Lawrence Livermore National Laboratory
LLNL-MI-625433

Earth System Grid Federation

The Earth System Grid Federation (ESGF, esgf.org) is an international collaboration for the software that powers most global climate change research, notably assessments by the International Panel on Climate Change. The portal below is for the Program for Climate Modeling Diagnosis and Intercomparison at Livermore, one of dozens federated around the world (http://pcmdi9.llnl.gov).

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

1. ORGANIZATION CONTACT INFORMATION

A. Submitter Information

Organization Name:	Lawrence Livermore National Laboratory
Contact Name:	Dean N. Williams
Address:	7000 East Avenue, L-103
City/State:	Livermore, California
Zip/Postal Code:	94550-9234
Country:	United States of America
Phone:	(925) 423-0145
Fax:	(925) 422-7675
Email:	williams13@llnl.gov
Web URL:	http://www-pcmdi.llnl.gov/about/staff/Williams/index.php

Indicate relationship to product being submitted:
__ Public Relations or Marketing Agency
X Product Developer
__ Other (explain) _____

(Provide marketing and media contacts in Appendix B)

B. Developer Organization Information

Organization Name:	**Lawrence Livermore National Laboratory (LLNL)**
Contact Name:	Dean N. Williams
Address:	7000 East Avenue, L-103
City/State:	Livermore, California
Zip/Postal Code:	94550-9234
Country:	United States of America
Phone:	(925) 422-1100
Fax:	(925) 422-1370
Email:	williams13@llnl.gov
Web URL:	www.llnl.gov

What role did this organization play in development of technology?
X Principal developer organization (All or majority of development responsibility)
__ Co-developer organization (Equal share of development responsibility with other organizations)
__ Supporting developer organization (Contributed less than 50% of development responsibility)

Provide details of development role: LLNL led the coordination of the Coupled Model Intercomparison Project, phase 5, (CMIP5) and the international development of the Earth System Grid Federation (ESGF) delivery production environment for managing and accessing petascale climate data used in the 2013 Intergovernmental Panel on Climate Change (IPCC) Fifth Assessment Report (AR5), among others. More specifically, LLNL was the overall coordinator and primary framework architect, and

assisted in the development of the compute node, security services, and user interface web development.

Organization Name: Jet Propulsion Laboratory (JPL)

What role did this organization play in development of technology?
__ Principal developer organization (All or majority of development responsibility)
__ Co-developer organization (Equal share of development responsibility with other organizations)
X Supporting developer organization (Contributed less than 50% of development responsibility)

Provide details of development role: For ESGF, JPL is leading the NASA observation for Model Intercomparison Projects (obs4MIPs) effort, which provides essential satellite observational data for model intercomparison and analysis. JPL reprocessed and reformatted observational data for comparison with the CMIP5 model output. JPL shared in the role of primary framework architect and has led in the development of federated security services, which allow single sign-on and distributed access control across the federation, and the search services for efficient, federation-wide discovery of distributed data holdings.

Address:	4800 Oak Grove Drive
City:	Pasadena
State:	California
Zip/Postal:	91109
Country:	United States of America
Web URL:	http://www.jpl.nasa.gov/
Contact Name:	Luca Cinquini
Phone:	(818) 309-0680
Fax:	(818) 393-6141
Email:	Luca.Cinquini@jpl.nasa.gov

Organization Name: Oak Ridge National Laboratory (ORNL)

What role did this organization play in development of technology?
__ Principal developer organization (All or majority of development responsibility)
__ Co-developer organization (Equal share of development responsibility with other organizations)
X Supporting developer organization (Contributed less than 50% of development responsibility)

Provide details of development role: ORNL led the user interface design and implementation. The front-end design accommodates a wider community of potential users that may be interested in various data sets (e.g., scientists interested in model, observational, and reanalysis data for comparison). ORNL is also assisting in providing ESGF access to petabytes of data located on deep-archives (or tape storage systems) such

as the High Performance Storage Systems (HPSSs) found at ORNL and Lawrence Berkeley National Laboratory.

Address:	1 Bethel Valley Road
City:	Oak Ridge
State:	Tennessee
Zip/Postal:	37831
Country:	United States of America
Web URL:	http://www.ornl.gov
Contact Name:	Galen Shipman
Phone:	(865) 576-2672
Fax:	(865) 574-6076
Email:	gshipman@ornl.gov

Organization Name: Pacific Marine Environmental Laboratory (PMEL)

What role did this organization play in development of technology?
__ Principal developer organization (All or majority of development responsibility)
__ Co-developer organization (Equal share of development responsibility with other organizations)
X Supporting developer organization (Contributed less than 50% of development responsibility)

Provide details of development role: PMEL led the development of the product-services user interface, which includes the integration of the Live Access Server (LAS) user interface and data services with the ESGF web portal, and interaction with the ESGF security infrastructure. PMEL improved the Ajax communication interfaces to facilitate cleaner and more user-friendly data discovery via the ESGF portal.

Address:	7600 Sand Point Way NE
City:	Seattle
State:	Washington
Zip/Postal:	98115
Country:	United States of America
Web URL:	http://www.pmel.noaa.gov/
Contact Name:	Steve Hankin
Phone:	(209) 526-6080
Fax:	(209) 526-6239
Email:	Steven.C.Hankin@noaa.gov

Organization Name: Argonne National Laboratory (ANL)

What role did this organization play in development of technology?
__ Principal developer organization (All or majority of development responsibility)
__ Co-developer organization (Equal share of development responsibility with other organizations)

X Supporting developer organization (Contributed less than 50% of development responsibility)

Provide details of development role: ANL participated in the Globus Online integration and management of federated security infrastructure. The Globus Online integration facilitates leveraging solutions for data transfer and management, thus focusing on climate-specific infrastructure and integration pieces. Management of certificates and trust roots is key to defining and binding the federation, since it enables interoperability for authentication, authorization, and data download. ANL also participated in the development of the installation procedures for the system grid infrastructure.

Address:	9700 South Cass Avenue
City:	Lemont
State:	Illinois
Zip/Postal:	60439
Country:	United States of America
Web URL:	http://www.anl.gov
Contact Name:	Ian Foster
Phone:	(630) 252-4619
Fax:	(630) 252-5986
Email:	foster@mcs.anl.gov

Organization Name: **Lawrence Berkeley National Laboratory (LBNL)**

What role did this organization play in development of technology?
__ Principal developer organization (All or majority of development responsibility)
__ Co-developer organization (Equal share of development responsibility with other organizations)
X Supporting developer organization (Contributed less than 50% of development responsibility)

Provide details of development role: LBNL developed the storage resource management, called the Berkeley Storage Manager (BeStMan), which provides the front-end for various disk-based systems and mass-storage systems such as the High-Performance Storage System (HPSS). BeStMan makes it possible for ESGF sites to access all remote deep storage data sources through a uniform interface. LBNL led the research, development, and testing of the end-to-end capabilities of the Advanced Network Initiative (ANI) test bed to ensure that the software, services, and applications associated with massive climate data sets can scale to the anticipated 100-Gbps network infrastructure.

Address:	1 Cyclotron Road
City:	Berkeley
State:	California
Zip/Postal:	94720
Country:	United States of America

Web URL:	http://www.lbl.gov
Contact Name:	Alex Sim
Phone:	(510) 495-2290
Fax:	(510) 486-4004
Email:	ASim@lbl.gov

Organization Name: **German Climate Computing Center (DKRZ)**

What role did this organization play in development of technology?
__ Principal developer organization (All or majority of development responsibility)
__ Co-developer organization (Equal share of development responsibility with other organizations)
X Supporting developer organization (Contributed less than 50% of development responsibility)

Provide details of development role: DKRZ assisted in the integrated system work for both old and new WGet download scripts generated from ESGF portals. DKRZ also assisted in part of the data replication effort, for expanded use and testing of the ESGF replication client, and the quality control effort to verify the correctness of the data submitted to the federated archive.

Address:	Bundesstraße 45a
City:	Hamburg
State:	
Zip/Postal:	D-20146
Country:	Germany
Web URL:	http://www.dkrz.de/
Contact Name:	Michael Lautenschlager
Phone:	+4940-460094-118
Fax:	
Email:	lautenschlager@dkrz.de

Organization Name: **Euro-Mediterranean Center on Climate Change (CMCC)**

What role did this organization play in development of technology?
__ Principal developer organization (All or majority of development responsibility)
__ Co-developer organization (Equal share of development responsibility with other organizations)
X Supporting developer organization (Contributed less than 50% of development responsibility)

Provide details of development role: CMCC developed a unique web-based dashboard capability showing the worldwide system metrics of users, downloads, federated resources, and published data. The dashboard allows interactive monitoring of the health and state of the federation, as well real-time service availability.

Address:	Via Augusto Imperatore 16
City:	Lecce
State:	Lecce
Zip/Postal:	73100
Country:	Italy
Web URL:	http://www.cmcc.it/
Contact Name:	Sandro Fiore
Phone:	+39 328 7525119
Fax:	
Email:	sandro.fiore@unisalento.it

Organization Name: **British Atmospheric Data Centre (BADC)**

What role did this organization play in development of technology?
__ Principal developer organization (All or majority of development responsibility)
__ Co-developer organization (Equal share of development responsibility with other organizations)
X Supporting developer organization (Contributed less than 50% of development responsibility)

Provide details of development role: BADC led the development of the model metadata schema and related infrastructure to gather and access this information from a centralized server. BADC also assisted in the development of the ESGF helpdesk and the ESGF replication client, and collaborated in the definition of the security APIs and data organization structure.

