# DOE SciDAC's Earth System Grid Center for Enabling Technologies Final Report

D. N. Williams

September 29, 2011

**Disclaimer**

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

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

# DOE SciDAC's Earth System Grid Center for Enabling Technologies
## *Final Report*
## *October 1, 2006 through September 30, 2011*

***Principal Investigators***
Dean N. Williams[3], Ian T. Foster[1], and Don E. Middleton[6]

***The Earth System Grid Center for Enabling Technologies Team:***
Rachana Ananthakrishnan[1], Neill Miller[1], Frank Siebenlist[1]
Mehmet Balman[2], Junmin Gu[2], Vijaya Natarajan[2], Arie Shoshani[2], Alex Sim[2],
Gavin Bell[3], Robert Drach[3], Michael Ganzberger[3],
Jim Ahrens[4], Phil Jones[4],
Daniel Crichton[5], Luca Cinquini[5],
David Brown[6], Danielle Harper[6], Nathan Hook[6], Eric Nienhouse[6], Gary Strand[6], Hannah Wilcox[6],
Nathaniel Wilhelmi[6], Stephan Zednik[6],
Steve Hankin[7], Roland Schweitzer[7],
Meili Chen[8], John Harney[8], Ross Miller[8], Galen Shipman[8], Feiyi, Wang[8],
Peter Fox[9], Patrick West[9], Stephan Zednik[9],
Ann Chervenak[10], and Craig Ward[10]

**Climate simulation data are now securely accessed, monitored, cataloged, transported, and distributed to the national and international climate research community.**

[1] Argonne National Laboratory (ANL)
[2] Lawrence Berkeley National Laboratory (LBNL)
[3] Lawrence Livermore National Laboratory (LLNL)
[4] Los Alamos National Laboratory (LANL)
[5] National Aeronautics and Space Administration (NASA)
[6] National Center for Atmospheric Research (NCAR)
[7] National Oceanic and Atmospheric Administration (NOAA)
[8] Oak Ridge National Laboratory (ORNL)
[9] Rensselaer Polytechnic Institute (RPI)
[10] University of Southern California, Information Sciences Institute (USC/ISI)

# 1    Executive Summary

The mission of the Earth System Grid Federation (ESGF) is to provide the worldwide climate-research community with access to the data, information, model codes, analysis tools, and intercomparison capabilities required to make sense of enormous climate data sets. Its specific goals are to (1) provide an easy-to-use and secure web-based data access environment for data sets; (2) add value to individual data sets by presenting them in the context of other data sets and tools for comparative analysis; (3) address the specific requirements of participating organizations with respect to bandwidth, access restrictions, and replication; (4) ensure that the data are readily accessible through the analysis and visualization tools used by the climate research community; and (5) transfer infrastructure advances to other domain areas.

For the ESGF, the U.S. Department of Energy's (DOE's) Earth System Grid Center for Enabling Technologies (ESG-CET) team has led international development and delivered a production environment for managing and accessing ultra-scale climate data. This production environment includes multiple national and international climate projects (such as the Community Earth System Model and the Coupled Model Intercomparison Project), ocean model data (such as the Parallel Ocean Program), observation data (Atmospheric Radiation Measurement Best Estimate, Carbon Dioxide Information and Analysis Center, Atmospheric Infrared Sounder, etc.), and analysis and visualization tools, all serving a diverse user community. These data holdings and services are distributed across multiple ESG-CET sites (such as ANL, LANL, LBNL/NERSC, LLNL/PCMDI, NCAR, and ORNL) and at unfunded partner sites, such as the Australian National University National Computational Infrastructure, the British Atmospheric Data Centre, the National Oceanic and Atmospheric Administration Geophysical Fluid Dynamics Laboratory, the Max Planck Institute for Meteorology, the German Climate Computing Centre, the National Aeronautics and Space Administration Jet Propulsion Laboratory, and the National Oceanic and Atmospheric Administration.

The ESGF software is distinguished from other collaborative knowledge systems in the climate community by its widespread adoption, federation capabilities, and broad developer base. It is the leading source for present climate data holdings, including the most important and largest data sets in the global-climate community, and—assuming its development continues—we expect it to be the leading source for future climate data holdings as well.

Recently, ESG-CET extended its services beyond data-file access and delivery to include more detailed information products (scientific graphics, animations, etc.), secure binary data-access services (based upon the OPeNDAP protocol), and server-side analysis. The latter capabilities allow users to request data subsets transformed through commonly used analysis and intercomparison procedures. As we transition from development activities to production and operations, the ESG-CET team is tasked with making data available to all users seeking to understand, process, extract value from, visualize, and/or communicate it to others. This ongoing effort, though daunting in scope and complexity, will greatly magnify the value of numerical climate model outputs and climate observations for future national and international climate-assessment reports. The ESG-CET team also faces substantial technical challenges due to the rapidly increasing scale of climate simulation and observational data, which will grow, for example, from less than 50 terabytes for the last Intergovernmental Panel on Climate Change (IPCC) assessment to multiple Petabytes for the next IPCC assessment. In a world of exponential technological change and rapidly growing sophistication in climate data analysis, an infrastructure such as ESGF must constantly evolve if it is to remain relevant and useful.

Regretfully, we submit our final report at the end of project funding. To continue to serve the climate-science community, we are currently seeking additional funding. Such funding would allow us to maintain and enhance ESGF production and operation of this vital endeavor of cataloging, serving, and analyzing ultra-scale climate-science data.

At this time, the entire ESG-CET team would like to take this opportunity to *sincerely thank* our funding agencies in the DOE Scientific Discovery through Advanced Computing (SciDAC) program and the Office of Biological and Environmental Research (OBER)—as well as our national and international collaborators, stakeholders, and partners—for allowing us to work with you and serve the community these past several years.

# Contents

## 2   Overview

### 2.1   Problem Statement

The Earth System Grid Center for Enabling Technologies (ESG-CET) was established to address the needs of modern-day climate data centers and climate researchers. Specifically, ESG-CET addresses the needs of both data centers and researchers for interoperable discovery, distribution, and analysis of large and complex data sets. Under the leadership of the Department of Energy's Program for Climate Model Diagnosis and Intercomparison (PCMDI) at Lawrence Livermore National Laboratory (LLNL) and in partnership with Argonne National Laboratory (ANL), the National Center for Atmospheric Research (NCAR), Lawrence Berkeley National Laboratory (LBNL), Oak Ridge National Laboratory (ORNL), Los Alamos National Laboratory (LANL), the National Aeronautics and Space Administration (NASA), the National Oceanic and Atmospheric Administration (NOAA), and others in the national and international community—including centers in the U.K., Germany, France, Italy, Japan, and Australia—an internationally federated, distributed data archival and retrieval system was established under the name Earth System Grid Federation (ESGF). Although this development effort is coordinated internationally, the ESG-CET team is the primary contributor to the ESGF software stack. ESGF work has resulted in production of an ultra-scale data system, empowering scientists to engage in new and exciting data exchanges that could ultimately lead to breakthrough climate-science discoveries. Through this ESGF effort, the team was able to achieve its proposal goals:

1.  Sustain the successful existing Earth System Grid (ESG) system;
2.  Address projected scientific needs for data management and analysis;
3.  Extend ESG to support the major Intergovernmental Panel on Climate Change (IPCC) assessment in 2011;
4.  Support the Climate Science Computational End Station at the DOE Leadership Computing Facility at ORNL; and
5.  Support climate model evaluation activities for future DOE climate applications and projects.

One of climate science's most difficult challenges remains managing and understanding massive amounts of global atmospheric, land, ocean, and sea-ice model data generated by ever more complex computer simulations, and driven by ever larger qualitative and quantitative observations (Overpeck 2011). Because of rapid increases in technology, storage capacity, and networks and the need to share information, communities are providing access to federated open-source collaborative systems that everyone (scientists, students, policymakers, etc.) can use to explore, study, and manipulate large-scale data. The ESGF software stands out from these emerging collaborative knowledge systems in the climate community along multiple dimensions: the amount of data provided (hundreds of terabytes), the number of global participating sites (over a few dozen), the number of users (over 25,000), the amount of data delivered to users (over 1.5 petabyes), and the sophistication of its software capabilities, and is therefore considered the leader for both present and future data holdings.

ESG was critical to the successful archiving, delivery, and analysis of the Coupled Model Intercomparison Project (CMIP), phase 3 (CMIP3) data for the Fourth Assessment Report (AR4) of the IPCC. It will prove equally important in meeting the data management needs of CMIP, phase 5 (CMIP5), which will provide petascale data informing the 2013 IPCC's Fifth Assessment Report (AR5). Although the ESGF has been indisputably important to CMIP, its current and future impact on climate is not limited only to this high-profile project. ESGF has been used to host data for a number of other projects (see data archive Table 7 in section 4), including CCSM and NARCCAP at NCAR, C-LAMP at ORNL, POP at Los Alamos National Laboratory (LANL), and AMIP at LLNL. These data archives have been augmented with observational data sets (for example, ARMBE, CDIAC, National Aeronautics and Space Administration (NASA) satellite observation data sets [CloudSat, MLS, MISR, AIRS, and TRMM], and NASA/NOAA reanalysis data sets [MERRA, CERES]).

ESGF requires integration of software and hardware resources spread across all-important worldwide institutions carrying out climate research. The United States (U.S.) participants in the ESG Federation include Argonne National Laboratory (ANL), National Oceanic and Atmospheric Administration (NOAA)/Geophysical Fluid Dynamics Laboratory (GFDL), LANL, LBNL, LLNL, NASA/Jet Propulsion Laboratory (JPL), NASA/Godard, NCAR, ORNL, Pacific Northwest National Laboratory (PNNL), NOAA/Pacific Marine Environmental Laboratory (PMEL), Rensselaer Polytechnic Institute (RPI), and the University of Southern California (USC) to name a few. The members of this group, led by DOE's LLNL, work across institutional boundaries to contribute to the development and integration of disparate software that facilitates climate research. Over the years, ESG (now ESGF) has seen its data archives grow substantially, and now boasts between 1.5 to 2.0 petabytes of data distributed to the community, comprising over 25,000 registered users. Close group collaborations and marathon national and international face-to-face meetings, teleconferences, coding and debugging sessions met the federation requirements for such an ambitious endeavor. Group software development included the following:

- Metadata expansion, for search services
- Data access, for data of commercial and non-commercial users
- Federated security, for single sign-on authentication and authorization
- Data services, for access and movement of large data sets
- Product services, for the generation and return of data products (e.g., visualization, reduced data sets, etc.)
- Framework, for the easy installation of the software stack
- Notification, for informing users of data changes and status
- Monitoring, for up to the minute condition of the federated system
- Metrics, for reporting data and system use, and
- Help desk, for helping users address questions and issues pertaining to the ESGF enterprise system.

ESG-CET's success in disseminating climate data has exceeded all expectations, and ESG (now ESGF) is recognized for its critical role in U.S. and international climate research.

## 2.2    Summary of Project History

Since production began in 2004, the ESG has housed and distributed significant and often extremely large data collections for many well-known efforts in climate science. The most notable example is the analyses of data contained in ESG's CMIP3 data archive that resulted in over 600 peer-reviewed scientific publications. Some of these works contributed to the 2007 Nobel Prize-winning IPCC AR4 efforts. As of the mid-point of the ESG-CET project (April 2009), the ESG production system had over 14,000 registered users and was managing over 237 TB of model data, comprising the contents of archives at five sites around the U.S. Also at the time, ESG users had downloaded more than 700 TB of data. The ESG-CET project incurred its mid-term assessment review in 2009. Conducted by the Offices of Advanced Scientific Computing Research (ASCR) and Biological and Environmental Research (BER) program management, six distinguished panel reviewers knowledgeable in climate, computer, and computational science, and software design and development gave thoughtful and insightful feedback on the ESG-CET project. The panel was presented with seven questions (see Figure 1 below) and rated the project based on the questions (see Figure 2). Reviewers gave an overall score ranging from 1, being the lowest possible score, to 10, the highest.

