95%) across the distribution of individual models (shading), based on annual means. The 1.64 standard deviation range based on the 20-yr averages of 2081–2100, relative to 1986–2005, are interpreted as likely changes for the end of the twenty-first century. The numerals (i.e., 42, 39, 25, 42, 32, 12, 17, and 12) in the plot represent the number of models used to generate the RCP projections (IPCC 2013).

Modeling groups want credit for the data they produce. Researchers also need to document what data were used in published research, necessitating the use of digital object identifiers (DOIs). To simplify these tasks, ESGF needs to capture the provenance of its data from start to finish. Provenance allows scientists to see how a dataset was created (which version of a model was used, which variables were chosen, etc.), how two results differ in the way they were created, and how a set of results could be recreated. Without knowing the origin and quality of published data, scientists have no way of trusting the results. This is analogous with the need to improve the quality of data and data services within ESGF.

EMBRACING A BROADER SWATH OF RESEARCH AND DATA. As climate community processes and needs have evolved over the past decade, data scientists have adapted ESGF's flexible software architecture (Fig. 7). In the coming years, ESGF will continue to improve and serve as the foundation for international climate data management, as the federation develops and implements the next generation of research tools. The federation's ambitious vision calls for a scalable, cloud-based ecosystem that supports collaborative, extreme-scale science (Advanced Scientific Computing Advisory Committee Subcommittee 2010) across a range of communities, projects, and disciplines (see sidebar "A test bed for new climate models"). Components for navigating, managing, manipulating, and sharing data will be made available as open-source modules that can be customized and extended by user communities. For instance, the federation is currently collaborating with private and public institutions and universities to build user access tools and modes to ensure smooth access by policy makers and social scientists to the most popular data products and analyses. Data scientists are also working to support a wider range of data types to meet the needs of the broader international Earth system research community by supplying the most popular data such as temperature and precipitation for regional climate study.

For reproducibility and efficiency purposes, tomorrow's climate data infrastructure must integrate, automate, and record every possible stage in the data life cycle. Achieving this goal requires the use of workflows. Workflows standardize scientific experiments, data processing, and analysis and can be used as building blocks to support more complex problems. They can also automatically record provenance information. Related efforts are under way to make more data manipulation activities occur remotely, with minimal movement, through a coordination of data management, security, and analysis tools. This would,