Address:	Didcot OX11 0QX
City:	Harwell Oxford
State:	
Zip/Postal:	
Country:	United Kingdom
Web URL:	
Contact Name:	Stephen Pascoe
Phone:	+44 (0) 1235 445980
Fax:	
Email:	stephen.pascoe@stfc.ac.uk

Organization Name: **Goddard Space Flight Center (GSFC)**

What role did this organization play in development of technology?
__ Principal developer organization (All or majority of development responsibility)
__ Co-developer organization (Equal share of development responsibility with other organizations)
X Supporting developer organization (Contributed less than 50% of development responsibility)

Provide details of development role: GSFC helped in the development of the ESGF installation script and WGET script.

Address:	8800 Greenbelt Road
City:	Greenbelt
State:	Maryland
Zip/Postal:	20771
Country:	United States of America
Web URL:	http://www.nasa.gov/centers/goddard/
Contact Name:	Zed Pobre
Phone:	(301) 286-7542
Fax:	
Email:	zed.pobre@nasa.gov

Organization Name: Institut Pierre Simon Laplace (IPSL)

What role did this organization play in development of technology?
__ Principal developer organization (All or majority of development responsibility)
__ Co-developer organization (Equal share of development responsibility with other organizations)
X Supporting developer organization (Contributed less than 50% of development responsibility)

Provide details of development role: IPSL led the development of the Common Information Model (CIM) viewer, a plugin client-side tool that integrates with the ESGF web portal and allows retrieval and display of complex metadata describing the scientific climate modeling processes.

Address:	4 Place Jussieu
City:	Paris
State:	Tour 45-44, Salle #207
Zip/Postal:	75005
Country:	France
Web URL:	http://www.ipsl.fr/
Contact Name:	Mark Morgan
Phone:	+33 (0) 1 44 27 49 10
Fax:	
Email:	momipsl@ipsl.jussieu.fr

C. Principal/primary investigator, developer, inventor, or team leader
List the individual(s) who are designated as principal/primary investigator, developer, inventor, or team leader for this entry. These individuals will receive special recognition if the entry wins. Copy the template to add additional individuals. List all team members in Appendix A.

Principal Investigator Name:	Dean N. Williams
Title:	Principal Investigator
Organization:	Lawrence Livermore National Laboratory
Address:	7000 East Avenue, L-103
City/State:	Livermore, California
Zip/Postal:	94551
Country:	United States of America
Phone:	925 423-0145
Fax:	925 422-7675
Email:	williams13@llnl.gov

Principal Developer Name:	Gavin Bell
Title:	Team Lead
Organization:	Lawrence Livermore National Laboratory
Address:	7000 East Avenue, L-103
City/State:	Livermore, California
Zip/Postal:	94551
Country:	United States of America
Phone:	925-422-7436
Fax:	925 422-7675
Email:	gavin@llnl.gov

Principal Developer Name:	Luca Cinquini
Title:	Team Lead
Organization:	Jet Propulsion Laboratory
Address:	4800 Oak Grove Drive
City/State:	Pasadena, California
Zip/Postal:	91109
Country:	United States of America
Phone:	818 309-0680
Fax:	818 393-6141
Email:	Luca.Cinquini@jpl.nasa.gov

Principal Developer Name:	John Harney
Title:	Team Lead
Organization:	Oak Ridge National Laboratory
Address:	1 Bethel Valley Road
City/State:	Oak Ridge, Tennessee
Zip/Postal:	37831
Country:	United States of America
Phone:	865 576-1162
Fax:	
Email:	harneyjf@ornl.gov

2. PRODUCT INFORMATION

A. Product brand name: Earth System Grid Federation—www.esgf.org

B. Generic description of product (spectrometer, battery, chemical, etc.)
An open information and knowledge infrastructure for peer-to-peer access to internationally distributed geospatial data for climate research.

C. Product photo

3. EXECUTIVE SUMMARY

The Earth System Grid Federation (ESGF) is a successful international collaboration that manages the first-ever decentralized database for handling climate science data, with multiple petabytes of data at dozens of federated sites worldwide. ESGF is recognized as the leading infrastructure for the management and access of large distributed data volumes for climate change research. ESGF supports the Coupled Model Intercomparison Project (CMIP), whose protocols enable the periodic assessments carried out by the Intergovernmental Panel on Climate Change (IPCC). Using a system of geographically distributed peer nodes—independently administered yet united by common protocols and interfaces—the ESGF community holds the premier collection of simulations and observational and reanalysis data for climate change research. ESGF is in the early stages of being adapted for use in two additional domains: biology (to accelerate drug design and development) and energy (infrastructure for California Energy Systems for the 21st Century (CES21)).

4. INTRODUCTION DATE

The infrastructure allowing direct peer-to-peer interaction became fully available for the International Panel on Climate Change Assessment Report community in 2012. With this infrastructure, scientists, resource managers, policymakers, and other users can obtain—through federated portals—disparate distributed climate data from a multitude of worldwide climate research projects. Work to establish the system of geographically distributed peer nodes began in 2010, and the transition from the use of client servers to the peer node system was complete in 2012. Modeling and data analysis to support the assessment by the 2012 IPCC was performed using this most recent iteration of the Earth System Grid Federation.

Available documentation:

1. IEEE special issue, *Future Generation Computing Systems*:
   Luca Cinquini, Daniel Crichton, Chris Mattmann, Gavin M. Bell, Bob Drach, Dean Williams, John Harney, Galen Shipman, Feiyi Wang, Philip Kershaw, Stephen Pascoe, Rachana Ananthakrishnan, Neill Miller, Estanislao Gonzalez, Sebastian Denvil, Mark Morgan, Sandro Fiore, Zed Pobre, Roland Schweitzer, "The Earth System Grid Federation: An Open Infrastructure for Access to Distributed Geospatial Data," accepted, due out in 2013.

2. IEEE special issue, *Cutting-Edge Research in Visualization*:
   Dean N. Williams, Time Bremer, Charles Doutriaux, John Patchett, Galen Shipman, Blake Haugen, Ross Miller, Brian Smith, Chad Steed, E. Wes Bethel, Hank Childs, Harinarayan Krishnan, Michael Wehner, Claudio T. Silva, Emanuele Santos, David Koop, Tommy Ellqvist, Huy T. Vo, Jorge Poco, Berk Geveci, Aashish Chaudhary, Andrew Bauer, Alexander Pletzer, Dave Kindig, Gerald L. Potter, Thomas P. Maxwell, "The Ultra-scale Visualization Climate Data Analysis Tools: Data Analysis and Visualization for Geoscience Data," accepted, due out in 2013.

5. PREVIOUS R&D 100 ENTRY

Has this product or an earlier version been entered in the R&D 100 awards competition previously? Did it win an R&D 100 Award? List product name and year.

No.

6. PRICE

There is no cost associated with using the open-source ESGF software. The Department of Energy and other U.S. agencies as well as several international research institutions bear the cost to develop and improve ESGF.

7. PATENTS

There are no patents associated with ESGF.

8. PRODUCT DESCRIPTION

A. What does it do?

As global, broad-based climate change projections have become more useful, effectively managing the vast accompanying volumes of data represents a major challenge for the computational scientists who support the projections. For climate change science, this "big data" challenge is being met with the Earth System Grid Federation (ESGF), an international collaboration led by Lawrence Livermore National Laboratory.

ESGF was established to support the Coupled Model Intercomparison Project (CMIP), which is the worldwide-standard experimental protocol for studying the output of coupled atmosphere–ocean general circulation models. Established in 1995 by the World Climate Research Programme's Working Group on Coupled Modeling, ESGF provides a community-based infrastructure in support of climate model diagnosis, validation, intercomparison, documentation, and data access. Through ESGF, scientists are able to analyze general circulation models in a systematic fashion, a process that serves to facilitate model improvement. Virtually the entire international climate modeling community has participated in CMIP since its inception. CMIP simulation model runs are key components for assessments of the International Panel on Climate Change (IPCC). The IPCC, including thousands of scientists whose research supports IPCC assessments, was named co-winner of the 2007 Nobel Peace Prize.

The internationally distributed, peer-to-peer ESGF "data cloud" archive represents the culmination of an effort that began in the late 1980s to develop efficient, community-based tools to obtain relevant meteorological and other observational data, develop custom computational models, and export analysis tools for climate-change simulations, such as those used in IPCC. ESGF portals are gateways to scientific data collections hosted at sites around the globe that allow the user to register and potentially access the entire ESGF network of data and services. (See Figure 1.) Currently more than 25 portals are in use, including Livermore's Program for Climate Model Diagnosis and Intercomparison (PCMDI).