**At a Glance, mid-term reviewers' rating of the ESG-CET project.**

Seven questions were posed to the reviewers:
1. Scientific and/or technical merit of the project
2. Appropriateness of the proposed methods or approach
3. Competency of the key personnel and adequacy of the proposed resources
4. Performance under existing award
5. Reasonableness and appropriateness of the budget
6. How well does the project advance the SciDAC goals?
7. Additional Comments

**Figure 1:** *Mid-term scores of the ESG-CET project.*

Reviewers rated the overall ESG-CET project, based on the above questions, on a scale from 1 to 10, where 1 being the lowest possible score and 10 the highest:
1. Reviewer (1) – scored ESG-CET: 9
2. Reviewer (2) – scored ESG-CET: 9
3. Reviewer (3) – scored ESG-CET: 9
4. Reviewer (4) – scored ESG-CET: 8
5. Reviewer (5) – scored ESG-CET: 8
6. Reviewer (6) – scored ESG-CET: 9
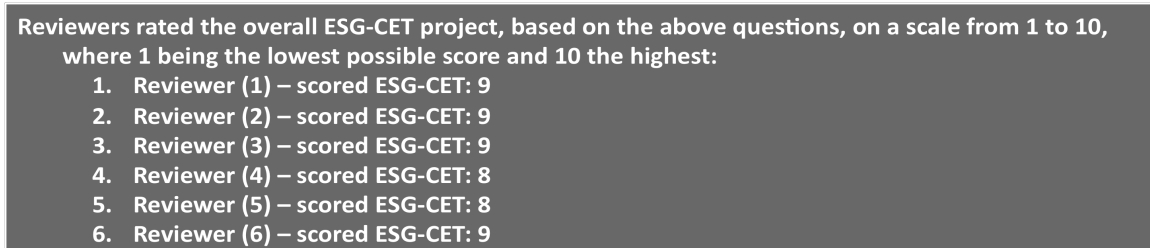
**Figure 2:** *Reviews 2009 mid-term score for the ESG-CET project.*

The following are a few of the numerous positive comments made by the review panel:

- "An important piece of distributed science infrastructure for the global climate community"

- "This project is especially advancing the SciDAC goal of creating a scientific computing software infrastructure that bridges the gap between applied mathematics & computer science and computational science in the environmental sciences"

- "ESG has become an indispensable tool for stakeholders to get their research done and to distribute the petabytes of data generated by climate modeling."

In late 2009 and early 2010, ESG-CET transitioned to work more closely with other international leading world climate data centers. Work with the international Global Organization for Earth System Science Portals (GO-ESSP) led to the establishment of a global federation, founded to provide data archival and access for CMIP5 data. This work culminated in the formation of the international ESGF, which currently consists of eight climate gateways—each of which provides for user registration and management and allows users to search, discover, and request data—and 25 data nodes, each of which allows data maintained on disk or through tertiary mass store (i.e., tape archive) to be published (or exposed) to any gateway. ESGF, a coordinated international collaboration of people and institutions, works to build a long-term, open-source software infrastructure to manage and analyze Earth system science data. Figure 3 shows the current configuration of the ESGF system.
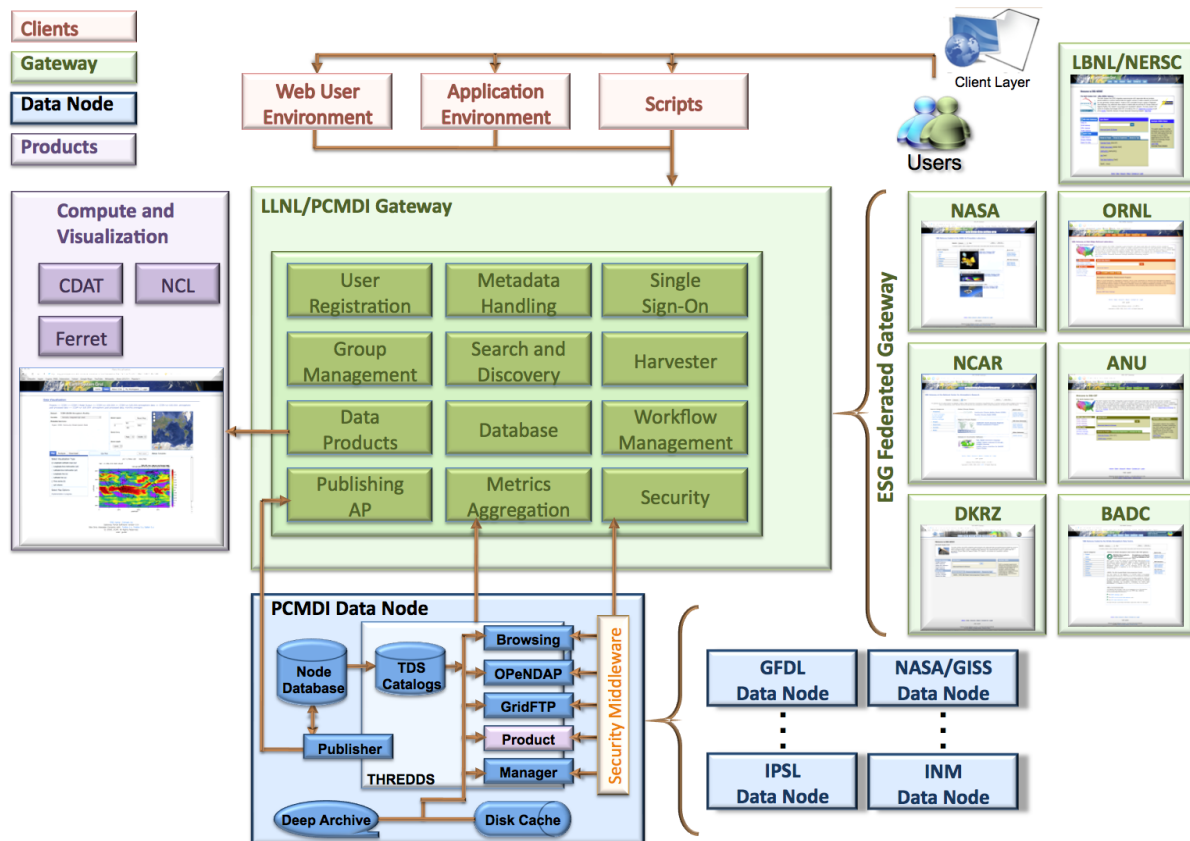
**Figure 3**: *Users can access ESGF data using Web browsers, scripts, and client applications. ESGF is separated into gateways (green) and data nodes (blue). Gateways handle user registration and management and allow users to search, discover, and request data. Data nodes are located where the data resides, allowing data to be published (or exposed) on disk or through tertiary mass store (i.e., tape archive) to any gateway. They also handle data reduction, analysis, and visualization. ESGF currently comprises eight national and international gateways, four of which hold special status in housing CMIP5/AR5 replication data sets: LLNL/PCMDI, the British Atmospheric Data Center (BADC), the German Climate Computing Centre (DKRZ), and the Australian National University (ANU) National Computational Infrastructure (NCI). Users have access to all data from the federation regardless of which gateway is used.*

## 2.3    Project Success and Accomplishments

Nationally and internationally, the software developed by the ESG-CET team (and indeed the community) provides a spread of model and observational data results, which are used by a myriad of climate projects (see Table 7 for ESGF's National and International Data Archive"). The climate-modeling community recognizes the value of model and observational data intercomparison activities and has devoted enormous resources to running large-scale computer simulations and collecting vast amounts of observational data for use in national and international assessment reports such as the U.S. National Climate Assessment Report and the IPCC AR5, respectively. The scientific value of such large-scale events would have been extremely limited without a software infrastructure that allows scientists outside modeling centers and observational collection agencies to access extreme scale data output. ESG-CET, along with its international partners, provides the essential infrastructure element that allows everyone equal access to large disparate data that would have otherwise been accessible only with great difficulty. Through ESG-CET 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.

In continuation of the success of the project from 1999–2005, ESG-CET sustained and expanded the CMIP3 data archive at LLNL. Early in the funding period, the CMIP3 archive was expanded from 23 to 25 global coupled ocean-atmospheric models. Still in use today, the archive continues to serve more than 4,000 registered users that have downloaded over 1.2 PB of data. To date, approximately 600 publications have resulted from the analysis of this data, and this archive is associated with the 2007 Nobel Prize-winning IPCC AR4 effort.

For the upcoming CMIP5 data archive for the IPCC AR5, the team has put into production the ESG gateway and peer-to-peer (P2P) software stacks (described below in section 2.4). These software systems are expected to provide data and analysis results that will be fundamental to the 2013 scientific assessment of climate science by the IPCC community. The protocol for model runs calls for a wider variety of experiments than ever before and a more comprehensive collection of model output. For an up-to-the-minute overview of the data availability via ESGF, please visit http://cmip-pcmdi.llnl.gov/cmip5/esg_tables/transpose_esg_static_table.html, which shows the IPCC model runs for designated experiments. For the CMIP5 archive status, see http://esgf.org/wiki/Cmip5Status/ArchiveView, which summarizes the modeling centers, ESG gateways, ESGF nodes, and the number of data sets in the global CMIP5 archive and the total size of the archive, currently over 240 TB.

Today, ESGF delivers data for a wide variety of purposes and projects, including both model simulation data and observational data from BER field programs. Our goal has been to address and deliver particular requirements and ultra-scale capabilities needed to catalog, access, and analyze large data sets that DOE's Office of Biological and Environmental Research (BER) is responsible for generating and distributing to the community. Several archives in particular provided a focus for the work:

- The IPCC CMIP3, CMIP5, and other multi-model data sets managed and distributed by the BER-funded Program for Climate Model Diagnosis and Intercomparison (PCMDI);

- The Community Climate System Model (CCSM) as it prepares for the AR5 and beyond;

- The Atmospheric Radiation Measurement (ARM) Climate Research Facility, Carbon Dioxide Information Analysis Center (CDIAC), and AmeriFlux observational data hosted by ORNL for the BER Atmospheric Science Research (ASR) and Terrestrial Carbon and Ecosystem Research programs, as well as the observational data from NASA-sponsored ORNL Distributed Active Archive Center (ORNL DAAC) for Biogeochemical Dynamics;

- The model data from the Carbon-Land Model Intercomparison Project (C-LAMP) and the Advanced Very High-resolution climate model simulations resulting from BER's Earth System Modeling (ESM) and Regional and Global Modeling (RGM) programs; and

- The North American Regional Climate Change Assessment Program (NARCCAP) and the Community Climate System Model (CCSM) hosted at NCAR.

These archives are important for BER science missions and are representative of climate-data sources used for advanced climate science. The ESG-CET project has met all milestones for these archives on or ahead of schedule. Our requests for future project funding, if granted, will allow us to stay in sync with new release of the repositories to deliver analysis and visualization and data products at scales appropriate for the research community. Tables 1 and 2 below show ESGF global production gateways and data nodes in use, providing data to the greater community.

**Table 1:** *ESG operational/production gateways.*

| Institute | Gateway URL | Version | Comment | Project | Contact |
|-----------|-------------|---------|---------|---------|---------|
| Argonne | http://www.esg.anl.gov/gateway | 1.3.1 | Production | HOMME | neillm@mcs.anl.gov |
| BADC | http://cmip-gw.badc.rl.ac.uk/ | 1.2.0 RC1 | Production | CMIP5, TAMIP2 | Maurizio.Nagni@stfc.ac.uk |
| DKRZ | http://ipcc-ar5.dkrz.de | 1.3.2 | Production | CMIP5, EUCLIPSE | estanislao.gonzalez@zmaw.de |
| JPL/NASA | http://esg-gateway.jpl.nasa.gov/ | 1.3.0 | Production | AIRS, MLS, TES | Luca.Cinquini@jpl.nasa.gov |

| | | | | | |
|---|---|---|---|---|---|
| LBNL/NERSC | http://esg.nersc.gov/esgcet | 1.2.0. RC1 | Production | CLIMES | mbalman@lbl.gov |
| LLNL/PCMDI | http://pcmdi3.llnl.gov/esgcet | 1.3.1 | Production | CMIP5 | drach1@llnl.gov |
| NCAR | http://www.earthsystemgrid.org | 1.3.1 | Production | CSSM, NARCCAP, PCM | esg-support@earthsystemgrid.org |
| NCI | http://esg.nci.org.au/esgcet | 1.3.1 | Production | CMIP5 | muhammad.atif@anu.edu.au |
| ORNL | http://esg2-gw.ccs.ornl.gov/ | 1.3.0 | Production | C-LAMP | rgmiller@ornl.gov |

**Table 2:** *ESGF operational/production data nodes.*

| Institute | THREDDS URL | Version | Comment | Contact |
|---|---|---|---|---|
| Argonne | http://esg.anl.gov/thredds | 1.0.4 | Production | neillm@mcs.anl.gov |
| BADC | http://cmip-dn.badc.rl.ac.uk/thredds | 1.0.4 | Production | Stephen.Pascoe@stfc.ac.uk |
| BCC | http://bcccsm.cma.gov.cn/thredds | 1.0.4 | Production | zhangli@cma.gov.cn |
| CCCMA | http://dap.cccma.uvic.ca/thredds | 1.0.4 | Production | cccma_info@ec.gc.ca |
| CNRM | http://esg.cnrm-game-meteo.fr/thredds | 1.0.4 | Production | contact.cmip5@meteo.fr |
| DKRZ | http://bmbf-ipcc-ar5.dkrz.de/thredds | 1.0.4 | Production | estanislao.gonzalez@zmaw.de |
| IPSL | http://vesg.ipsl.fr/thredds | 1.0.4 | Production | ipsl-cmip5@ipsl.jussieu.fr |
| LBNL/NERSC | http://esg-datanode.nersc.gov/thredds | 1.0.4 | Production | mbalman@lbl.gov |
| LLNL/PCMDI | http://pcmdi3.llnl.gov/thredds | 1.0.4 | Production | drach1@llnl.gov |
| NASA/JPL | http://esg-datanode.jpl.nasa.gov/thredds | 1.0.4 | Production | luca.cinquini@jpl.nasa.gov |
| NCAR | http://tds.ucar.edu/thredds | 1.0.4 | Production | esg-support@earthsystemgrid.org |
| NCC | http://norstore-trd-bio1.hpc.ntnu.no/thredds | 1.0.4 | Production | Ingo.Bethke@uni.no |
| NCCS | http://esg.nccs.nasa.gov/thredds | 1.0.4 | Production | tsanah@gmail.com |
| NCI | http://esgnode1.nci.org.au/thredds/catalog.html | 1.0.4 | Production | muhammad.atif@anu.edu.au |
| ORNL | http://esg2-sdn1.ccs.ornl.gov/thredds | 1.0.4 | Production | rgmiller@ornl.gov |

For more detailed project successes and accomplishments achieved along the way, please visit the ESG-CET website for "Accomplishments" and "Bi-Annual Progress Reports" (URL: http://esg-pcmdi.llnl.gov/). In particular, the progress reports highlights the overall success of the project in meeting its deliverables.