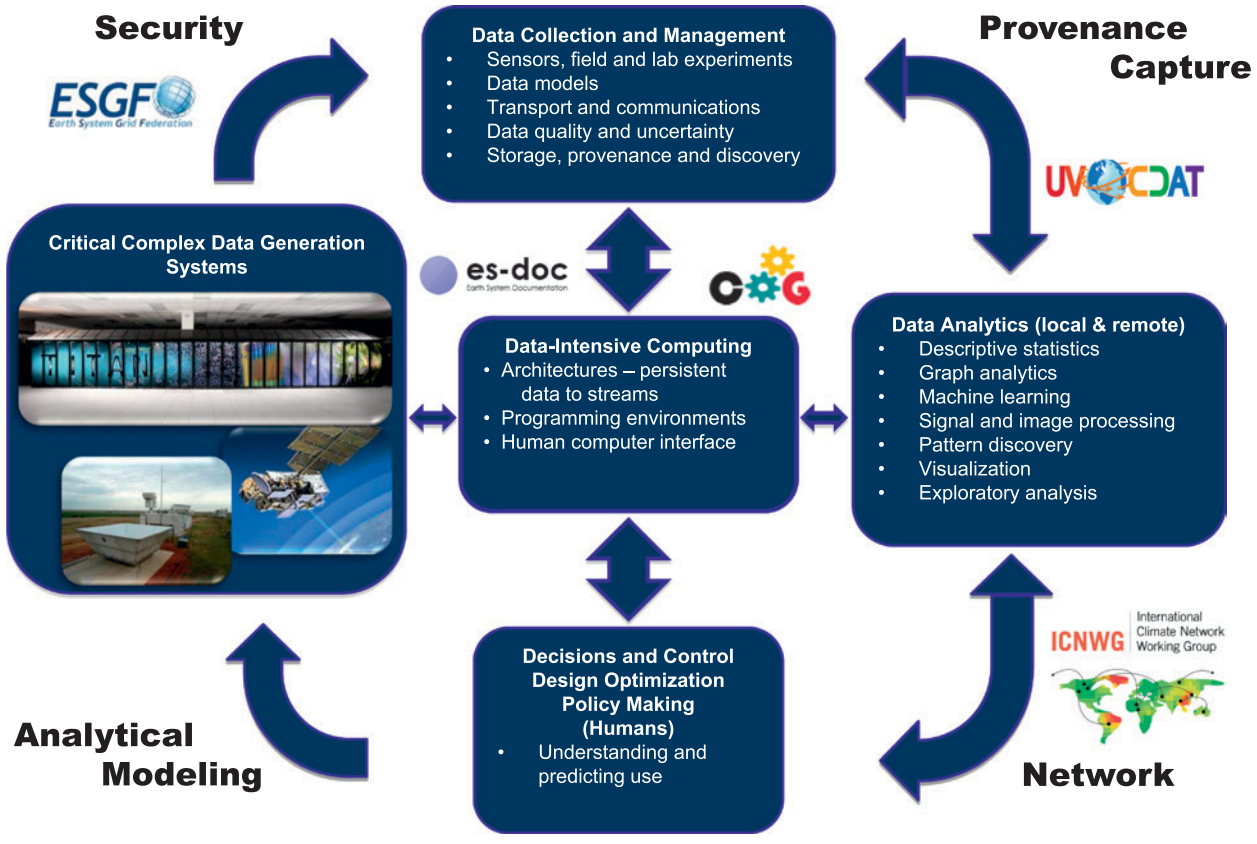

Fig. 7. The proposed components of the climate community's integrated data ecosystem and workflow currently under development, with comprehensive provenance capture. This integrated data ecosystem tightly interweaves every aspect of climate data research, from model development through interpretation and dissemination of research results.

for instance, obviate the need to download massive multifile data collections for an analysis that may require only a small subset of the data.

The federation's enhanced ecosystem for climate and Earth system research will require rapid and reliable access to data (Fig. 8). To ensure data availability even during partial system outages, efforts are under way to implement smart caching, dynamic data replication, and publication of data at multiple locations. Identical copies of climate datasets will reside in supernodes hosted by seven climate data centers on four continents—Beijing Normal University, BADC, DKRZ, LLNL, NCI, IPSL, and the University of Tokyo—from which the data will be made available to the international climate community. These ESGF data replication supernodes will be paired with compute servers, to further minimize data movement demands. A fast and secure data network connecting

A TEST BED FOR NEW CLIMATE MODELS

In 2014, eight Department of Energy (DOE) national laboratories, four academic institutions, one company, and the National Center for Atmospheric Research combined forces in a project called Accelerated Climate Modeling for Energy (ACME), which is designed to accelerate coupled Earth system model development for climate and energy applications. Over a planned 10-year span, the project will conduct simulations and modeling on DOE's most powerful high-performance computing systems. Initial focus will be on three key climate change science drivers: the atmosphere, land surface, and ocean and sea ice. It is an example of what the community would call coupled modeling.

ACME's model test bed will be integrated into the federation's next-generation climate data infrastructure. Access to ESGF's model data storage, data analysis and visualization, workflow automation, and provenance capture features will expedite efforts by collaborating DOE scientists to develop, test, evaluate, and use sophisticated new models.