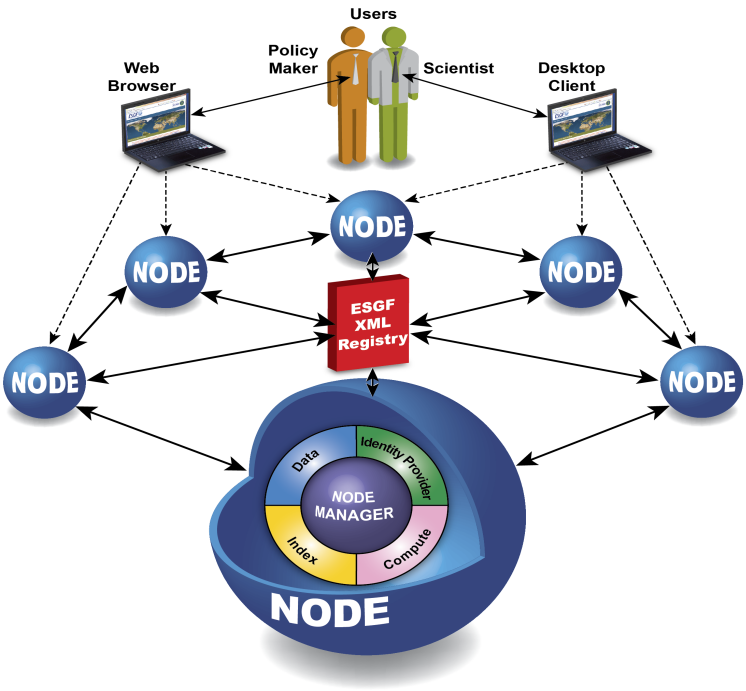


Figure 1. *The Earth System Grid Federation's peer-to-peer architecture is based on the principles of modular components and standard protocols. Each system node, or software stack (shown in blue), can be configured to posses one or more "software types," each entailing a specific functionality: "Data" publishes data and serves data through a variety of protocols; "Index" harvests metadata and enables data discovery; "Identity Provider" registers, authenticates, and authorizes users; and "Compute" handles computational resources for data reduction, analytics, and visualization. All federated nodes interact as equals, so no single points of failure arise.*

B. Describe the principal applications of this product.

ESGF allows international climate research teams to work in highly distributed research environments, using unique scientific instruments, exascale-class computers, and extreme amounts of data. (See Figure 2.) A key to ESGF's success is its ability to effectively produce, validate, and analyze research results collaboratively, so that, for example, new results generated by one team member are immediately accessible to the rest of the team, who can annotate, comment on, and otherwise interact around those results. ESGF supports the latest incarnation of CMIP, which is CMIP phase 5 (or CMIP5). CMIP5 simulation models runs are essential for the upcoming 2013 IPCC Fifth Assessment Report.

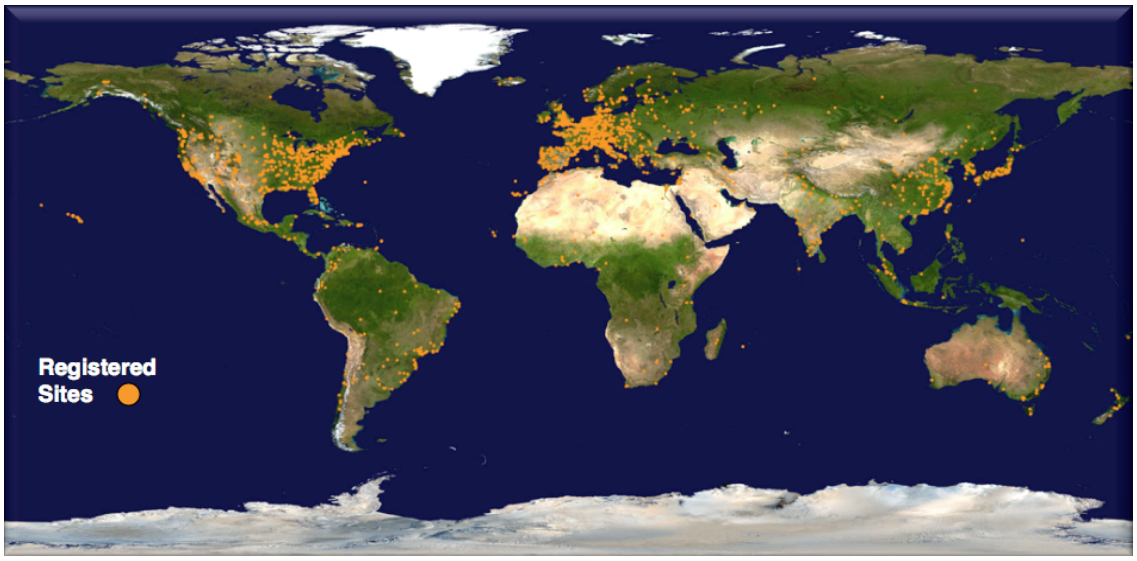


Figure 2. *Climate scientists anywhere in the world can access the ESGF research centers and climate data products. Over 25,000 registered sites.*

C. How will this product benefit the market that it serves?
ESGF has been benefiting the climate research community since its inception. Users access all data as if they are working on one centralized archive system. To date, ESGF has made available over 60 large-scale CMIP5 simulation model runs (more than 1.8 petabytes of data) from 25 worldwide climate research centers spanning 21 countries. ESGF also provides access to 18 highly visible national and international climate data products, with more on the way.

Because CMIP5 is a high-profile community effort, other projects have been attracted by the software infrastructure that enables data intercomparison and validation. By participating in CMIP5, these projects and their data become instantly accessible to tens of thousands of users. Virtually all climate researchers now use ESGF.

D. List all other applications for which your product can now be used.
Rapid advances in experimental, sensor, and computational technologies are driving exponential growth in the volume, acquisition rate, variety, and complexity of scientific data across all scientific domains. This new wealth of scientifically meaningful data has tremendous potential for scientific discovery. However, to achieve breakthroughs, these data must be exploitable—they must be analyzed effectively and efficiently, and the results shared and communicated easily with the wider community. The explosion in data complexity and scale makes these tasks exceedingly difficult to achieve, particularly given that an increasing number of disciplines are working across techniques, integrating simulation and experimental or observational results. We need new approaches to data management, analysis, and visualization that provide research teams with easy-to-use, end-to-end solutions. These solutions must facilitate—and where feasible, automate—every stage in the data lifecycle.

ESGF is in the early stages of being adapted for use in two domains: biology (to accelerate drug design and development) and energy (infrastructure for California Energy Systems for the 21st Century (CES21)). Required core functionalities are the same among science communities, but customization is necessary to adapt to specific needs and fit into research and analysis workflows. For biology and CES21, ESGF's large-scale data management infrastructure will be augmented with specific community tools for data mining, analysis, and visualization. ESGF's architecture can also be leveraged for accessing data from numerous other scientific domains, such as astrophysics, chemistry, combustion, fusion, material science, and nuclear energy.

9. TECHNOLOGY DESCRIPTION

A. How does the product or technology operate?

The ESGF peer-to-peer architecture is based on a dynamic system of nodes that interact on an equal basis and offer a broad range of user and data services, depending on how each is set up. This extensible and scalable system supports geospatial and temporal searches and includes a dashboard that shows system metrics and a rich set of climate analysis tools to help manipulate the data. For example, the Ultra-scale Visualization Climate Data Analysis Tools (UV-CDAT) have workflow scripts that automate scientific analysis and visualization, making it easy for users to re-run analyses and to work together, which encourages collaboration and openness in scientific enquiry.

A typical use case scenario is that of a climate scientist who wishes to analyze the predicted change in some physical field (for example, sea-surface temperature) for the next 20 years under different emission scenarios. (See Figure 3.) The scientist can start her workflow at any of the ESGF portals and use the search interface to select monthly average data sets from the CMIP5 project that contains the field in question. Because of the distributed nature of the system, the search results will contain data sets stored across four continents and more than 25 data nodes. To reduce the number of results returned, the scientist can additionally select a specific model (e.g., "GFDL-ESM2M") and/or a specific emission scenario (e.g., "1% CO_2 increase").

Each data set result will contain a link to "compute," which the scientist can click to visualize the field over the surface of the Earth or create a line plot versus time, latitude, or longitude. Another link will display the detailed metadata describing both the model and the experiment that were used to generate the output. Once the scientist has decided that she has indeed located data relevant to her research, she may either download the files from the browser interface (if only a few) or click on the "wget" link to generate a script that will run on her desktop to download all the files at once. At this point, she will be able to use an analysis package of choice (e.g., UV-CDAT, IDL, MatLab, etc.) to take a more in-depth look at the data.

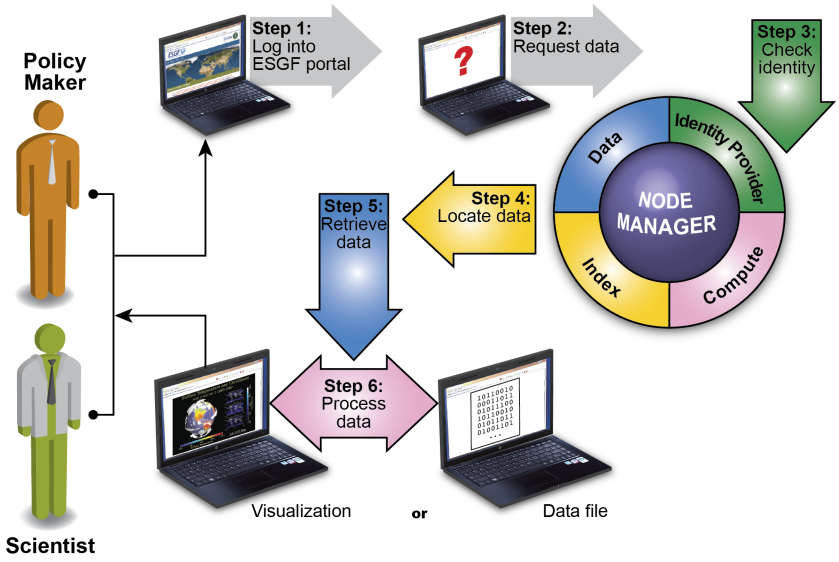


Figure 3. *Different types of users can easily access ESGF data sets, enabling the exploration of data by different means and evaluation from different viewpoints.*

B. What scientific theories, if any, are involved?

We are using a novel peer-to-peer implementation to provide informatics and analytics capabilities to ESGF. (See Figure 4.) The system is designed to be resilient to membership changes over the set of participating federated nodes. The system uses a push, cascading, peer-to-peer network model as its gossip[1] protocol such that all nodes maintain a consistent view of all other federated node members. The algorithm provably converges over time and the implementation of the gossip protocol minimizes the occurrences of disconnected cliques and accelerates the discovery process. ESGF also takes advantage of secured network transport protocols to create secured node group and sub-group partitions. The peer-to-peer protocol underpins the ESGF system by providing a resilient substrate from which higher-level services can leverage for coordination. In addition to offering resilience, the peer-to-peer protocol has the advantage of lowering the overhead of configuration and maintenance for node administrators.