## 2.4    Balanced Research Ecosystem: Overall Architecture Design

The Earth System Grid Federation is a system of cooperating sites distributed across four continents (North America, Europe, Asia, and Australia) that collectively represent a global archive composed of hundreds of petabytes of data, both model output and observations. From the software perspective, the system is the result of the integration of two distinct architectures:

- **Peer-to-Peer (P2P)***:* An innovative paradigm in which all participating sites interact as equal partners, can be flexibly configured to expose different sets of services, and can act as consumers or providers of services depending on circumstances
- **Client-server**: A more traditional model, based on a specialized gateway application that acts as a broker towards services provided by a data node

The client-server architecture was developed first to fit the specific needs of the CMIP5 data archive and the ongoing current IPCC-AR5 activities. The P2P architecture represents an evolution of the former architecture intended to increase its flexibility, scalability and dynamicity, with the goal of opening up the system to other

possible climate and scientific domains (nuclear energy, biology, remote satellite observations, etc.) that may be less forgiving about the lack of data formats and conventions.

The two architectures seamlessly integrate with each other because of federation-wide agreements on protocols, trust relations, and application programming interfaces (APIs). The result is an ecosystem of software archives and services, geographically distributed and locally managed where users and clients can access as if they were a single standalone system (see Figure 4).
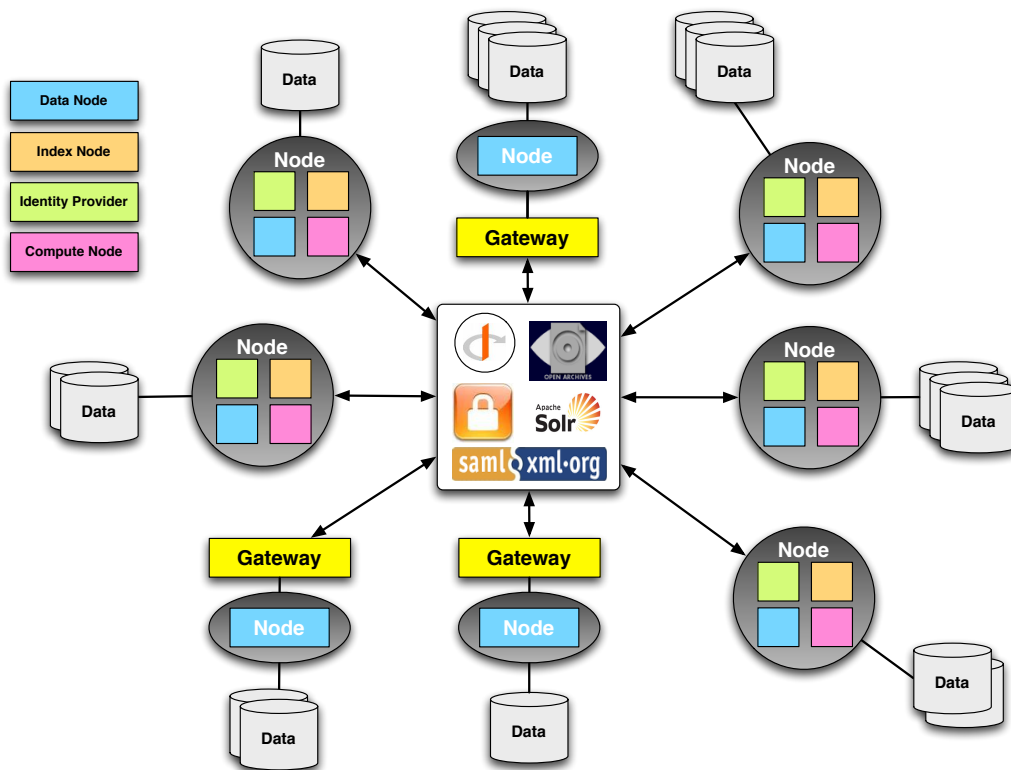


**Figure 4:** *The Earth System Grid Federation combines the new P2P architecture with the more traditional client-server model. The P2P system was expressly designed for extensibility and scalability, and it supports geospatial and temporal search, dashboard showing system metrics, user interface for notification, and a rich set of climate analysis tools for data manipulation. For backward compatibility and interoperability, the data node component of the P2P software stack services both the ESG gateway (in yellow) and the ESGF node. These nodes form the network of a geographically distributed global federation that is based on standard protocols and application programming interfaces—such as OpenID, Solr, Security Assertion Markup Language (SAML), and Open Archive Initiative (OAI)—thus allowing seamless access to a large and diverse user community.*

### 2.4.1  Peer-To-Peer Architecture

The ESGF Peer-To-Peer (P2P) architecture is based on the concept of a dynamic system of nodes, which interact on an equalitarian basis and can offer a broad range of user and data services depending on how they are set up. Each ESGF node is composed of some form of data storage (online disk or deep archive) and of a configurable set of applications and services to publish, search, download, and analyze the data. Specifically, an ESGF node can be configured to posses one or more "software types", each entailing a specific functionality:

- **Data Node***: contains services to publish and serve data through a variety of protocols such as HTTP, GridFTP, OPeNDAP, etc.
- **Index Node***: contains services to harvest metadata and enable data discovery
- **Identity Provider***: contains facilities to register, authenticate, and authorize users

- **Compute Node***: contains application servers for data reduction, analyzing and visualizing the data

Internally, each node is based on the integration of modular applications and servers, often developed independently by the community, such as the THREDDS Data Server (TDS), Live Access Server (LAS), Hyrax, GridFTP, and MyProxy. A fully featured node, configured to possess all software types, can operate as a standalone and expose its data directly to the users. Alternatively, a minimally configured data node can be front-ended by a specialized web application called a gateway, which then assumes the responsibility of exposing higher-level services for user management and data search and discovery.

The core component of each ESGF node is the Node Manager, a Java utility that runs as a separate process and allows common operations tasks for a managed server, regardless of its location with respect to its administration server. The Node Manager provides critical functionality for hosting applications and high availability requirements; it allows starting and stopping the servers remotely from the command line and provides automatic restart of services after an unexpected failure. The Node Manager is also the component that allows for P2P networking. That is, in this distributed environment each ESGF node advertises its capabilities (public keys, functionality, service endpoints etc.) in a registry document, which is continuously propagated to all other nodes through a "gossip" protocol. A gossip protocol is a style of computer-to-computer communication protocol inspired by the form of gossip seen in social networks. As a consequence, each node is constantly aware of the full state of the federation, and nodes can join and leave the federation dynamically, without impacting the availability of data and services at any other node. Others are using the P2P architecture structure from the industry in a variety of ways by businesses, consumers, government agencies, academic institution and others to distribute quickly large amount of data and information.

From the client perspective, ESGF node users can interact with the system either through a traditional Web browser or through a growing number of rich desktop applications, such as UV-CDAT (Ultra-scale Visualization Climate Data Analysis Tools), DML (Data Mover-Lite), the CDX (Climate Data Exchange) toolkit, and any script or program based on OPeNDAP data access.

### 2.4.2   Gateway Architecture

Discipline specific gateways represent entry points for users (and user clients) into a federated set of scientific data collections and services. A gateway is a web-accessible application that includes a user interface and high-level services including user identity, user administration and security management, search, browse, data publication and versioning, data-replica management, rich-model metadata, data-use metrics, and other aggregation functions. Gateways may also act as a broker for data requests sent to the data node application servers, which then serve data over HTTP, or GridFTP protocols. Logically, a gateway provides a set of aggregated functions enabling a user to access data from one or more associated data nodes. Since gateways exchange metadata with each other and are federated through a common security infrastructure, a user can seamlessly find and access data that is served from anywhere in the system.

### 2.4.3   Ecosystem: Both Systems Co-existing

As a requirement, the P2P and gateway and data-node systems must work together to provide an integrated and reliable level of services to the research community. This interoperability is made possible because of two key factors. First of all, the ESG gateway utilizes the data-node components of an ESGF node (see Figure 5), thus providing the same data services, albeit accessed through a different user interface. Most importantly, all ESGF nodes and ESG gateways in the system can interoperate with each other because of the adoption of common federation protocols, which allow users to interact with the overall system as if they were a single application. The following areas are key to enable interoperability within the federation:

- **Data search and discovery***: in the current model, ESGF nodes and the ESG gateway harvest each other's metadata holdings through the OAI (Open Archive Initiative) protocol. Work is under way to define a common query API so that nodes could query each other in real time by executing a distributed search.

- **Data access***: data can be accessed as geographic and temporal subsets, irrespective of the source location and underlying format, via the OPeNDAP protocol. Additionally, data can be downloaded as full files through standard HTTP and GridFTP servers.
- **Security***: users can register with any ESGF node or ESG gateway and use their assigned OpenID to login via a browser at any other site in the federation. Programmatic access to data (i.e. via rich desktop clients and scripts) is supported through the Secure Sockets Layer (SSL) and Public Key Infrastructure (PKI). Also, each node can authorize users to access specific data collections and encode this information as digitally signed SAML (Security Assertion Markup Language) statements that are honored by all other nodes in the system.

By offering two systems, data providers have more flexibility to deliver their data products to a wider community. Additionally, the modularity and flexibility of the P2P architecture represent a clear path for the adoption of the system, in its entirety or in parts, by other communities and scientific domains.
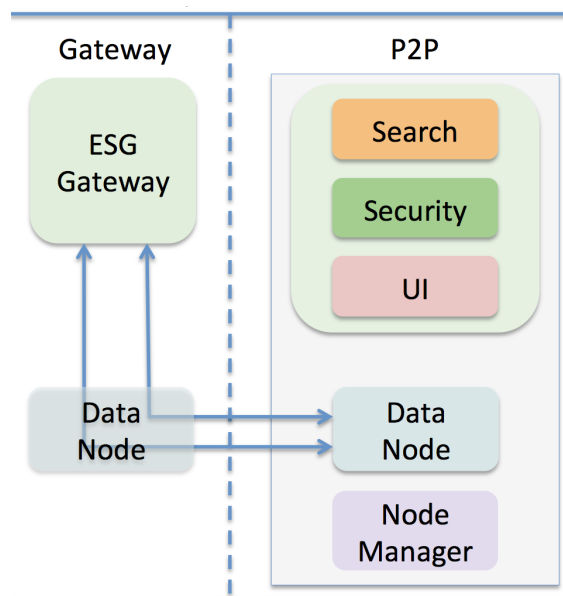


**Figure 5:** *Ecosystem: co-exiting stable systems*

## 2.5    State of the Ecosystem

Table 3 below shows the current state of the ESGF P2P nodes, listing national institutions that are gearing up to use it to serve future DOE, NASA, and NOAA climate simulation and observational data. Table 3 complements the above Table 1, which lists the nine national and international institutions with operational ESG gateways, predominately used for serving CMIP5 data. Table 2 shows a listing of several ESGF data nodes that are harvesting data to either an ESG gateway or ESGF node, or both.