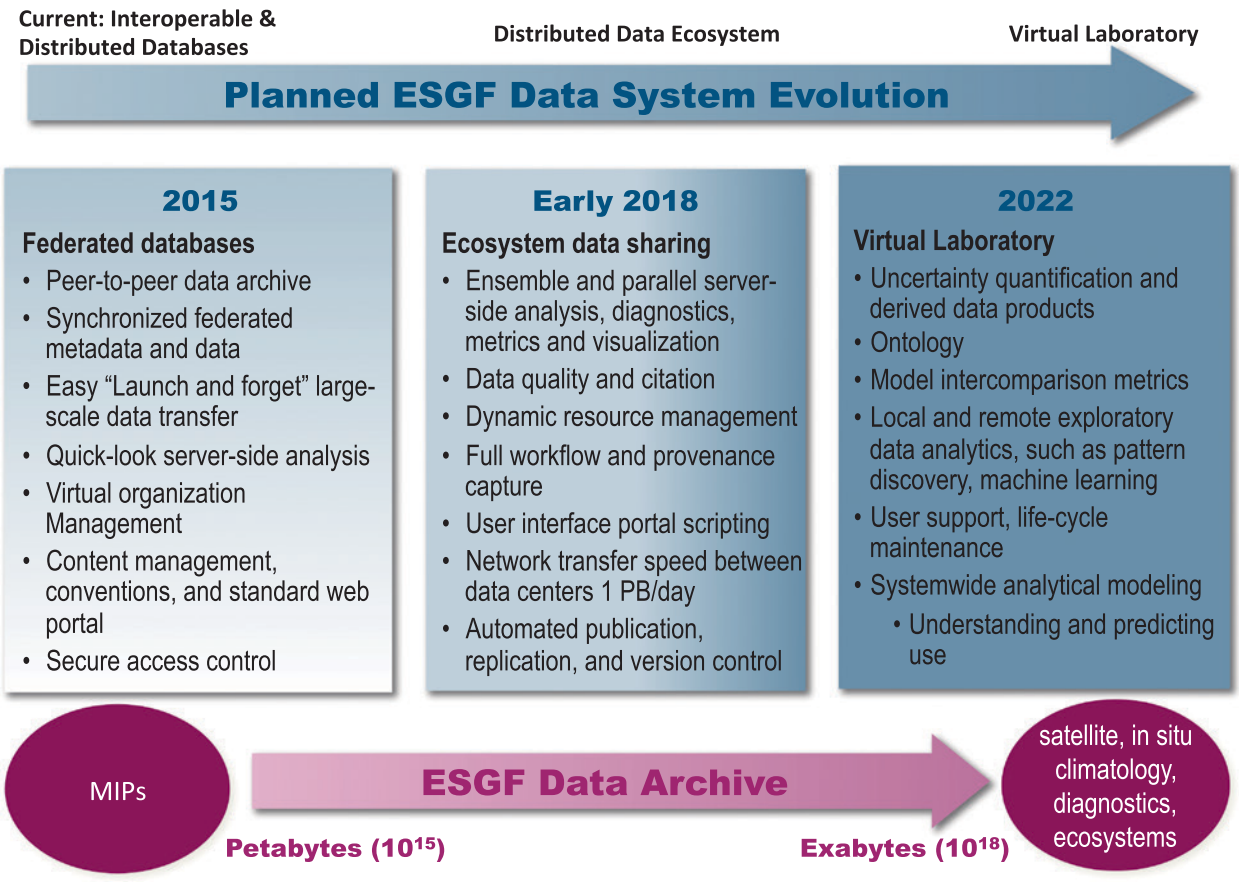

Fig. 8. A close collaboration between data scientists, climate scientists, and scientists from other domains is enabling the development of a data ecosystem that can support a broad variety of data and disciplines. Ecosystem features will be incorporated into the ESGF system incrementally, over the next 5 yr.

the primary climate data and modeling centers will also prove essential for those inevitable situations when data will need to be moved. It is projected, for example, that tens of petabytes of data will need to be replicated among data centers for CMIP6.