[1] A style of computer-to-computer communication protocol inspired by social networks.

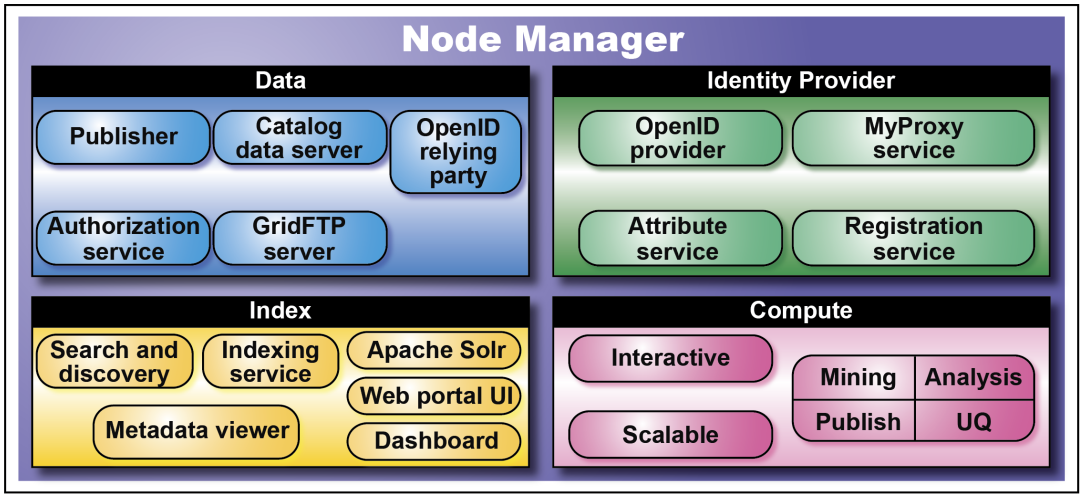


Figure 4. *ESGF delivers a comprehensive, end-to-end, and top-to-bottom environment for current petascale and emerging exascale science domains. The figure emphasizes data services at each level for the node architecture. The production of ESGF for climate products is evidence that a distributed dynamic federated system is flexible enough to support a wide range of heterogeneous data (i.e., simulation, observation, and re-analysis) and application tools. Current application of ESGF architecture to biology and energy science affirms this flexibility.*

ESGF employs a global security infrastructure that enables federation-wide services such as single sign-on and distributed access control, while maintaining local administrative control over data and computational resources. Local policy statements, which match classes of resources to groups of users bestowed with the necessary permissions, control access to data holdings and analysis algorithms. User authorization is encoded as digitally signed statements, which are respected by other nodes in the federation because of the establishment of mutual trust relationships. ESGF security is based on open standards and application programing interfaces, such as OpenID, X.509, SSL and SAML, to guarantee interoperability.

The ESGF data grid is arguably the world largest system for managing and accessing distributed scientific data (currently, around 2 petabytes and growing). Data and metadata are published, stored, and served from tens of nodes around the globe (controlled by different authorities), yet they are searchable and accessible as if they were stored in a single global archive. In order to achieve satisfactory performance in service to the global scientific community, metadata holdings at each site are partitioned into high-level, frequently queried discovery information and low-level inventory information. Queries issued to any node are automatically distributed to all other nodes in the federation, so that complete, up-to-date results are always returned to the user. Additionally, automatic and incremental metadata replication technologies create identical, synchronized local copies of remote metadata catalogs, to speed up query performance over large geographic distances.

C. What are the building blocks or your technology?

It was an important design criterion that this system be an open source system and as such takes advantage of open source tools. The system leverages web servers and a few key protocols to provide necessary communication transport. The codebase was written

to be easy to read and understand. Most of the system is written to take advantage of the power of the JVM (Java Virtual Machine), allowing ESGF to mitigate operating system idiosyncrasies and maximize portability. The system uses the Representational State Transfer (REST)[2] services where feasible, allowing for future expansion and interoperability. The system is not only the cornerstone platform for climate research and information sharing and collaboration but is also designed to be a growing research platform to further investigation into new computing paradigms, including smart, autonomic, self-managing, and self-repairing computing infrastructures.

10. PRODUCT COMPARISON

A. Competitors

The world is full of large-scale data management and retrieval enterprise systems. In the U.S., for climate science alone there are the Global Change Master Directory, Network-Object Mobile-Agent Dynamic System (NOMADS), Atmospheric Radiation Measurement Archive, Regional Climate Model Evaluation System, and the NASA Distributed Active Archive Centers, to name a few. None is a distributed data system, however, and none allows interoperability among disparate data sets (i.e., simulations, observations, and reanalysis data) for assessment study.

The Open Science Grid (OSG), which handles data from experiments at the Large Hadron Collider (LHC) at the European Organization for Nuclear Research (CERN), uses grid architecture similar to ESGF. Grid computing, in which a federation of computer resources from multiple administrative domains work toward a common goal, has been a great success for such projects as SETI@Home, BOINC, Einstein@Home, etc. These projects use the downtime of a vast network of desktop computers to crunch observational data. OSG's grid is far more sophisticated. Collision data from LHC experiments are anticipated to be produced at an unprecedented rate of tens of petabytes per year and analyzed by a grid-based computer network. In 2012, the LHC grid was the world's largest, comprising over 170 computing facilities in a worldwide network across 36 countries.

A distributed infrastructure that is similar to ESGF's is that used by the search engine Google. Google operates the largest data centers in the world and handles more data than any other single entity on the planet. However, Google's data centers each serve a specific geographic area. A data search in London will likely produce different results from an identical search in Rio de Janeiro, Cape Town, Sydney, or New York City.

B. Matrix

The Earth System Grid Federation is alone in combining grid-based computing with a distributed architecture and interoperability among disparate data sets. Its infrastructure keeps participating members sovereign while simultaneously linking them together. To achieve that, ESGF developers created a unique system of nodes that requires very little explicit coordination while still providing a robust "data space" for storage and computation.

[2] A style of software architecture for distributed systems using web services.

C. Describe how your product improves upon competitive products or technologies.
In comparing OSG and ESGF, OSG's focus is on distributed computing, while ESGF's is on federated catalogs and higher-level analysis and visualization. However, OSG is really not so much a competitor as a sister project, and OSG and ESGF are discussing an integrated path forward.

With regard to Google, their use cases are radically different from ESGF's science-based use cases. In addition, with virtually unlimited financial resources, Google maintains total administrative control over their data at search sites. In contrast, ESGF with limited resources, partners with the science communities to set data policies, giving individual data owners' complete control over their data for access and use.

Several features make ESGF unique with respect to both OSG and Google:

- ESGF enables complex scientific queries (keywords, scientific facets, geospatial and temporal constraints) over federated holdings, with excellent performance.
- ESGF has a flexible distributed architecture, where administrators can decide which nodes to replicate, and which nodes to query directly. In other words, local administrators maintain control over their resources and services.
- ESGF enables higher-level services such as analysis and visualization.
- ESGF is completely based on open source technologies.

ESGF's real distinction is that it is a concert of many technologies integrated such that the whole is truly greater than the sum of its parts—while enabling next-level climate insight and experimentation.

D. Describe limitations of your product. What criticisms would your competitors offer?
Because ESGF is not under the administrative control of a single organization, the availability of data and metadata throughout the system and network performance are subject to the downtimes and outages of multiple centers. It is true that ESGF relies mostly on voluntary work by the data node administrators, who are not explicitly funded to support the system. To help offset this challenge, ESGF allows the capability to replicate highly valuable resources and data products at multiple sites. In addition, ESGF is augmenting its technology to allow smart caching, dynamic data replication, and publication of data at multiple index sites, thus ensuring more confidence in the availability of data and metadata throughout. Because climate data is so valuable, the World Climate Research Programme's Working Group on Coupled Modeling is working to fund ESGF administrators, so centers can be held more accountable for outage of their data holdings.

A non-researcher might say that ESGF is only for the technically savvy—because ESGF was first and foremost designed for the climate scientist. To combat this criticism, ESGF is working with international private and non-private institutions and universities to build multiple user access tools and modes that are useful for policy makers and other non-researchers. The primary goal is to create mechanisms to make extremely valuable data

available to all, including enabling the commercial use of the data. Facilitating widespread accessibility of these data products can hugely augment the importance of ESGF data holdings. As a result, the broad community of researchers and non-researchers will be able to efficiently access the most popular data products and analysis for planning and better decision-making.