**Table 3:** *Operational ESGF P2P nodes.*

| Institute | Portal URL | Version | Project | Contact |
|---|---|---|---|---|
| Argonne | http://dev.esg.anl.gov/esgf-web-fe/ | 1.0.0 beta | CSSEF, HOMME | neillm@mcs.anl.gov |
| LLNL/PCMDI | http://esgf-node3.llnl.gov/esgf-web-fe/ | 1.0.0 beta | CSSEF, CMIP5, obs4MIPs | bell51@llnl.gov |
| NASA/JPL | http://esg-datanode.jpl.nasa.gov/esgf-web-fe/ | 1.0.0 beta | AIRS, MLS, TES, TRMM, CERES, MERRA, MODIS, QuikSCAT, PO.DAAC, AVISO, AMSR-E | Luca.Cinquini@jpl.nasa.gov |
| ORNL | http://esg-vm-demo02.ccs.ornl.gov/esgf-web-fe/ | 1.0.0 beta | CSSEF, ARMBE, CDIAC, | harneyjf@ornl.gov |

| | | | AmeriFLUX | |
|---|---|---|---|---|
| **PNNL** | http://esg1-gw.pnl.gov/esgf-web-fe/ | 1.0.0 beta | CSSEF | cathy.smith@pnnl.gov |
| **SNL** | Coming soon… | 1.0.0 Beta | CSSEF | csafta@sandia.gov |

## 2.6    Administration, Operations, and Plans for Maintenance

ESGF management is critically important for the continued service of high profile (and other equally important, if lower profile) data sets. The ESGF management consists of distinct parts but functions as a cohesive entity, with the workload distributed evenly among the working partners. From our experience with maintaining CMIP3, we know that future ESGF management must address five distinct functions: operations, maintenance, engineering, training, and administration. These are the common building blocks for any large software effort. The needed work is to sustain the management of existing ESGF tools, algorithms, and components used to distribute large climate data sets in the federation:

- **Operations** to ensure effective implementation and control of ESGF operation activities. This work includes the status and monitoring of hardware, network, and software systems to ensure secure and reliable processing operations and data retrieval
- **Maintenance** to ensure effective implementation and control of software, hardware and network activities to provide optimum performance levels needed by the community.
- **Engineering** to ensure effective implementation and control of technical support. This work entails proper design review, required development to conform to new standards and technologies, and documentation of changes for others to follow.
- **Training** to ensure effective implementation and control of training activities and user support. This work includes the training necessary for new projects, data providers and the diverse user community. It also includes the ESGF help desk, Frequently Asked Questions (FAQ), software-use tutorials (such as the Publisher Tutorial), support mailing lists (such as the esg-support mailing list), and websites and wiki sites (such as the esgf.org site). See section 2.7 for the help desk summary.
- **Administration** to ensure the establishment of policies and the planning and control of activities. This work is especially necessary for the overall operations and monitoring of the ESGF performance and interfacing with stakeholders, program sponsors, team members, data centers, and programs.

For all these distinct functions, operator knowledge and performance will help to support a reliable ESGF system.

ESGF will never be a static system. As the platforms on which it operates—server hardware, networks, operating systems, and browsers—evolve, ESGF software must be adapted. ESGF is also a collaborative development with components from several quasi-independent projects. As one component advances, adaptations in others will be required to support the updated functionality.

## 2.7    CMIP5 Help Desk Weekly Traffic

Our BADC partners and the ESG-CET established a community help desk to assist users with information and resources and to help troubleshoot problems with data, the ESGF ecosystem, and products. This help desk service is provided in the form of software, website, and e-mail. Traffic on the CMIP5 help desk is increasing significantly as the CMIP5 archive increases in volume and the ESGF ecosystem increases in use. Figure 6 shows the summary of support tickets created and closed since April 2011; note that the number of tickets per day continues to grow. ESG-CET team members are essential to keeping this operation going. This is a lot of traffic considering user responses back and forth with help desk providers and follow-on questions by other users. Scientists at BADC, PCMDI, and DKRZ are charged with addressing CMIP5 data and science questions, while technical staffs at these institutions (including NCAR and NASA/JPL) are charged with addressing ESGF system questions. These personnel spend a portion of their day scanning the list of new questions, divvying up the workload and responding to users. Questions that are resolved are place on the ESGF FAQ list to circumvent

answering repeat questions. About half of these queries come directly to cmip5-helpdesk@stfc.ac.uk and the other half to the esg-support@earthsystemgrid.org mailing list. We are moving away from the esg-support mailing list to the help desk for greater control and better tracking of questions. To do this, we are redirecting users from esg-support mailing list to the help desk.
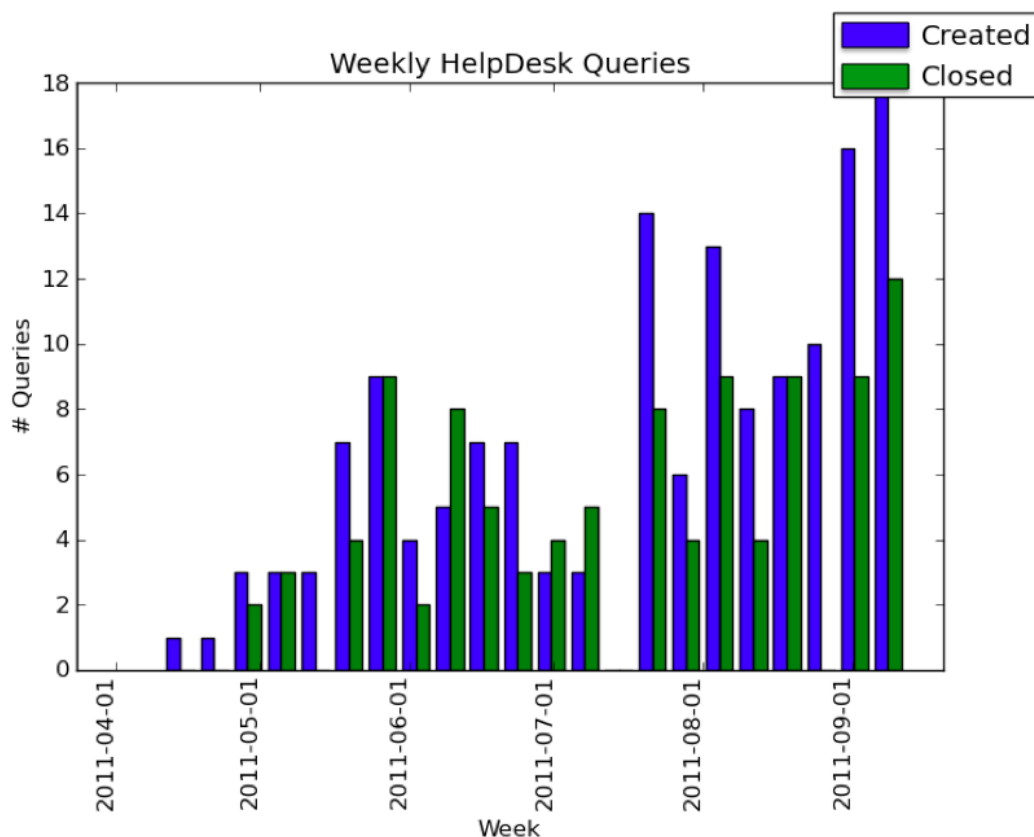


**Figure 6:** *Summary of tickets created and closed since April 2011*

## 2.8    The Future Funding of ESGF

The ESGF effort has begun to look outside the climate community for additional funding opportunities. These efforts intend to broaden the ESGF node development for other scientific domains, such as energy, material, nuclear energy, biology, chemistry, fusion, and others. This by no means indicates that we will abandon the climate community; we seek to strengthen the entire scientific community with advances that will be beneficial for all. Besides our usual trademark advances in federation, analysis, and visualization, our efforts will look to expand in the areas of data mining, provenance and metadata, high-performance computing (HPC), data movement, and data ontology. These are all challenging areas related to the management, manipulation, storage, access, analysis, and visualization of large-scale scientific data. With that said, we have come to the realization that some communities (such as climate) have a good handle on their data thanks to efforts such as ESGF. However, satisfying the need for substantial investments in data-driven software and technologies is paramount, as future computing platforms and archives expand and reach extraordinary speeds and capacity. Our goal for these communities is to deliver a comprehensive end-to-end solution for the overall increase in scientific productivity, as has been as our continual effort for climate science.

ESGF is being funded in pieces by smaller, less focused activities with some success. Some of these activities include the Climate Science for a Sustainable Energy Future (CSSEF) to transform the climate model development and testing process; the NASA ACCESS proposal to produce an ESGF node capable of delivering

NASA data products to the community; the Climate100, an Advanced Networking Initiative (ANI) research project to develop and test end-to-end capabilities of the next generation networks in collaboration with the climate community via the ESGF nodes; the Large-scale Data Systems for Nuclear Energy (LDSNE), to provide the nuclear energy community with access to data and information, model codes, analysis tools, and intercomparison capabilities for deeply analyzing nuclear energy simulations; and the NCAR effort to archive and disseminate Artic data. Internationally, small ESGF development efforts are also happening in countries such as in the U.K., Australia, Germany, France, Italy, and Japan. In addition, small business innovation research (SBIR) and other commercial efforts are funding their research and development efforts on the access and dissemination of the ESGF infrastructure and data archive.

Many challenges remain, and new ones will emerge. Through such efforts mentioned above, we hope to continue and promote the sharing of knowledge, software, and tools among partners that combines model output, observational data, and analysis and visualization tools and the facilities needed to ensure access by all user communities.

> *"The development and integration of a federated data repository from multiple sources (including fixed servers and a virtual network of servers) is a major enterprise that must be designed carefully and is recommended for future large-scale interdisciplinary scientific advances." - Dean*

# 3    Innovative Technology and Integration

## 3.1    Gateway Research & Development

### 3.1.1    Overview

Gateways are the main entry points for users to access the data and services offered by the ESG federation. Gateways allow users to browse and search for data, examine detailed metadata, download and subset files, request high level data products such as analysis and visualization, and register and apply for specific group memberships. Local administrators manage community groups at the gateway, managing user membership and access control. Additionally, gateways are federated, so that services are available to the all users seamlessly throughout the system, no matter where they first registered or where they started their working session. A gateway is typically configured, customized, and populated with resources to serve the specific needs of a project, an institution, or a community. Each gateway acts as a broker to access the data holdings served by one or more associated data nodes, because it provides infrastructure to publish and register the data, authorize user access, expose the data services, and (in the case of higher level data products) to formulate and submit the data request to the data node on behalf of the user.
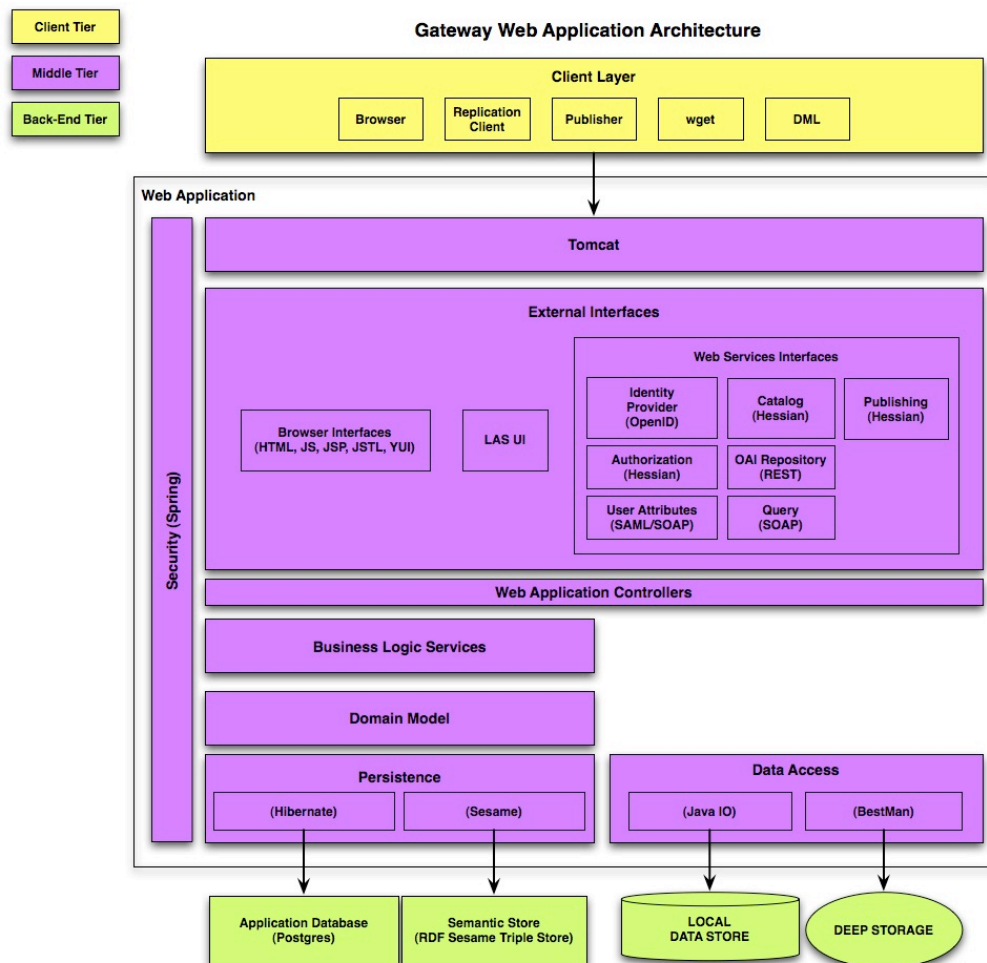
**Figure 7:** *The ESG gateway provides a suite of services for petascale scientific data management accessible from common tools using well-established protocols.*

A gateway is structured as a multi-layered web application (see Figure 7). Metadata persists in a relational database and an RDF triple store. A stack of data access, service, and control layers exposes gateway functionality to users and clients either as a suite of user interfaces or as web service endpoints (via Hessian, SOAP, or REST protocols). In particular, a gateway acts as an OAI-PMH Provider and Harvester to enable exchange of metadata records with other gateways and partner data centers. Also, a gateway may act as its own Identity Provider for web-browser user authentication, or it may be connected to an external Identity Provider. A MyProxy server allows users to obtain short-lived digital certificates that may be used by client applications to directly access data holdings from the data nodes in the system. The current ESG-CET federation is composed of eight gateways, including those at NCAR and PCMDI. This operational federation of gateways provides critical infrastructure serving a rapidly growing and diverse community of users including hundreds of scientists in the IPCC AR5 and CMIP5 efforts (CMIP5 data volume will be 2–10 PB) as well as regional model product communities such as NARCCAP. Several stand-alone gateways are also in production serving specific communities and projects including CADIS and Curator. The gateway version 1.3.2 was released in August 2011 and is in production use. A gateway 2.0-beta release was made available as a community preview in August 2011, intended for production readiness in the near future. The gateway software source code is Open Source under the Apache2 license.