ENHANCEMENTS TO SCIENTIFIC WORKFLOW AND SYSTEM PERFORMANCE. Realizing the next stage of climate data management will take significant coordination and several years of work by participating data scientists, progress is under way to significantly revamp the ESGF infrastructure—with some associated features becoming available as soon as this year. Several of these near-term improvements will streamline tasks for scientists using the system. For instance, the publication tool, which allows users to move data into an archive for storage and hosting, presently requires the user to complete a series of tedious tasks to publish a file (Fig. 9). Data scientists will soon be implementing a new "launch and forget" data transfer service model that automates many steps and simplifies the process. Enhancements are also under way for the ESGF dashboard, which provides users with high-level, real-time metrics for the federation such as usage and service availability in a customizable format. The upgraded dashboard will display network information and other managed resources from the major climate data centers and will feature an improved user interface that enables scientists to host, manage, and share scientific projects seamlessly, from any location.

Data scientists are enhancing several components that operate behind the scenes, too, including security and node management and network performance. Recognizing that balancing ease of use with data security is a common challenge for information systems, ESGF data scientists are working with Amazon and studying organizations such as Facebook to glean new ideas for improving the ESGF security process (see sidebar "Lessons from CMIP5"). More immediately, ESGF will be transitioning to an open standard for its authorization security framework for sharing credentials, which will enable the use of a single sign-in system via a simple web interface. ESGF data

Fig. 9. A snapshot from the current version of the dashboard, which provides users with status updates on federation data and services.

LESSONS FROM CMIP5

From the user's perspective, the initial CMIP5 Earth System Grid Federation (ESGF) was an imperfect system. A slow search interface, a complex and restrictive user authentication method, and network and system availability issues were a few of the pitfalls users faced. ESGF developers organized into smaller subgroups to tackle specific technical issues and concerns. The "scaled agile" development process—highlighting individual roles, teams, and activities—allowed for the adjustment of schedules and priorities as necessary to quickly provide new solutions and meet the customers' ever-changing demands.

Also a concern was the stability and governance of ESGF, as it is funded by disparate government agencies, each wanting to achieve their own projects' aims. By setting up an official ESGF governance board (http://esgf.llnl.gov/governance.html) with representatives from the funding agencies (i.e., the Steering Committee) and their respective principal investigators (i.e., the Executive Committee), ESGF is now receiving the direction and oversight needed to support climate science community projects.

To avoid duplicate data transfers within single institutions, several CMIP5 communities set up "snapshots" of the archive for local use, or use within collaborative user communities. Snapshots serve the dual purpose of providing a local archive for close-knit communities to share and a fixed reference point (as the name snapshot implies) to compare with the continually updated federated archive. The IPCC Working Group I Snapshot (http://wiki.c2sm.ethz.ch/Wiki/CMIP5) was begun by ETH-Zürich to provide a reference dataset for a large number of CMIP5 studies (e.g., Knutti and Sedláček 2013), while the AR5 Reference Snapshot (www.ipcc-data.org/sim/gcm_monthly/AR5/Reference-Archive.html) contains the "dataset of record" for CMIP5 content that has undergone rigorous quality control and been issued DOIs. Many other centers have used these snapshots to construct private repositories and reduce pressure and duplicate downloads from ESGF data nodes. A community-developed replication tool (synda; https://github.com/Prodiguer/synda) was built to facilitate snapshot creation and replication. One lesson for CMIP6 and beyond is for the ESGF community to accommodate snapshots into the ESGF design.

In response to recent security attacks at lead institutions on ESGF servers, the installation team has incorporated security scans into the ESGF software release process. Scans entail both finding and fixing security infractions and code violations within individual software components and monitoring dynamic web portals used as scientific gateways into the ESGF archive. Through these scans, ESGF is fulfilling the security requirements mandated by many of the modeling and data centers.

scientists are also designing a new node manager, an organizational tool for gathering metrics, sharing node information across the federation of nodes, and managing user groups. The next version will have a hierarchical structure that increases the tool's scalability and fault tolerance for system reliability.

Finally, data scientists are tackling the issue of network performance. Most ESGF data must traverse four or five network domains, often owned and operated by different organizations in different countries, before reaching their destination, which can make it difficult to locate and diagnose bottlenecks in the data transfer process and get the responsible parties to fix the issues. The International Climate Network Working Group (ICNWG) has deployed tools to track network speed and usability, beginning with the networks that connect internationally hosted ESGF nodes, and is partnering with institutions to identify local network, server, and storage components to upgrade to achieve reliable high-speed data transfer. By the end of 2016, five major climate research centers expect to achieve a transfer rate of 1 GB a second—as much as 100 times faster than some sites were experiencing at the start of the project in 2014.