11. SUMMARY

Progress in understanding and predicting climate change requires advanced tools to securely store, manage, access, process, analyze, and visualize enormous and distributed data sets. Only then can climate researchers understand the effects of climate change across all scales and use this information to inform policy decisions.

With the advent of major international climate modeling intercomparisons, a need emerged within the climate-change research community to develop efficient, community-based tools to obtain relevant meteorological and other observational data, develop custom computational models, and export analysis tools for climate-change simulations. While many nascent efforts to fill these gaps appeared, they were not integrated and therefore did not benefit from collaborative development. Sharing huge data sets was difficult, and the lack of data standards prevented the merger of output data from different modeling groups.

Thus began one of the largest-ever collaborative data efforts in climate science, resulting in the Earth System Grid Federation (ESGF), which is now used to disseminate model and observational data for research assessed by the Intergovernmental Panel on Climate Change (IPCC). Today, ESGF is a community-driven activity led by Lawrence Livermore National Laboratory as an open-source, operational code-base with secure, petabyte-level data storage and dissemination of the resources essential for studying climate change on a global scale. ESGF is designed to remain robust even as data volumes grow exponentially. It is used by virtually all climate science researchers in the world.

More recently, the decentralized approach to ESGF has changed from a client-server model to a more robust peer-to-peer (P2P) approach already proven for distributing large amounts of data and information. ESGF comprises a system of geographically distributed peer nodes, independently administered yet united by common protocols and interfaces allowing access to global atmospheric, land, ocean, and sea-ice data generated by satellite and in-situ observations and complex computer simulations for use in national and international assessment reports. Scientists are accessing climate data more efficiently and robustly through newly developed user interfaces, distributed or local search protocols, federated security, server-side analysis tools, and other community standards—all for improving our understanding of climate change.

ESGF ensures equal access to large disparate data sets, which in the past would have been accessible across the climate science community only with great difficulty. The ESGF infrastructure enables scientists to evaluate models, understand their differences,

and explore the impacts of climate change through a common interface, regardless of the location of the data.

ESGF's architecture can easily be leveraged for accessing data from other scientific domains, such as astrophysics, biology, chemistry, combustion, energy, fusion, material science, and nuclear energy. ESGF is now in the early stages of being adapted for use in two domains: biology (to accelerate drug design and development) and energy (infrastructure for California Energy Systems for the 21st century). Satisfying the need for substantial investments in software and technologies across many data-driven domains is paramount as future computing platforms and archives expand and reach extraordinary speeds and capacity.

12. AFFIRMATION

By submitting this entry to *R&D Magazine* you affirm that all information submitted as a part of, or supplemental to, this entry is a fair and accurate representation of this product. You affirm that you have read the instructions and entry notes and agree to the rules specified in those sections.

APPENDIX A: DEVELOPMENT TEAM INFORMATION

Team Member Name:	Gavin Bell
Title:	Team Leader
Organization:	Lawrence Livermore National Laboratory
Address:	7000 East Avenue, L-103
City/State:	Livermore, California
Zip/Postal:	94551
Country:	United States of America
Phone:	925 422-7436
Fax:	925 422-7675
Email:	gavin@llnl.gov

Team Member Name:	Charles Doutriaux
Title:	Computer Scientist
Organization:	Lawrence Livermore National Laboratory
Address:	7000 East Avenue, L-103
City/State:	Livermore, California
Zip/Postal:	94551
Country:	United States of America
Phone:	925 422-1487
Fax:	925 422-7675
Email:	doutriaux1@llnl.gov

Team Member Name:	Robert Drach
Title:	Computer Scientist
Organization:	Lawrence Livermore National Laboratory
Address:	7000 East Avenue, L-103
City/State:	Livermore, California
Zip/Postal:	94551
Country:	United States of America
Phone:	925 422-8303
Fax:	925 422-7675
Email:	drach1@llnl.gov

Team Member Name:	Matthew Harris
Title:	Computer Scientist
Organization:	Lawrence Livermore National Laboratory
Address:	7000 East Avenue, L-103
City/State:	Livermore, California
Zip/Postal:	94551
Country:	United States of America
Phone:	925 423-8978
Fax:	925 422-7675
Email:	harris112@llnl.gov

Team Member Name: Michael Ganzberger
Title: Computer Scientist
Organization: Lawrence Livermore National Laboratory
Address: 7000 East Avenue, L-103
City/State: Livermore, California
Zip/Postal: 94551
Country: United States of America
Phone: 925 423-4513
Fax: 925 422-7675
Email: ganzberger1@llnl.gov

Team Member Name: Renata McCoy
Title: Computer Scientist
Organization: Lawrence Livermore National Laboratory
Address: 7000 East Avenue, L-103
City/State: Livermore, California
Zip/Postal: 94551
Country: United States of America
Phone: 925 424-5237
Fax: 925 422-7675
Email: painter1@llnl.gov

Team Member Name: Jeffrey Painter
Title: Computer Scientist
Organization: Lawrence Livermore National Laboratory
Address: 7000 East Avenue, L-103
City/State: Livermore, California
Zip/Postal: 94551
Country: United States of America
Phone: 925 422-8114
Fax: 925 422-7675
Email: painter1@llnl.gov

Team Member Name: Dean N. Williams
Title: Principal Investigator
Organization: Lawrence Livermore National Laboratory
Address: 7000 East Avenue, L-103
City/State: Livermore, California
Zip/Postal: 94551
Country: United States of America
Phone: 925 423-0145
Fax: 925 422-7675
Email: williams13@llnl.gov

Team Member Name: Luca Cinquini
Title: Team Leader

Organization: Jet Propulsion Laboratory
Address: 4800 Oak Grove Drive
City/State: Pasadena, California
Zip/Postal: 91109
Country: United States of America
Phone: 818 309-0680
Fax: 818 393-6141
Email: Luca.Cinquini@jpl.nasa.gov

Team Member Name: Galen Shipman
Title: Site Principle Investigator
Organization: Oak Ridge National Laboratory
Address: 1 Bethel Valley Road
City/State: Oak Ridge, Tennessee
Zip/Postal: 37831
Country: United States of America
Phone: 865 576-1162
Fax:
Email: gshipman@ornl.gov

Team Member Name: John Harney
Title: Team Leader
Organization: Oak Ridge National Laboratory
Address: 1 Bethel Valley Road
City/State: Oak Ridge, Tennessee
Zip/Postal: 37831
Country: United States of America
Phone: 865 576-1162
Fax:
Email: harneyjf@ornl.gov

Team Member Name: Roland Schweitzer
Title: Computer Scientist
Organization: Pacific Marine Environment Laboratory
Address: 7600 Sand Point Way NE
City/State: Seattle, Washington
Zip/Postal: 98115
Country: United States of America
Phone: (209) 526-5080
Fax: (209) 526-6239
Email: roland.schweitzer@noaa.gov

Team Member Name: Ian Foster
Title: Site Principle Investigator
Organization: Argonne National Laboratory
Address: 9700 S. Cass Avenue

City/State: Argonne, Illinois
Zip/Postal: 60439
Country: United States of America
Phone: (630) 252-2020
Fax:
Email: foster@anl.gov

Team Member Name: Rachana Ananthakrishnan
Title: Computer Scientist
Organization: Argonne National Laboratory
Address: 9700 S. Cass Avenue
City/State: Argonne, Illinois
Zip/Postal: 60439
Country: United States of America
Phone: (630) 252-2020
Fax:
Email: ranantha@mcs.anl.gov

Team Member Name: Neill Miller
Title: Computer Scientist
Organization: Argonne National Laboratory
Address: 9700 S. Cass Avenue
City/State: Argonne, Illinois
Zip/Postal: 60439
Country: United States of America
Phone: (630) 252-2020
Fax:
Email: neillm@mcs.anl.gov

Team Member Name: Alex Sim
Title: Site Principle Investigator
Organization: Lawrence Berkeley National Laboratory
Address: 1 Cyclotron Road
City/State: Berkeley, California
Zip/Postal: 94720
Country: United States of America
Phone: (510) 495-2290
Fax: (510) 486-4004
Email: ASim@lbl.gov

Team Member Name: Estanislao Gonzalez
Title: Computer Scientist
Organization: German Climate Computing Center
Address: Bundesstraße 45a
City/State: Hamburg
Zip/Postal: D-20146

Country: Germany
Phone: +4940-460094-118
Fax:
Email: gonzalez@dkrz.de

Team Member Name: Sandro Fiore
Title: Computer Scientist
Organization: Euro-Mediterranean Center on Climate Change
Address: Via Augusto Imperatore 16
City/State: Lecce
Zip/Postal: 73100
Country: Italy
Phone: +39 328 7525119
Fax:
Email: sandro.fiore@unisalento.it

Team Member Name: Stephen Pascoe
Title: Computer Scientist
Organization: STFC Rutherford Appleton Laboratory
Address: Didcot OX11 0QX
City/State: Harwell Oxford
Zip/Postal:
Country: United Kingdom
Phone: +44 (0) 1235 445980
Fax:
Email: stephen.pascoe@stfc.ac.uk

Team Member Name: Philip Kershaw
Title: Computer Scientist
Organization: STFC Rutherford Appleton Laboratory
Address: Didcot OX11 0QX
City/State: Harwell Oxford
Zip/Postal:
Country: United Kingdom
Phone: +44 (0) 1235 445980
Fax:
Email: philip.kershaw@stfc.ac.uk

Team Member Name: Zed Pobre
Title: Computer Scientist
Organization: Goddard Space Flight Center
Address: 8800 Greenbelt Road
City/State: Greenbelt Maryland
Zip/Postal: 20771
Country: United States of America
Phone: (301) 286-7542