### *3.1.2 Gateway User Interface*

Each gateway within the ESGF provides a web-accessible user enabling user interaction with the services provided by the system interface (see Figure 8). The gateway user interface allows end users to search and discover scientific data from the entire federated system, browse data collection hierarchies from a discipline specific gateway, download data collection files individually or in bulk, sub-set collection files, visualize data collections, track deep storage file download requests, and access user profile information.
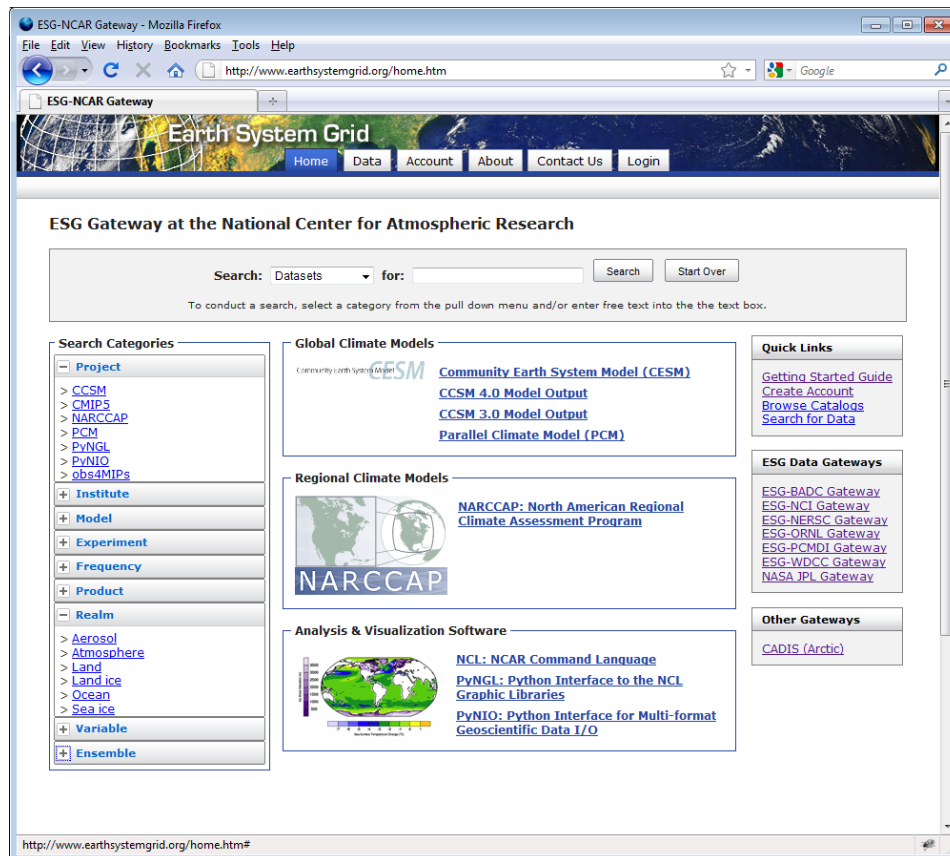


**Figure 8:** *The ESG gateway User Interface allows users to access a global collection of scientific data from a federated portal.*

User interface components have been developed for the key functional areas listed below. Significant refinement has been achieved in each component area based on production use of the system and user feedback. The design goals of the gateway user interface include exposing the base system functionality, providing a consistent and intuitive user experience, and supporting a flexible and maintainable framework for future enhancements and revisions.

- The **home page** provides visitors with general information about the discipline-specific gateway, starting points for discovering data collections, direct access to notable data collections, important notices regarding system status, and access to login and account request functions. The home page allows a gateway to customize its format to include discipline-specific information, data browser entry points, logo images, and color palette.

- **User registration** offers a multi-step workflow for account creation, approval, and validation. The resulting account may be used to authenticate any user at any gateway in the federation.

- **User and group management** allows registered users to change account settings and request access to privileged data collections, and provides tools for group administrators to approve group requests and manage group membership.

- **Login** allows registered users to authenticate with the federated gateway system with an OpenID user identifier. Users may request their password delivered in email in case of lost credentials or make inquiries to gateway support staff.

- **Data browsing** provides a variety of ways in the data browsing component. Support for file system-like hierarchies and high level associative arrangements such as experiment and project related listings are provided.

- **Data search** is the primary data-discovery method for most users. This component provides simple and familiar text-based search as well as faceted navigation for exploration based metadata inquiry.

- **Data download** allows both the download of individual files via hyperlink and bulk downloads requests from the file listing interface using generated wget scripts and the Globus Online hosted data movement service. If data collections are restricted and under access control, the user is directed to authenticate prior to data download.

- **Data transfer** allows registered and authorized users to request and manage groups of files from deep storage systems throughout the federation. Users can access real time status reports and are notified by email when transfers are complete.

- **Data visualization and sub-setting** provides an interface for requesting charts, plots, and data sub-set downloads. Users may choose variables of interest and select sub-regions geo-spatially with an interactive map and temporally with time controls.

- **Model metadata trackback** allows detailed inspection of the metadata associated with models, model components, and model component frameworks (including model sub-components, grid specifications, scientific configuration, and model run input). This novel new capability was developed under support from NOAA's Global Interoperability Program (GIP) under the auspices of the Earth System Curator project, with heavy collaboration with the EU METAFOR project.

### 3.1.3 Gateway Milestones and Software Releases

**Table 4:** *Gateway software releases.*

| Release Version | Release Date | Milestone |
|---|---|---|
| 1.0.0 | May 6, 2010 | Replacement system for ESG-II data portal based on Spring Framework |
| 1.0.0 | May 6, 2010 | Automated database migration support using Liquibase |
| 1.0.0 | May 6, 2010 | Deployed in production as NCAR ESG gateway |
| 1.0.0 | May 6, 2010 | Data Versioning support |
| 1.1.0 | July 15, 2010 | OpenID login support |
| 1.1.0 | July 15, 2010 | Basic data set replication support |
| 1.1.0 | July 15, 2010 | Internet Explorer browser fixes |
| 1.1.0 | July 15, 2010 | Data set management improvements |
| 1.1.0 | July 15, 2010 | Database integrity improvements |

| 1.1.1 | September 7, 2010 | Login and other user interface improvements |
| 1.1.1 | September 7, 2010 | Configuration file fixes |
| 1.2.0 | February 11, 2011 | Baseline release for CMIP5 data management |
| 1.2.0 | February 11, 2011 | Federation-wide X509 certificate and OpenID based authorization |
| 1.2.0 | February 11, 2011 | Full CMIP5 DRS metadata capture |
| 1.2.0 | February 11, 2011 | Automated OAI based federation metadata exchange using RDF |
| 1.2.0 | February 11, 2011 | Basic MetaFor CIM ingestion |
| 1.2.0 | February 11, 2011 | Streamlined gateway installation and configuration |
| 1.2.0 | February 11, 2011 | MyProxyLogon WebStart application integrated into data download user interface (UI) |
| 1.2.0 | February 11, 2011 | Missing DRS attributes captured during publication |
| 1.2.0 | February 11, 2011 | Over 75 bug fixes and improvements |
| 1.3.0 | June 25, 2011 | Centralized gateway registry for federation support |
| 1.3.0 | June 25, 2011 | Improved user registration workflow |
| 1.3.0 | June 25, 2011 | Spring 3.0 framework upgrade |
| 1.3.0 | June 25, 2011 | Spring security upgrade |
| 1.3.0 | June 25, 2011 | MyProxy service improvements |
| 1.3.0 | June 25, 2011 | Data access authorization workflow fixes |
| 1.3.1 | August 1, 2011 | MetaFor CIM and Trackback improvements (CIM v1.5) |
| 1.3.1 | August 1, 2011 | REST services for data sets, ESG metadata, projects |
| 1.3.1 | August 1, 2011 | Authorization of OPeNDAP aggregations |
| 2.0.0 Beta | August 24, 2011 | Significant reimplementation of search system |
| 2.0.0 Beta | August 24, 2011 | Search user interface improvements based on user feedback |
| 2.0.0 Beta | August 24, 2011 | Improved performance and scalability |
| 1.3.2 | August 29, 2011 | OpenID registration streamlining |
| 1.3.2 | August 29, 2011 | Security fixes |

## 3.2    Data Node Research & Development

### 3.2.1    Overview

Originally, the ESG data node was developed to act as the data services back-end to the user interface provided by a gateway. In the past 12 months, the data node has been completely re-factored and evolved into the ESGF P2P node—a fully featured system that contains the whole functionality originally provided by the combined gateway and data node system. Internally, the ESGF P2P node (a.k.a., ESGF node) is composed of completely separate components, each of which can be deployed independently to provide a specific functionality (data access, search, user registration, access control, etc.). This modular approach was chosen both to allow parallel development of the various modules, and to facilitate infusion of technology into other collaborations and scientific disciplines. Depending on which components are installed, the ESGF node may assume different flavors in combination: a *Data Node*, an *Index Node*, an *Identity Provider*, and/or a *Compute Node*.

As described earlier, ESGF nodes interact with each other as peers – i.e. no node is privileged over the others, and each node can join or leave the federation independently, without affecting the working state of the other nodes. All nodes are aware of each other through the ESGF registry – a document that each node produces and continuously updates by collecting information from all other nodes. Users and rich clients can start from any

ESGF node and find, download and analyze data from any other node. This works because all nodes interoperate through the adoption of common services and APIs, which are based on industry standards such as OpenID, SSL, PKI, SAML, OPeNDAP, and Solr.
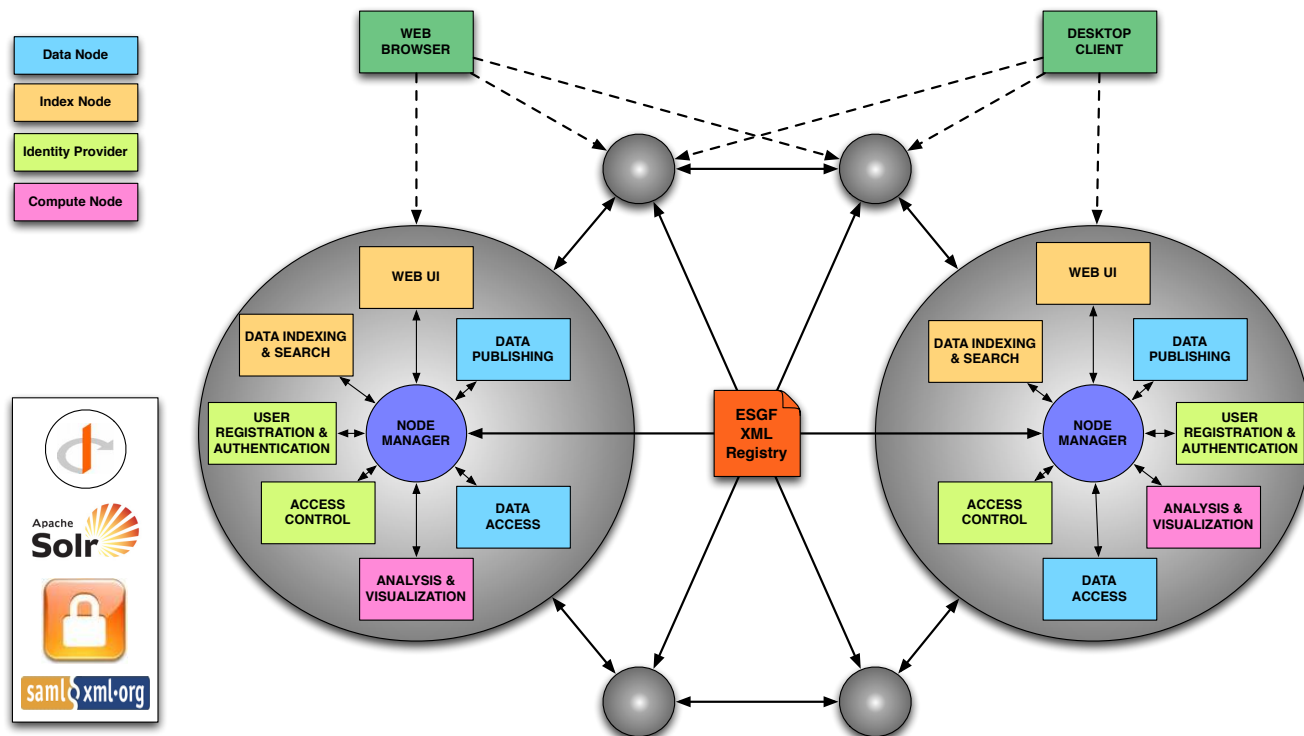


**Figure 9:** *The ESGF Peer-To-Peer architecture is based on the principles of modularity and equality. Each ESGF node can offer different services depending on how it is configured, nodes with different flavors can be scaled differently (for example, to provide increased computational resources, or failover search capabilities), and all nodes interact as equals, so there is not a single point-of-failure.*

### 3.2.2   The Node Manager

The Node Manager (indicated in blue in Figure 9) is the most fundamental building block of the ESGF application stack.  It monitors the node and houses the logic for the recording of the access traffic.  It monitors the overall "health" of the node and is the heart of the peer-to-peer system that gives ESGF its ability to simultaneously maintain organizational sovereignty and create a unified data space from which all can share and leverage. The Node Manager provides these services to the higher-level applications in the ESGF node stack.  The P2P registry is the center of internode knowledge with regards to whitelisting, routing, searching and commissioning.