Fig. 10. Data experts from roughly two dozen organizations in half a dozen countries participated in the 2014 ESGF and UV-CDAT Conference, held in Livermore, CA, where the climate community's data management needs were discussed, and updates of progress in data tools and infrastructure since the last conference were shared. The annual event is one of many methods climate data scientists are using to engage with their user communities, solicit feedback, and share their breakthroughs. The full conference report is available online (Williams et al. 2015).

INVOLVING COMMUNITIES. Eighteen international research projects use the ESGF infrastructure to globally manage and distribute their data. ESGF data scientists selected several of these research communities to help provide focus for the climate ecosystem development effort. One of these targeted communities encompasses the global climate modeling groups in dozens of countries that contributed to CMIP5 and are now participating in CMIP6. CMIP researchers typically run the same prescribed set of climate change scenarios on the most powerful available supercomputers, to produce datasets containing hundreds of physical variables and spanning tens or hundreds of years. Participants in the Coordinated Regional Climate Downscaling Experiment (CORDEX) use

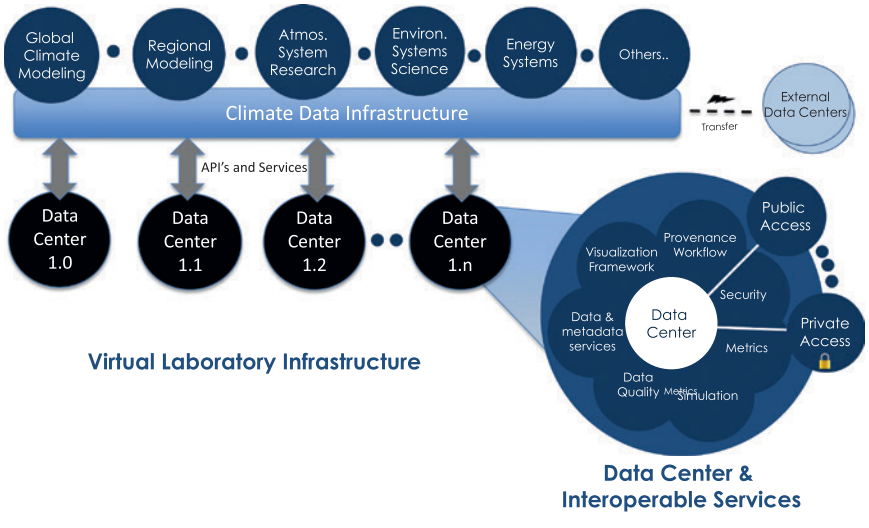

Fig. 11. The integrated cyber infrastructure under development will leverage core global climate resources to enable discovery, analytics, simulation, and knowledge innovation in Earth system research.

various techniques to simulate Earth's climate system at a higher spatial resolution over more limited areas than CMIP. CORDEX scientists are working to forecast how Earth's climate may change regionally. To ensure data coordination, CORDEX recently decided to adopt the same ESGF infrastructure as CMIP.

The Accelerated Climate Modeling for Energy (ACME; ACME Council 2014) project (see sidebar "A test bed for new climate models"), designed to create and operate a test bed for advanced Earth system model development, has among the most varied data management needs. ACME scientists will be performing many short model runs with rapid turnaround during the model development phase, more computationally demanding UQ and optimization work for model refinement, and massive data runs on leading Department of Energy (DOE) supercomputers with the full array of ESGF features once the models are in production.

The last community featured in this article is the most nebulous, as it encompasses various scientific domains affected by climate. As an example, many diseases have strong correlations with short- and long-term weather patterns, from the simple—cold noses are more susceptible to the common cold—to the complex—weather, acorn crops, acorn-eating mice, and mice ticks all contribute to the spread of Lyme disease to human populations. By integrating disease outbreak and climate data within the data ecosystem's flexible infrastructure, epidemiologists and public health experts will be able to better understand how the data relate and to develop predictive models for individual diseases.