Fax:
Email: zed.pobre@nasa.gov

Team Member Name: Sébastien Denvil
Title: Computer Scientist
Organization: Institut Pierre Simon Laplace
Address: UPMC, Case 101, 4 place Jussieu
City/State: Paris
Zip/Postal: 75252
Country: France
Phone: 33 1 44 27 21 10
Fax: 33 1 44 27 39 02
Email: sebastien.denvil@ipsl.jussieu.fr

Team Member Name: Mark Mogan
Title: Computer Scientist
Organization: Institut Pierre Simon Laplace
Address: 4 Place Jussieu
City/State: Paris, Tour 45-44, Salle #207
Zip/Postal: 75005
Country: France
Phone: +33 (0) 1 44 27 49 10
Fax:
Email: momipsl@ipsl.jussieu.fr

APPENDIX B: MAREKTING AND MEDIA INFORMATION

1. MARKETING

Contact Person to Handle All Arrangements

Name:	Connie Pitcock, Marketing and Business Development Associate
Organization:	Lawrence Livermore National Laboratory
Address:	7000 East Ave, L-795
City:	Livermore
State:	CA
ZIP Code:	94551
Country:	USA
Phone:	925-422-1072
Fax:	925-423-8988
Email:	pitcock1@llnl.gov

2. MEDIA

Contact Person for Media and Editorial Inquiries

Name:	Connie Pitcock, Marketing and Business Development Associate
Organization:	Lawrence Livermore National Laboratory
Address:	7000 East Ave, L-795
City:	Livermore
State:	CA
ZIP Code:	94551
Country:	USA
Phone:	925-422-1072
Fax:	925-423-8988
Email:	pitcock1@llnl.gov

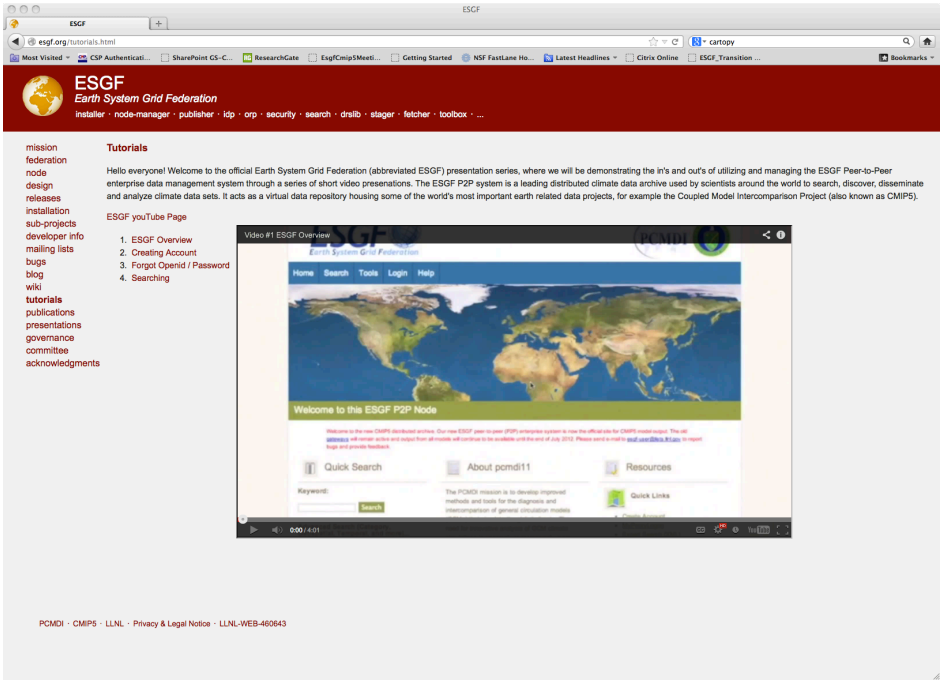
APPENDIX C: SUPPORTING MATERIALS

Publications

Cinquini, L.,et al (2012) "The Earth System Grid Federation: An open infrastructure for access to distributed geospatial data," *E-Science, 2012 8th International Conference*, 1–10.
http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6404471&tag=1

Williams, D.N., et al (2009) "The Earth System Grid: Enabling Access to Multi-Model Climate Simulation Data," *Bulletin of the American Meteorological Society*, January. http://ams.allenpress.com/perlserv/?request=get-abstract&doi=10.1175/2008BAMS2459.1

Video Link



http://esgf.org/tutorials.html

- ESGF Overview Video
- Creating an Account
- Forgot OpenID / Password
- Searching

APPENDIX D: LETTERS OF RECOMMENDATION

Department of Energy
Office of Science
Washington, DC 20585

April 12, 2013

R&D 100 Awards
100 Enterprise Drive
Suite 600, Box 912
Rockaway, NJ 07866-0912

RE: Letter of Reference for ESGF

Dear R&D 100 Awards Committee:

We are happy to endorse the outstanding research and engineering capabilities of the Earth System Grid Federation (ESGF). The ESGF is internationally recognized as a relatively new and highly significant research and engineering capability, that pioneered developments in federated and distributed data infrastructure, data management, and analysis and visualization tools. ESGF has extensively served the climate community at large for more than a decade. Data disseminated through the ESGF has undoubtedly allowed a more rapid advance of the understanding of climate models and has created a paradigm shift in how the developers of climate models view ownership of the model output.

The Department of Energy (DOE) has supported the Earth System Grid since its inception in the late 1990s. Since then, the Earth System Grid provided the community data dissemination, access, and discovery tools for the exploration of climate research through model intercomparison projects. Fueled by the Coupled Model Intercomparison Project (CMIP)—which in turn supplies data to the Intergovernmental Panel on Climate Change (IPCC) periodic assessment reports—the Earth System Grid expanded to a federated system. This ESGF team devised a revolutionary technology for housing, managing, searching, and disseminating data on a scale that is critical to climate science knowledge management and discovery.

Because the IPCC CMIP simulation archives are high profile community efforts, other projects within the climate and earth system community have become attracted by ESGF software infrastructure. As such, ESGF provides access to dozens of highly visible national and international climate and earth system data products. The ESGF distributed archive now includes model simulation output, observational, and reanalysis data sets from major U.S. agencies (e.g., Department of Energy (DOE), National Aeronautics and Space Administration (NASA), National Oceanic and Atmospheric Administration (NOAA), and National Science Foundation (NSF)) and major international agencies (InfraStructure for the European Network for the Earth System Modeling (IS-ENES), Australian National Computational Infrastructure (NCI), and the University of Tokyo).

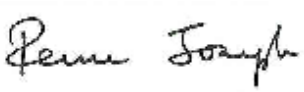
By leveraging current and evolving technologies to manage distributed climate and earth system data in a unified virtual environment, the ESGF project is promoting data sharing between international research centers and diverse users. In transforming these data into a collaborative community resource, ESGF is changing the way global climate research is conducted. Through unifying team efforts, the ESGF infrastructure enables scientists to evaluate models, understand their differences, and explore the impacts of climate change through a common interface, regardless of the location of the data. The total ESGF federated archive currently serves tens of thousands of users, and it contains multiple petabytes of disparate data sets that have enabled publication of more than one thousand scientific journal articles.

Like many others in the field, we believe that ESGF is a major asset in support of the climate change community at large—and deserving of the R&D 100 honor. Therefore, it is with great pleasure and enthusiasm that we write this letter of support for the ESGF technical team.

Sincerely,

Dr. Gerald Geernaert
DOE CESD Division Director

Dr. Wanda Ferrell
DOE CESD Program Manager



Dr. Renu Joseph
DOE CESD Program Manger

National Center for Atmospheric Research
Climate and Global Dynamics Division • Climate Change Research Section
1850 Table Mesa Drive • Boulder, CO 80305
Tel: 303-497-1321 • Fax: 303-497-1715 • Email: wmw@ucar.edu
Web: http://www.cgd.ucar.edu/ccr/warren/

20 March 2013

R&D 100 Awards
100 Enterprise Drive
Suite 600, Box 912
Rockaway, NJ 07866-0912

Dear R&D 100 Awards:

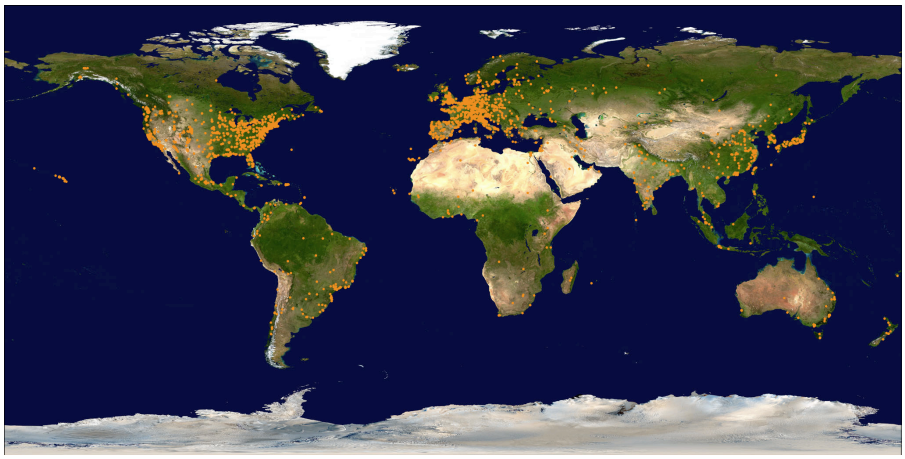
We are writing to support the selection of the Earth System Grid Federation (ESGF) as an R&D 100 awardee for 2013.