The Node Manager is needed to provide the unifying layer that the data space is built on.  The Node Manager seamlessly handles nodes joining and leaving the federation and maintains a federation wide singular view of the changing topology.  The Node Manager also keeps track of information access and provides notification support. The Node Manager is what allows there to be no single point of failure in the system.  The Node Manager is also the entity that allows for massive scale out because of its ability to coordinate among a large number of nodes. One could easily build out an entire farm of nodes and handle them as easily as a single node instance.  The Node Manager also greatly reduces if not eliminates many configuration tasks encountered in the older ESG architecture.

The Node Manager is used by the higher-level applications of the ESGF node stack. The security system uses it for whitelisting and attribute finding. The security system is able to fail over to another attribute or authorization service because of the information provided by the Node Manager. The search system is able to use the Node Manager to allow it to distribute queries across the federation. The compute system will be able to use the Node Manager to query nodes in order to post work requests for distributed computation - finding out where the most storage is, where the fastest central processing units (CPUs) are, who has the most bandwidth, who has the most memory, who has the longest up time, etc. The Node Manger provides an external representation of its state as an XML document that is readable by anyone on the web. This means that services not actively participating in the distributed data space made possible by the Node Manager can still find federation-wide information and leverage it as well.

With the Node Manager, you can create a highly scalable, distributed, data space, spanning geographical locations. Also, you can provide smart replication and content distribution. The Node Manager provides the basic interconnectivity of nodes that higher-level applications can leverage. That is, before there was a central entity that had to be communicated with, if that central entity failed the system failed or would have to be reconfigured manually to recover. Also in the past we had to send information through some sort of out-of-band process so an administrator could put the information into a registration database and post that information on a known central location. And finally, before we could not find out when nodes were up and when nodes were down, we could not set up mechanisms to track this information to build reliability metrics for nodes. With the Node Manager we can do all of these more efficiently and gracefully in a timely manner throughout the P2P system.

The Node Manager state is ephemeral so when the system goes down or something happens it is still fine. The information about the rest of the network is re-learned by its peers. There is no intrinsic state that is required other than having at least one 'seed' peer's address that is durably written to file in the configuration. Any information or statistics gathered is durably written to a file or database and easily reconstructed if needed. Thus, providing the Node Manager with fail over safety features for ease of system crashes and recoveries.

There is some tuning that will need to be done to the protocol. This work is ongoing as systems change and data archives expand. There is also a fair bit of self-tuning that can be programmed into the system so that it may learn the characteristics of the federation. These are future developmental work that will allow better federation among nodes.

### 3.2.3 The Security Services

Several ESGF institutions (ANL, NASA/JPL, LLNL/PCMDI) have worked in strict collaboration with their European partners (BADC, DKRZ) to define and implement a state-of-the-art security infrastructure, which allows secure access to distributed data holdings that are served by generic application servers written in either Java or Python. Because this infrastructure is based on a rigorous API definition, and on the adoption of industry standards such as SSL, PKI, OpenID and SAML, it allows clients and servers written in different languages to communicate with each other, within a global distributed environment where access control is decentralized and managed by separate institutions. For example, a Python script running on a NASA/JPL machine can securely request data sets from a Java THREDDS Data Server at LLNL/PCMDI, or a client application based on the NetCDF Java library running on a scientist's laptop can access data from a BADC PyDAP server.

The ESGF security infrastructure is based on several modular components that can be deployed independently or in concert:

- **Identity Provider** allows users to use their OpenID to login at any node in the federation, by being redirected to enter the password at the one node where they originally registered.

- **Attribute Service** is responsible for delivering user information to a trusted client (a minimal set of personal information, and most importantly the user's access control attributes), encoded as a digitally signed SAML statement.

- **Authorization Service** is responsible for executing authorization decision about the right of a user to access a given resource, and delivering this information to a trusted client in the form of a signed SAML statement.

- **OpenID Relying Party** is invoked as part of a data access request sent to a data node, and fulfills three different purposes:

  o In the case of programmatic data access, it can validate a client certificate and establish a session-scope authentication cookie.

  o In the case of web-browser access, it provides an interface where a user can enter their OpenID for authenticating at their *Identity Provider*, eventually establishing an authentication cookie.

  o In the case a user has authenticated, but lacks the proper access control privileges to access a given data collection, it exposes a user interface to register the user with the proper remote registration service

- **Access Control Filters** represent a composable framework for securing a generic Java servlet application, by front-ending the application with servlet filters that intercept a data request, and implement a configurable logic for authenticating and authorizing the request (redirecting to the OpenID Relying Party, or invoking one or more Authorization Services).

Figure 10 shows an example data access workflow, which illustrates the role played by the different components in authenticating and authorizing the request.
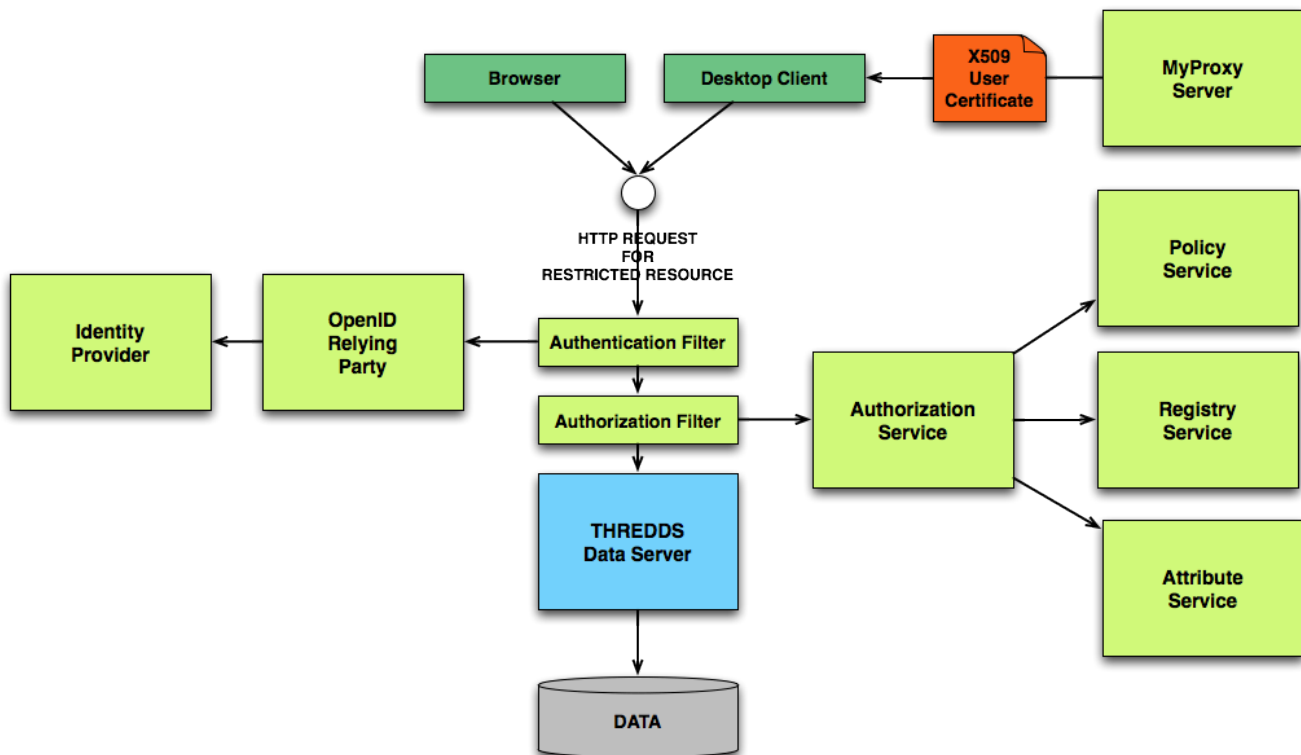


**Figure 10:** *Example of secure data access workflow. The user can utilize a browser or a rich desktop client to issue a request for data served by a TDS. The request is intercepted by the Access Control Filters, which establish authentication (by*

*validating the client certificate, or redirecting the user to the OpenID Relying Party) and authorization (by contacting a chain of Authorization Services).*

### 3.2.4 The Search Services

One of the key functionalities offered by an ESGF node is the capability to index metadata from multiple sources, and to expose its contents via web-accessible search services. A search client (either a web browser, or a rich desktop client such as UV-CDAT or DML) can query the metadata throughout the federation via a simple RESTful API – i.e. a self-describing, URL-based syntax. The API exposes a full-featured faceted search, as well as powerful text-based query syntax. For example, a client could ask for all the files in the system that were generated across all models to conform to a given CMIP5 scenario, contain a certain standard variable, and have monthly frequency.

Internally, the ESGF Search Services utilize the Apache Solr engine to store and query metadata. Solr is a reliable, scalable application, which is very popular in industry, and offers faceted search capabilities, as well as powerful text searches based on the underlying Lucene libraries. Solr is based on a flat model of (key, value) pairs of metadata units, so the ESGF infrastructure includes an extensible framework for parsing metadata documents conforming to multiple schemas (e.g., THREDDS, Directory Interchange Format (DIF), Global Change Master Directory (GCMD), etc.) into the Solr index. One of the big advantages of using Solr (besides performance and scalability) is that the metadata fields don't need to be pre-defined: as new search categories are defined by specific projects and entered into the metadata catalogs, they are automatically entered into the index and exposed by the API for faceted searching.

The search capabilities of the ESGF system augmented by the ESGF Node Manager allows new search (*index*) nodes that join the federation to immediately have their data be a part of the global (federation wide) *data space* search corpus without any configuration necessary from the node administrator.

### 3.2.5 Peer-to-Peer Web Front End

As the demand for robust and consistent scientific data distribution platforms increase, user interface (UI) design and implementation have become one of the more important tasks in the creation of any scientific data portal. Potential data consumers, and specifically end users of the ESGF node portal, are inevitably concerned with the manner in which services (e.g. node management, search, etc.) are presented.  In general, end users highly value *usability*.  An oft-studied, yet rarely perfected concept, usability requires careful consideration in designing interfaces that are easily navigable, flexible, and intuitive.  Moreover, it is imperative to conceive a design amenable to the evolving architecture and services offered by ESGF.   As a result, the ESGF P2P effort altered the construction of the current gateway to accommodate this changing infrastructure while maintaining optimal usability. Figure 11 illustrate the look and feel of the ESGF P2P front end design.

The ESGF P2P team has incorporated a number of components to the front end design to accommodate a wider community of potential users that may be interested in various data sets (e.g. scientists interested in model and observational data for comparison).   Like most modern search portals, text search has been enhanced with autocomplete capabilities to aid users unsure of specific search strings.  The temporal search tool leverages the highly-effective range search capabilities of the Apache Solr search backend, allowing users to extract data sets according to ranges constrained by time measurements via Solr's Lucerne based inverted-index technology. The faceted-based category navigation tools were expanded to include flexible terminology (e.g. "instruments" in observational terminology) while providing direct support for the structured terminology of CMIP-5.  Finally, the new design leverages the popular Google Maps engine to allow users to conduct searches over geospatial range criteria as illustrated in Figure 12.

Perhaps the most important design consideration, however, is the development of a search environment that leverages recent advances in web technology to promote efficiency and usability.  In this scheme, the ESGF P2P team created a single, intuitive search page, illustrated in Figure 13, which heavily utilizes the JQuery javascript framework and the emerging Ajax-style development pattern.  Data consumers need only visit *one composite*

*page*, where they pick and choose the components they would like to use to aid in searching for available data sets. Search components are encased in overlays, while search state is presented to the user in a clear and concise manner. This configuration has several inherent advantages. The overhead of clumsy "back and forth" static page navigation is eliminated. Full state of a specific user's search constraints is persisted to HTML5 local storage and presented to the user at any time throughout the duration of a search session. Consumers can also add search constraints in parallel. For example, a user can perform a geospatial range query while simultaneously viewing variables measured in that associated region. Finally, a simple data cart system for downloading data sets was introduced. A user needs to navigate only to the data cart tab to add, remove and view data sets chosen for download.