The common thread that links these various user communities and the data scientists developing the new generation of software infrastructure is communication. Extensive community input and involvement is needed to ensure that the system supports research communities in meeting their scientific missions, both as individual projects and as a federation of collective projects. Opportunities to provide feedback and define user requirements have included town hall meetings at major scientific societies, such as the American Geophysical Union, the European Geophysical Union, and the American Meteorological Society; community workshops held by the Global Organization for Earth System Science Portals; use, data, and modeling surveys; and an annual ESGF and UV-CDAT conference (Fig. 10). Information about the project has been shared through a succession of reports and presentations.

A GAME CHANGER. Despite the inevitable growing pains, ESGF—one of the most complex big-data systems in existence—has been a tremendous success. Climate scientists have hailed ESGF as a game changer. The American Meteorological Society in 2010 lauded ESGF for ushering in "a new era in climate system analysis and understanding," and ESGF and UV-CDAT have won Federal Laboratory Consortium awards for outstanding technical partnership each of the past three years. But ESGF is by no means a finished product because climate modeling is ever increasing in complexity and sophistication.

Scientific progress in the climate realm is critically dependent on the availability of a reliable infrastructure for data access and management (Fig. 11). The next-generation data ecosystem will help meet this need and ensure that the scientific investigation is completely transparent, collaborative, and reproducible, which are crucial attributes, given the field's high visibility and direct impact on climate research.

ACKNOWLEDGMENTS. The authors wish to thank the participants at the 2014 Earth System Grid Federation and Ultrascale Visualization Climate Data Analysis Tools Conference, whose presentations and conference report input helped considerably in the development of this article (Williams et al. 2015). This work was supported by the U.S. Department of Energy Office of Science/Office of Biological and Environmental Research under Contract DE-AC52-07NA27344 at Lawrence Livermore National Laboratory. VB is supported by the Cooperative Institute for Climate Science, Princeton University, under Award NA08OAR4320752 from the National Oceanic and Atmospheric Administration, U.S. Department of Commerce. Part of this work was undertaken with the assistance of resources from the National Computational Infrastructure (NCI), which is supported by the Australian Government. Part of this activity was performed on behalf of the Jet Propulsion Laboratory, California Institute of Technology, under a contract with NASA. Part of this activity was performed on behalf of the Goddard Space Flight Center, under a contract with NASA. This work was supported by ANR Convergence project (Grant Agreement ANR-13-MONU-0008). This work was supported by FP7 IS-ENES2 project (Grant Agreement 312979). The statements, findings, conclusions, and recommendations are those of the authors and do not necessarily reflect the views of Princeton University, the National Oceanic and Atmospheric Administration, or the U.S. Department of Commerce.