The impact of the ESGF on the advancement of climate science research and our understanding of climate change cannot be overstated. Our best estimates of future climate originate from the model projections performed by the world's leading modeling centers. Because no single model is demonstrably superior to all others, a result of both the extreme complexity of the climate system and scientific uncertainties in model formulations, the ability to examine the results of many models has proved to be crucial in improving of scientific understanding of future climate change.

Model simulations from the fifth phase of the Coupled Model Intercomparison Project (CMIP5) will underpin a major part of the upcoming Fifth Scientific Assessment by the Intergovernmental Panel on Climate Change. Through ESGF, thousands of researchers can access the CMIP5 database, which now exceeds 50 Petabytes (and will ultimately reach 100 Petabytes) distributed over the modeling center archives on 4 continents. The ability of thousands of researchers to easily access, search, discover and obtain the data from a large, multi-model archive of climate simulations is truly transformational. The ESGF system provides secure access to a scientific gold mine of information.

In the future, the ESGF system will enable major advances in our understanding of the past and future climate change. It will also allow the scientific community to provide the more detailed and more reliable forecasts of climate change that governments and society require for developing successful adaption and mitigation strategies.

To highlight the success of the ESGF, we show a world map of the current distribution of users.



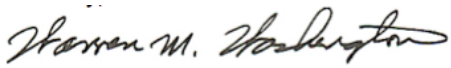
Climate scientists anywhere in the world can access the ESGF research centers and climate data products. Each dot represents a registered site, which total over 25,000.

The National Center for Atmospheric Research is operated by the University Corporation for Atmospheric Research under sponsorship of the National Science Foundation.

An Equal Opportunity/Affirmative Action Employer

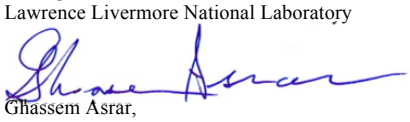
Because of these impacts, we firmly believe that the ESGF developers are deserving of an R&D 100 award.

Sincerely,

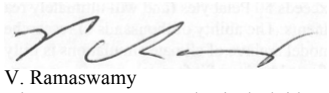


Warren Washington,
Member, National Academy of Engineering,
Former Chair, National Science Board
Senior Scientist, National Center for Atmospheric Science

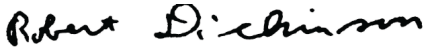
Benjamin Santer,
Member, National Academy of Science
Distinguished Member of the Technical Staff
Lawrence Livermore National Laboratory



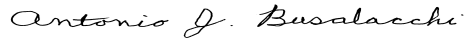
Ghassem Asrar,
Executive Director,
World Climate Research Programme



V. Ramaswamy
Director, NOAA Geophysical Fluid Dynamics Laboratory



Robert Dickinson,
Member, National Academies of Science and Engineering
Professor of Geological Sciences, University of Texas



Antonio J. Busalacchi
Professor and Director, U. of MD Earth System Science Interdisciplinary Center
Chairman of the Joint Scientific Committee for the World Climate Research Programme
Chairman of the National Research Council Board on Atmospheric Sciences and Climate

NATIONAL SCIENCE FOUNDATION

4201 Wilson Boulevard
Arlington, VA 22230

DIVISION OF ATMOSPHERIC AND GEOSPACE SCIENCES

To Whosever it may concern **March 22, 2013**

I was a program director at DOE Office of Science, Office of Biological and Environmental Research from 2004-2010. During that period, with the Office of Advanced Scientific Computing and Research, our office co-funded a project called the Earth System Grid Center for Enabling Technologies (ESG CET). The lead PI for this project was Dean Williams of Program for Climate Model Diagnosis and Intercomparison (PCMDI) located at Lawrence Livermore National Laboratory.

The goal of this project was to build a **"science gateway"** to climate resources that provides data, information, models, analysis, and visualization tools and computational capabilities for management and analysis. ESG-CET was a multi-institutional effort across various DOE national laboratories, National Center for Atmospheric Research, NOAA/Pacific Marine Environmental Laboratory), and University of Southern California, working to maximize the accessibility of climate simulation data by the international research community.

ESG was responsible for the data management of hundreds of terabytes of climate data ranging from high-resolution modeling, grand challenge-scale computations on leadership computing systems, regional climate modeling, coupled climate/carbon cycle modeling, land/biosphere modeling, atmospheric chemistry modeling, detection and attribution of climate change, and model intercomparison projects.[1,2]

The ESG-CET subsequently led to Earth System Grid Federation (ESGF[3]), a project being led by Dean Williams. To date, ESGF has made available over 60 large-scale CMIP5 simulation model runs (more than 1.8 petabytes of data) from 25 worldwide climate research centers spanning 21 countries. ESGF also provides access to 18 highly visible national and international climate data products, with more on the way. Climate scientists anywhere in the world can access the ESGF research centers and climate data products. There are over 25,000 registered sites, with currently more than 25 portals are in use, including one at PCMDI, and over 2.5 Petabytes downloaded by the community.

At NSF, in FY 11, we made ~16 small grants to university researchers for analysis of the Coupled Model Intercomparison Project (CMIP) simulations. They all accessed the data at PCMDI through the use of ESGF. In March 2012, a workshop

on *Coupled Model Intercomparison Project Phase 5 (CMIP5) Model Analysis* was held where PIs presented their results. The website for the Workshop is http://cmip5.wcrp-climate.org/workshop/sponsors.shtml A list of publications resulting from the CMIP project may be found at http://cmip.llnl.gov/cmip5/publications/allpublications

Anjuli Bamzai, Ph.D.

Program Director Climate and Large-Scale Dynamics

NSF Division of Atmospheric and Geospace Sciences

Directorate of Geosciences

National Science Foundation

References:

1. Data Management and Analysis for the Earth System Grid
Overview of climate data management challenges and server-side analysis in ESG. D. N. Williams, R. Ananthakrishnan, D. E. Bernholdt, S. Bharathi, D. Brown, M. Chen, A. L. Chervenak, L. Cinquini, R. Drach, I. T. Foster, P. Fox, S. Hankin, V. E. Henson, P. Jones, D. E. Middleton, J. Schwidder, R. Schweitzer, R. Schuler, A Shoshani, F. Siebenlist, A. Sim, W. G. Strand, N. Wilhelmi, M. Su, "Data Management and Analysis for the Earth System Grid", in the Journal of Physics: Conference Series, SciDAC '08 conference proceedings, volume 125. ([Citation:] D N Williams et al 2008 J. Phys.: Conf. Ser. 125 012072)

2. The Earth System Grid: Enabling Access to Multi-Model Climate Simulation Data . D N Williams, R Ananthakrishnan, D E Bernholdt, S Bharathi, D Brown, M Chen, A L Chervenak, L Cinquini, R Drach, I T Foster, P Fox, D Fraser, J Garcia, S Hankin, P Jones, D E Middleton, J Schwidder, R Schweitzer, R Schuler, A Shoshani, F Siebenlist, A Sim, W G Strand, M Su, N. Wilhelmi. Bulletin of the American Meteorological Society, February 2009.

3. *The Earth System Grid Federation: An Open Infrastructure for Access to Distributed Geospatial Data*, IEEE special issue, *Future Generation Computing Systems*: Luca Cinquini, Daniel Crichton, Chris Mattmann, Gavin M. Bell, Bob Drach, Dean Williams, John Harney, Galen Shipman, Feiyi Wang, Philip Kershaw, Stephen Pascoe, Rachana Ananthakrishnan, Neill Miller, Estanislao Gonzalez, Sebastian Denvil, Mark Morgan, Sandro Fiore, Zed Pobre, Roland Schweitzer; accepted, due out in 2013.

18 March 2013

Dr. Donald E. Anderson, Jr.
NOAA Climate Program Office
Program Manager, NOAA Environmental Software Infrastructure and Interoperability (NESII) Group, Modeling, Analysis Predictions and Projections (MAPP) Program, and the National Climate Predictions and Projections (NCPP) Platform

To whom it may concern:

I am writing to support the Earth System Grid Federation (ESGF) for an R&D award. As a program manager at the National Oceanic and Atmospheric Administration, I can attest to the need for the open sharing of climate data and metadata, for research, international assessments, and studies of local impacts. Many of the groups that I manage rely on ESGF for these functions.

In particular, ESGF was the vehicle for data collection and distribution for the Fifth Coupled Model Intercomparison Project (CMIP5), a scientific foundation for the next Intergovernmental Panel on Climate Change (IPCC) assessment (AR5). With contributions from many national and international partners, ESGF delivered a system that has made it possible for the worldwide community to access this data. New pilot projects with NASA, such as obs4MIPs, promise to bridge the gap between the observational and modeling communities and greatly improve the model evaluation process.

ESGF is of particular value to my programs through its integration in the Earth System CoG collaboration environment. Through CoG, ESGF provides data support for community-based activities such as the National Climate Predictions and Projections (NCPP) Platform, which provides information about the local and regional climate, an atmospheric Dynamical Core Model Intercomparison Project (DCMIP), and a Downscaling Intercomparison Project, hosted by NCPP, to be held this summer.

NOAA programs look forward to working with the ESGF effort as it evolves and hope that the team will be recognized for their contribution to the climate community.

Thank you,

Donald E. Anderson, Jr.