**Figure 11:** *The ESGF node installation script deploys the home page of a minimally configured P2P web front-end application, as it is shown. Included on the home page are search capabilities, information section pertaining to the node and organization, quick links and links to other ESGF P2P nodes, and analysis tools.*
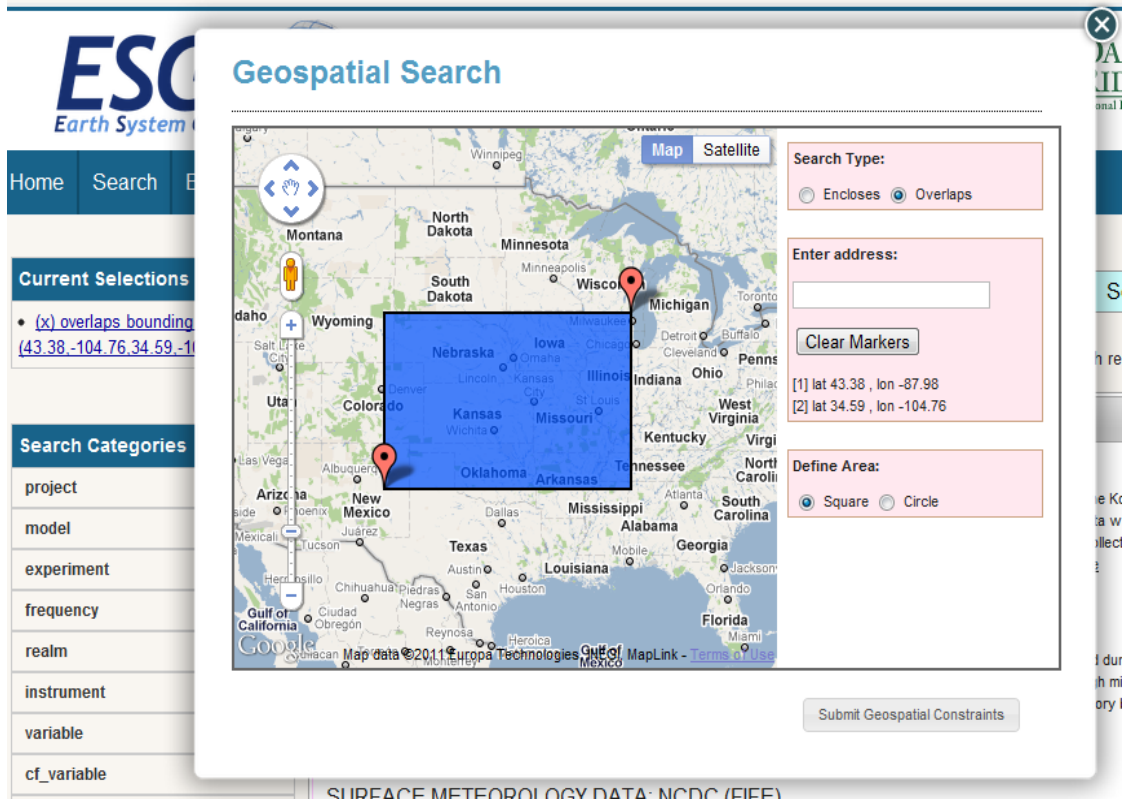
**Figure 12:** Faceted and geospatial based search/discovery



**Figure 13:** *Intuitive search and discovery*

### 3.2.6 *Data Publisher*

The Data Publisher is the component of the ESGF Data Node software that ingests data and metadata into the ESGF catalogs. This publication process involves several steps, including scanning data files and ingesting the information into the node database, gathering additional metadata not included in the data files, performing basic quality control, summarizing the metadata in one or more XML catalogs, notifying other node components that new catalogs have been created, and notifying the ESG gateway or ESGF index node that new data sets have been added.

For security, the Publisher supports SSL client authentication using short-lived X.509 certificates. The MyProxy credential management service is used to generate and manage certificates. We offer a number of data publication management utilities to perform MyProxy authentication, data-set queries, data-set publication, data set modification, and data-set deletion. In addition, we have developed an easy publication GUI (see Figures 14 and 15) that combines those functions into one convenient tool.
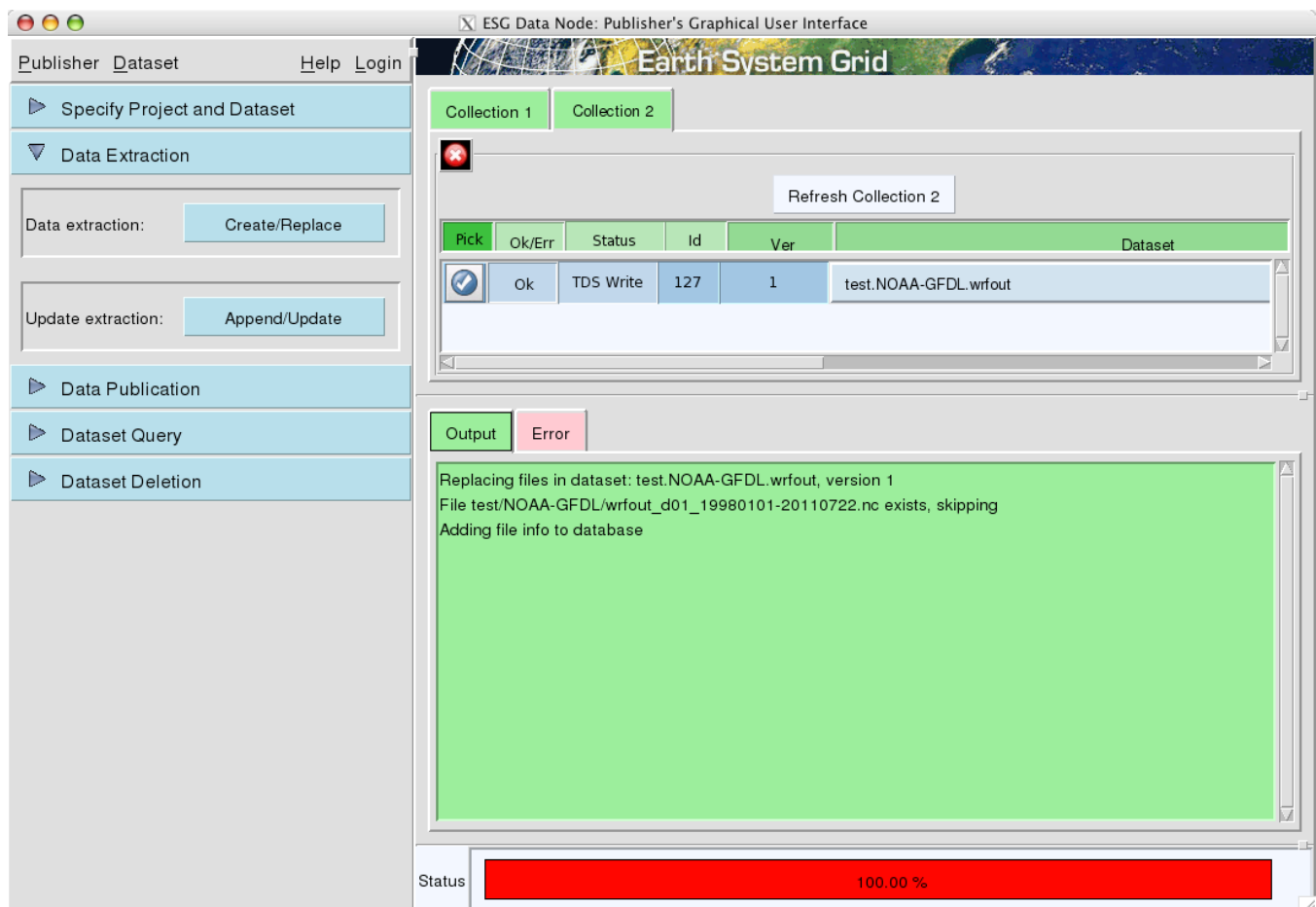


**Figure 14**: *A screen shot of the Publisher GUI displaying a NOAA GFDL data set that was published to the THREDDS catalog server.*
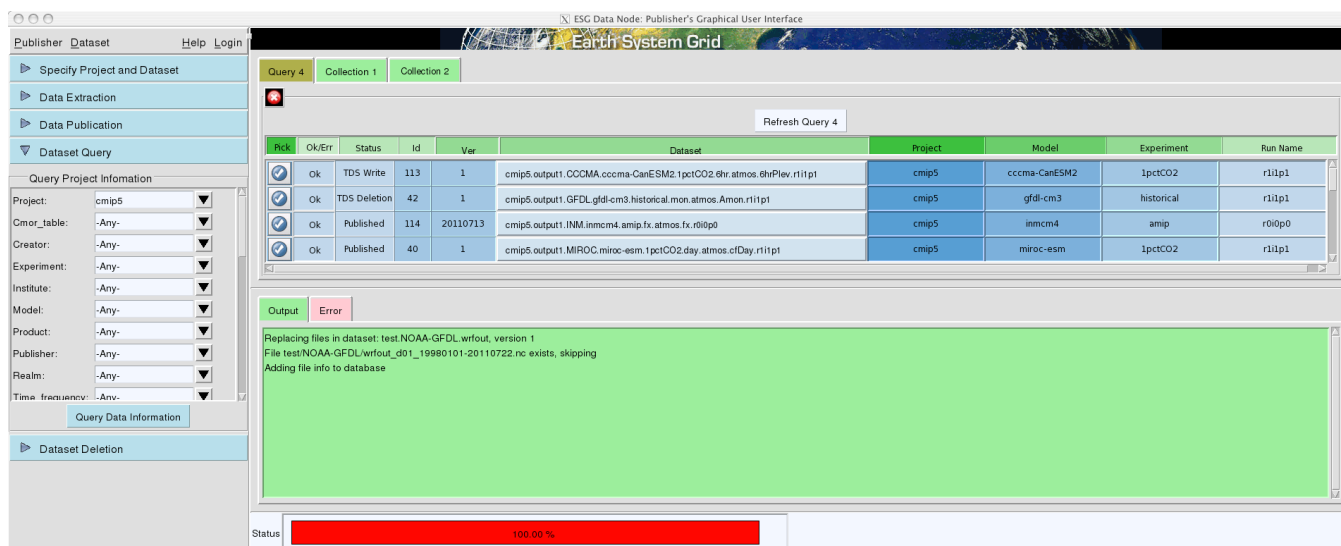
**Figure 15***: The Publisher GUI is used to conveniently perform a data set query*

For more information regarding the Publisher, please refer to:

- CMIP5 Best Practices for Data Publication, http://esg-pcmdi.llnl.gov/internal/esg-data-node-documentation/cmip5-best-practices

- Introduction to ESGF Data Node Configuration and Publishing, http://esg-pcmdi.llnl.gov/internal/esg-data-node-documentation/introduction-to-esg-data-node-configuration

- CMIP5 Controlled Vocabulary, http://esg-pcmdi.llnl.gov/internal/esg-data-node-documentation/cmip5_controlled_vocab.txt/view

- Customizing the ESGF Publisher, http://esg-pcmdi.llnl.gov/internal/esg-data-node-documentation/customizing-the-esg-publisher-with-handlers

- CMIP5 PCMDI Publisher Checklist, http://esg-pcmdi.llnl.gov/internal/esg-data-node-documentation/cmip5-pcmdi-publisher-checklist

- ESGF Publication GUI, http://www2-pcmdi.llnl.gov/Members/bdrach/.personal/esg-publication-gui/

- Publisher Tutorial, http://esg-pcmdi.llnl.gov/internal/esg-data-node-documentation/CMIP5_publication_tutorial.ppt/view