APPENDIX. *Glossary of terms.*

ACME	Accelerated Climate Modeling for Energy. DOE's effort to build an Earth system modeling capability tailored to meet the climate change research strategic objectives
CMIP5	Coupled Model Intercomparison Project, phase 5, sponsored by WCRP/WGCM, and the related multimodel database planned for the IPCC's Fifth Assessment Report (AR5; http://cmip-pcmdi.llnl.gov)
CORDEX	Coordinated Regional Climate Downscaling Experiment, providing global coordination of regional climate downscaling for improved regional climate change adaptation and impact assessment (http://wcrp-cordex.ipsl.jussieu.fr/)
Data node	Internet location providing data access or processing (http://en.wikipedia.org/wiki/Node-to-node_data_transfer)
DOE	Department of Energy, the U.S. government entity chiefly responsible for implementing energy policy (www.energy.gov/)
Ecosystem	A complex network of interconnected systems and resources such as high-performance computers, networks, storage facilities, software, and algorithms (see Fig. 7)
ESGF	Earth System Grid Federation, led by LLNL, a worldwide federation of climate and computer scientists deploying a distributed multipetabyte archive for climate science (http://esgf.llnl.gov)
ICNWG	International Climate Network Working Group, formed under the ESGF, is tasked to help set up and optimize network infrastructure for their climate data sites located around the world (http://icnwg.llnl.gov/)
IPCC	Intergovernmental Panel on Climate Change, a scientific body of the United Nations, which periodically issues assessment reports on climate change (www.ipcc.ch/)
LLNL	Lawrence Livermore National Laboratory, sponsored by the DOE (www.llnl.gov/)
Metadata	Data properties, such as their origins, spatiotemporal extent, and format (http://en.wikipedia.org/wiki/Metadata)
UQ	Uncertainty quantification is a method for determining how likely a particular outcome is, given the inherent uncertainties or unknowns in a system (http://en.wikipedia.org/wiki/Uncertainty_quantification)
UV-CDAT	Ultrascale Visualization Climate Data Analysis Tools provides access to large-scale data analysis and visualization tools for the climate modeling and observational communities (http://uvcdat.llnl.gov)
WCRP	World Climate Research Programme, which aims to facilitate analysis and prediction of Earth system variability and change for use in an increasing range of practical applications of direct relevance, benefit, and value to society (www.wcrp-climate.org/)
Web portal	A point of access to information on the world wide web (http://en.wikipedia.org/wiki/Web_portal)
WGCM	Working Group on Coupled Modeling (www.wcrp-climate.org/wgcm/)

REFERENCES

ACME Council, 2014: Accelerated Climate Modeling for Energy: Project strategy and initial implementation plan. ACME Council, Department of Energy, 25 pp. [Available online at http://climatemodeling.science.energy.gov/sites/default/files/publications/acme-project-strategy-plan.pdf.]

Advanced Scientific Computing Advisory Committee Subcommittee, 2010: The opportunities and challenges of exascale computing. ASCAC Subcommittee Rep. on Exascale Computing, ASCAC, Department of Energy, 71 pp. [Available online at http://science.energy.gov/~/media/ascr/ascac/pdf/reports/Exascale_subcommittee_report.pdf.]

Cinquini, L., and Coauthors, 2014: The Earth System Grid Federation: An open infrastructure for access to distributed geospatial data. *Future Gener. Comput. Syst.*, **36**, 400–417, doi:10.1016/j.future.2013.07.002.

IPCC, 2013: *Climate Change 2013: The Physical Science Basis*. Cambridge University Press, 1535 pp., doi:10.1017/CBO9781107415324.

Knutti, R., and J. Sedláček, 2013: Robustness and uncertainties in the new CMIP5 climate model projections. *Nat. Climate Change*, **3**, 369–373, doi:10.1038/nclimate1716.

Meehl, G. A., R. Moss, K. E. Taylor, V. Eyring, R. J. Stouffer, S. Bony and B. Stevens, 2014: Climate Model Intercomparison: Preparing for the next phase. *Eos, Trans. Amer. Geophys. Union*, **95** (9), 77–78, doi:10.1002/2014EO090001.

Overpeck, J. T., G. A. Meehl, S. Bony, and D. R. Easterling, 2011: Climate data challenges in the 21st century. *Science*, **331**, 700–702, doi:10.1126/science.1197869.

Williams, D. N., 2014: Visualization and analysis tools for ultrascale climate data. *Eos, Trans. Amer. Geophys. Union*, **95** (42), 377–378, doi:10.1002/2014EO420002.

——, G. Palanisamy, G. Shipman, T. A. Boden, and J. W. Voyles, 2014: Department of Energy strategic roadmap for Earth system science data integration. *Proc. Conf. on Big Data*, Washington, DC, IEEE, 772–777, doi:10.1109/BigData.2014.7004304.

——, and Coauthors, 2015: Fourth Annual Earth System Grid Federation and Ultrascale Visualization Climate Data Analysis Tools Conference report. Lawrence Livermore National Laboratory Tech. Rep. LLNL-TR-666753, 73 pp. [Available online at http://aims-group.github.io/pdf/2014-ESGF_UV-CDAT_Conference_Report.pdf.]