Lawrence Livermore National Laboratory
Office of Strategic Outcomes

March 13, 2013

Jason Paragas, PhD
Deputy Program Director for Bio-Security

Subject: *Support Letter for ESGF*

To Whom It May Concern:

The complexity of the life sciences is being unraveled through innovation brought by high throughput technologies, systems biology and more recently biocomputational approaches. As success in these areas grows, vast and disparate data silos are evolving that were derived from scared and rare resources. In fact, much of the research effort may be house in stand-alone computers, email correspondence, and paper notebooks. The extraordinary atomization of our collective research efforts in the life sciences dilutes the potential treasure. Effectively the majority of the work product from the life sciences is lost. The American Academy for the Advancement of Science reported that the Federal investment in research and development is approximately 147 billion in FY10[1]. The life sciences investment is approximately a third of that number. A slim fraction of that investment is transformed into an inconvenient form that collectively advances the field. We, as a life sciences community, are in essence returning copper back to the ground. The greatest opportunity in the life sciences today is not in the frontiers of science but to transform our research effort into a computable resource.

This is not a lofty idea. In fact it has been accomplished, but in the climate sciences field. The efforts of the Earth Sciences Grid Federation have transformed the community from a series of disparate efforts into a dynamic field that expends efforts in scientific frontiers and not in data impedance mismatches. Yet preserving the ability for individual teams to openly compete ideas.

Porting the principles of the Earth Sciences Grid Federation to the life sciences community will disruptively transform and accelerate our efforts. Furthermore, a life sciences application of the Earth Sciences Grid Federation will enable the

---

[1] AAAS Report XXXIV Research and Development FY 2010
[2]http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_me

vision from the recent Whitehouse directive to develop plans to increase access to the Federally funded research[2]. Extending the principles of Earth Sciences Grid Federation into the life sciences will accelerate innovation, contain costs, and preserve the data derived from sacred research resources.

Sincerely,

Jason Paragas, PhD

[2]http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Lawrence Livermore National Laboratory

Biosciences and Biotechnology Division

March 8, 2013

To the R&D 100 Awards Committee:

As the Division Leader for the Biosciences & Biotechnology Division in the Physical and Life Sciences Directorate at Lawrence Livermore National Laboratory (LLNL), I oversee the development of new technologies that will help to improve and accelerate biomedicine and core biotechnologies.

An area of particular interests is the building of an infrastructure for the accelerated therapeutic community. The goal here would be to extend and/or develop new technologies to improve accessing of biological data for the purpose of accelerating drug design and development. To achieve this goal, an underlying computational infrastructure is needed to facilitate the exchange of biological data to allow the integration of data from many existing experimental efforts in genome sequence, structural genomics, biochemical kinetics, and pharmacology. These data would be used in massive-scale systems biology representations of biochemical pathways (endogenous and xenobiotic metabolism, and regulatory mechanisms) within the human body, detailed as interconnected compartmentalized organs.

This infrastructure will allow an integrated system to be developed, including: 1) system-level bioinformatics data (e.g. genomics, transcriptomics, proteomics); 2) protein interaction data (e.g. kinetic characteristics of enzymes) based on atomistic protein structure function and cheminformatics); and 3) adverse outcome data (side effects) from clinical trials and elsewhere. The technical motivation for the infrastructure is based on the success of the Earth System Grid Federation (ESGF) and the status that it has achieved over the past decade in the climate domain—demonstrating the capability to help a scientific community self-organize to build an information and knowledge infrastructure that has revolutionized how climate modeling is done.

Looking forward, the biology community is looking to applying the successful ESGF infrastructure to meet our needs and the possible integration of cross-discipline study between the biology and climate.

Sincerely,

Kenneth W. Turteltaub
Division Leader
Biosciences and Biotechnology Division

National Aeronautics and
Space Administration
Goddard Space Flight Center
Greenbelt, MD 20771

March 7, 2013

Reply to Attn of: 606

R&D 100 Awards
100 Enterprise Drive
Suite 600, Box 912
Rockaway, NJ 07866-0912

Re: Support Letter for ESGF to Receive R&D 100 Awards

To Whom It May Concern:

It is with great pleasure and enthusiasm that we write this letter of recommendation and support for the Earth System Grid Federation (ESGF) to receive an R&D 100 Award. As members of the international observational and re-analysis data prepared working groups, we can attest to the critical importance of validating observational and re-analysis data against climate model simulations. Today, links between observational data experts and model developers/analysts are weak, and as a result, valuable data resources in understanding predictive climate change are underutilized and poorly understood.

As a result of this underutilization, two pilot projects have emerged to help bridge the gaps between the communities in order to better the model evaluation process: observation for Model Intercomparison Projects (obs4MIPs); and re-analysis for Model Intercomaprison Projects (ana4MIPs). These pilot projects aim to apply their observational and re-forecast data sets to the Coupled Model Intercomparison Project (CMIP) process for simulation data verification and validation. It is through the ESGF infrastructure that the co-mingling of data products can exist. At NASA, ESGF is one of the primary conduits in which we distribute NASA climate model output, observations, and re-analysis data to the climate change community. Conversely, it is the primary conduit in which we ascertain and access climate simulation and other related data products.

For the first time ever, ESGF is making it possible for disparate data and resources to be joined together under one knowledge discovery system, across different U.S. federal agencies (DOE, NOAA, NASA, NSF) and international borders. Without ESGF, it would be difficult to generate reports such as the Intergovernmental Panel on Climate Change (IPCC) Fifth Assessment Report (AR5), due out at the end of 2013.

We have been pleased to be part of this ongoing ESGF effort and hope that the team will

be recognized for their outstanding contribution to climate change science and other scientific domains.

Sincerely,

Dr. W. Phillip Webster
Chief, Computational and Information Sciences and Technology Office, NASA Goddard

Digitally signed by TSENGDAR LEE
DN: c=US, o=U.S. Government, ou=NASA, ou=People, cn=TSENGDAR LEE, 0.9.2342.19200300.100.1.1=tjlee1
Date: 2013.03.07 12:44:05 -05'00'

Dr. Tsengdar Lee
Program Manager, High End Computing, NASA Headquarters

California Energy Systems for the 21st Century

Dr. Noah Goldstein, Director, Livermore Energy Systems Informatics Capability
Lawrence Livermore National Laboratory

Letter of Support for ESGF March 1, 2013

To whom it may concern:

I relish this opportunity to recommend the Earth System Grid Federation (ESGF) for an R&D 100 award for its innovation and contributions. At LLNL's Livermore Energy Systems Informatics Capability (LESIC), we are tasked with identifying and implementing systems for integrating and managing large volumes of Energy Systems data. Among the challenges we face is how to structure data across local and regional repositories that can have varying degrees of sensitivity, yet as a whole will contribute greatly to our mission-based projects. In ESGF, we see a solution that will enable us to maximize data integration for simulation in a labor- and computationally- efficient manner.

In the upcoming months, we hope to begin a large integrated Advanced Computing project, called the California Energy Systems for the Twenty-First Century, or CES-21. This project will enable the California Electric and Gas utilities to leverage LLNL's simulation and modeling expertise to solve some of the most challenging problems of our Energy System modularization, including renewable power integration, streaming data, and massively expanding volumes of sensitive data. We have the responsibility of managing those data associated with that effort, as well as developing new techniques for seamless simulation and data analysis. In CES-21, we are looking to innovative approaches like ESGF to provide some leadership in creating solutions to our problems, especially in the domains of data integration and management, pre- and post-processing, as well as data analysis and visualization.

Here at LLNL and LESIC, we hope that ESGF will enable CES-21 and similar projects to help with creating a more efficient grid, a safer energy system, with enthusiastic support of our stakeholders and public partners. We are very pleased that ESGF exists as an easy-to-use model of data distribution, integration, and processing for some of society's most important problems. At LESIC, we look forward to applying ESGF's innovations in the Energy System domain.

Thank you,

Noah Goldstein
Goldstein8@llnl.gov

UNITED STATES DEPARTMENT OF COMMERCE
National Oceanic and Atmospheric Administration
GEOPHYSICAL FLUID DYNAMICS LABORATORY
Princeton University
Forrestal Campus, US Route 1
201 Forrestal Road
Princeton, New Jersey 08540

22 February 2013

Support Letter for ESGF

Ronald J Stouffer
Physical Scientist GFDL/NOAA
Head of the Climate Model Intercomparison Project (CMIP) Panel

To whom it may concern:

The CMIP activity is an organized effort by the world's climate modeling community to freely provide climate information to other scientists and anybody who is able to download and process the data. These data sets are being used for analysis and increased scientific understanding. They provides the scientific basis for many of the international climate reports, including the Intergovernmental Panel on Climate Change (IPCC) 5th Assessment report due out later this year. The scientific findings in the IPCC reports provide input to climate policy negotiations between and among countries. The CMIP database is therefore of extremely high value to society both inside the U.S. and the world.

The total amount of data under CMIP Panel oversight is about 4 Pb and is one of the larger databases in the world. This data does not reside on any one server but is distributed around the world across 10's of data servers. ESGF provides the infrastructure which allows the data to be useful to those who try to obtain data from the CMIP database. This software is extremely complex as in involves allowing various servers to access information on each other across the internet and the security issues associated with that process.

There is an urgent need for more funding support for the ESGF activity. Without the support, it is possible that serving the CMIP database as outlined above will no longer be possible which will greatly hinder the advancement of climate science.

Therefore I strongly support the ESGF proposal.