### *3.2.7 Client Analysis Tool Direct Access to the ESGF Archive*

Funded under BER in support of its science mission, the UV-CDAT framework is designed to integrate five analytical and visualization tools—CDAT, VisTrails, ParaView, VisIt, and R—under one application. Based on Python, it links disparate software subsystems and packages to form an integrated environment for analysis. UV-CDAT's design and openness permit the shared development of climate-related software by the collaborative climate community. In particular, the goals of the UV-CDAT project are to (1) prepare for the CMIP5 data archive and assessment process by developing derived data products and user-reproducible workflows and analysis archives; (2) develop capabilities to inter-compare ungridded observational data sets and model data for validation; (3) deliver efficient scalable analyses and visualization for high-resolution simulation data; (4) deliver data products in formats suitable for expert and non-expert users; and (5) build all capabilities on existing ESGF

node infrastructures. Figure 16 shows the new interactive UV-CDAT graphical user interface (GUI) in the background and the UV-CDAT ESGF search and browse GUI. This allows users to search and browse from UV-CDAT analysis tool as if they were on a web browser. Once data is located, it is downloaded directly to the UV-CDAT application.
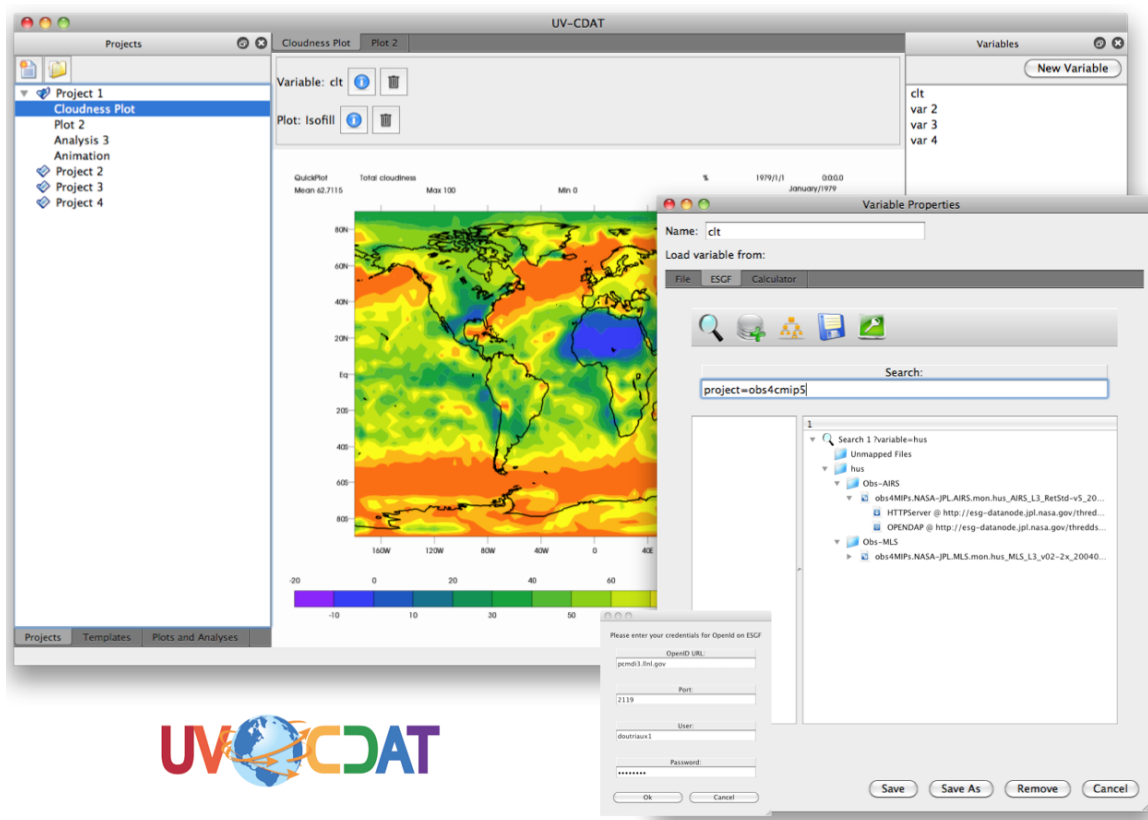


**Figure 16:** *UV-CDAT searches and accesses the ESGF node archive at LLNL/PCMDI and displays the results*

### 3.2.8   Installation Script

The installation script creates an entire ESGF node installation when given little more than the base operating system to start with. It is responsible for downloading, configuring, and coordinating all the elements of the ESGF application stack to produce the resulting amalgam that is the ESGF node.

The installation is achieved not by a singular script but by a series of scripts; each sub-script is responsible for the installation of its respective sub-module. Taking inspiration from the UNIX boot process, the main installation script coordinates the sub-scripts and when executed in concert culminates into the full installation. The installer checks the integrity of all files it downloaded to ensure they have not been tampered with. The install script, just like the UNIX operating system scripts, is written in the BASH scripting language and thus takes advantage of all the powerful UNIX tools when performing its tasks. Required files are downloaded, configuration files are edited, backups are created and programs are initialized resulting in the ESGF node. In the case of the ESGF node, the gestalt is truly greater than its parts.

The installation script is a necessary tool for building the ESGF node, which is embodied by the complete host of the node, and it can be thought of as an "image creator" for building Virtual Machines that encapsulate the entire ESGF node stack. This future goal will make setting up a node very easy and instant and immediately be able to take advantage of large cluster scale out physical infrastructures

directly. Because of the Node Manager, the ESGF node deployment scenario allows us to build instant clouds from ESGF nodes.

**Table 5**: Software comprising the ESGF Peer-to-Peer node software stack.

| Application | Name | Version Number |
|---|---|---|
| Node (Web Application) | ESGF Node Software Stack | 1.0.4.0 |
| Language | Java | 1.6.0.23 |
| Language | Python | 2.6.5.0 |
| Custom Component | Node Manager | 1.0.0.4 |
| Custom Component | Publisher | 2.8.2.0 |
| Custom Component | Versioning (Publisher) | 1.0.0.0 |
| Custom Component | Replication (Publisher) | 1.0.0.0 |
| Custom Component | Bulk Data Mover (Replication) | 1.0.0.0 |
| Custom Component | Product Server (LAS) | 7.1.2.0 |
| Custom Component | ESGF Search Services | 1.4.0 |
| Custom Component | Web UI Front-end | 0.9.6 |
| Custom Component | Installation | 1.5.0 |
| Security | SAML | 1.1.3.0 |
| Security | MyProxy | 4.9.0.0 |
| Security | Globus | 4.2.1.0 |
| Security | ESGF Security Services | 1.4.1 |
| Security | OpenID Relying Party | 1.2.2 |
| Security | Identity Provider | 1.1.6 |
| Analysis Tool | Ferret | 6.6.4.0 |
| Analysis Tool | UV-CDAT | 6.0.0 alpha |
| 3rd-Party Library | HDF5 | 1.8.4-patch1 |
| 3rd-Party Library | NetCDF | 4.1.1 |
| 3rd-Party Library | Readline | 6.1.0.0 |
| 3rd-Party Library | Xml2 | 2.7.6.0 |
| 3rd-Party Library | Xslt | 1.1.26.0 |
| 3rd-Party Library | Flex | 2.5.35.0 |
| 3rd-Party Library | NumPy | 1.4.1.0 |
| 3rd-Party Library | Pmw | 1.3.0.0 |
| 3rd-Party Library | PyOpenGL | 3.0.1.0 |
| 3rd-Party Library | Setuptools | 0.6c11 |
| 3rd-Party Library | Bison | 2.4.0.0 |
| 3rd-Party Library | Pam_Pgsql | 0.7.0.0 |
| 3rd-Party Library | Mhash | 0.9.9.9 |
| 3rd-Party Library | gSoap | 2.7.15 |
| 3rd-Party Software | Ant | 1.8.2.0 |
| 3rd-Party Software | Git | 1.7.3.0 |
| 3rd-Party Software | Curl | 7.20.1.0 |
| 3rd-Party Software | Postgres | 8.4.7.0 |

| 3rd-Party Software | Tomcat | 6.0.32 |
|---|---|---|
| 3rd-Party Software | THREDDS | 4.2.2.0 |
| 3rd-Party Software | GridFTP | 3.24.0.0 |
| 3rd-Party Software | Tcl/Tk | 8.5.8.0 |
| 3rd-Party Software | Solr | 3.1.0 |
| 3rd-Party Software | Spring | 3.0.5 |
| 3rd-Party Software | Google Map | 3.0 |
| 3rd-Party Software | JQuery | 1.4.2 |
| 3rd-Party Software | Apache Tiles | 2.1.4 |
| 3rd-Party Software | Ivy | 2.2.0 |
| 3rd-Party Software | Bash | 4.0.0 |

## 3.3 Data Movement and Storage Management

### 3.3.1 GridFTP and Globus Online for data replication and download

GridFTP servers are provided as download options in both architectures. In both cases the node classified as the Data Node, that serves the data, hosts the GridFTP server to provide access to the data. Globus Toolkit GridFTP server is distributed with the data node software, and two different configuration of the server is used, one for data download by end user and another for replication of data between sites. In all cases the GridFTP server is configured to trust CAs agreed by the ESG federation, ensure user is authorized to perform the transfer operation, record metrics of the operation, and restrict the server to read from a defined data directory.

GridFTP for end user download, is configured with custom security handlers, that allow use of ESG security services. User authentication is via use of X.509 certificate and PKI security, and a custom handler calls out to the ESG Authorization Service to determine if a user can download a particular file. The use of PKI and authorization callout, along with a callout that records metrics of the download, the GridFTP is customized to the needs of the federation.

For replication, the ESGF policies are governed by the presence of a user in a privileged group. Replication requests are expected to be made only by site administrators or their delegates, and each of those users are placed in a special group for bulk data movement. The authorization callout does not validate each file transfer, but it establishes that the user requesting transfer is a member of the privilege group, and allows access to all files. This reduces the overhead involved in validating access permissions for each file in a bulk data movement scenario.

While the GridFTP servers provide the server side of the transfer, Globus Online provides a client solution for data transfer.

Globus Online is a data transfer management software-as-a-service provided by ANL and University of Chicago that builds on the GridFTP data transfer mechanism and provides reliable and high- performance data transfer solutions. Globus Online lets users "fire-and-forget" their transfer request and the service will manage the entire operation—monitoring performance, retrying failed transfers, recovering from faults automatically when possible, and reporting request status. Further, Globus Online auto-tunes the transfer to ensure best performance based on the transfer size and the number of files per transfer.

ESG can outsource data transfer management to Globus Online and focus on domain-specific tools and solutions for climate science research. Some of the concepts and solutions for the integration were demonstrated at the SC10 conference.

Globus Online is provided as one of the options in the ESG gateway for downloading selected data. The user experience includes user login, search and selection of data to download, and a series of UI-driven steps to submit the transfer request to Globus Online. The transfer uses the user's credentials from an ESG MyProxy server, and

allows users to either download to their local machines or transfer to any other machine with a GridFTP server. This work has been merged with the Gateway source code, and is slated for release in the next production release.
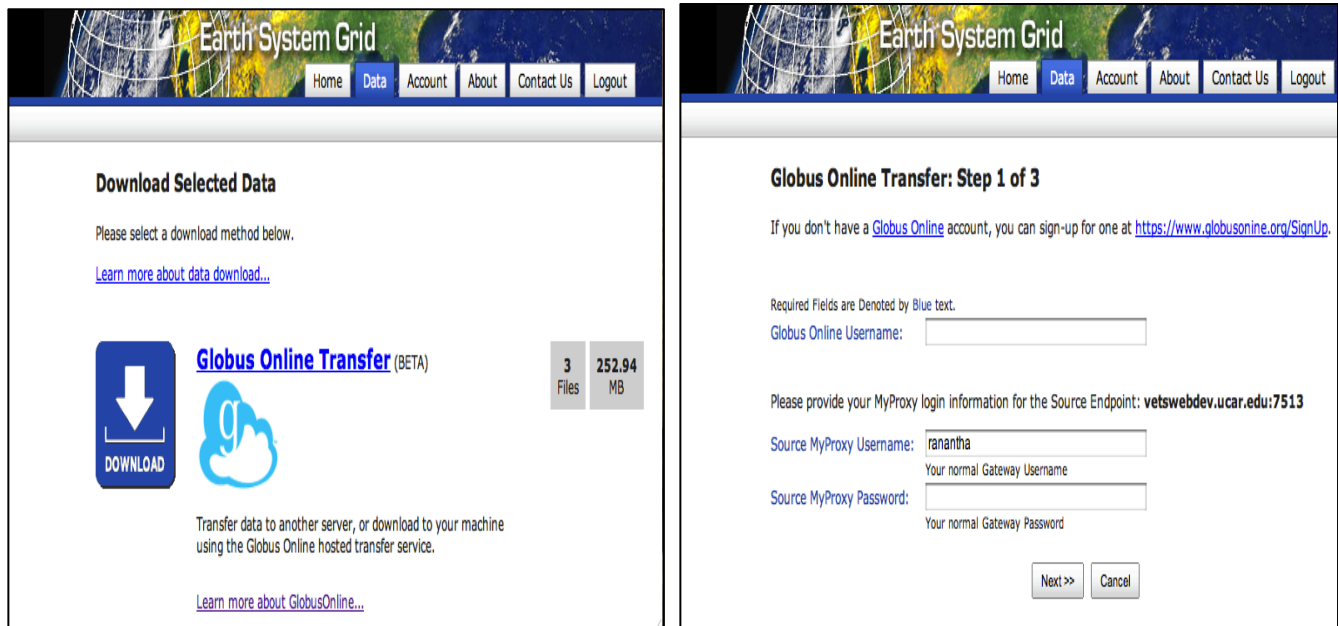


**Figure 17:** *Globus Online for file download in ESG gateway*

A similar effort to leverage Globus Online for replication use cases is underway. This work is focused on using the Secure Shell (SSH)-based client interface to Globus Online for automating the replication process by administrators. The replication tool has been enhanced to support Globus Online as a transfer option, and some testing of this option has been completed by the European partners.

The above efforts have resulted in ANL developing a Java client library for the Globus Online Transfer API, which can be leveraged by other ESGF tools and services to integrate data transfer functionality. In collaboration with other ESGF partners, these features will be integrated with other search and download tools and UV-CDAT in the coming months.

### 3.3.2 Berkeley Storage Manager (BeStMan) and Storage Resource Management

Storage management is one of the most important enabling technologies for large-scale scientific investigations. The ESG-CET project must access data from distributed heterogeneous storage systems, including Mass Storage Systems (MSSs), move data efficiently between storage systems, and manage the content of data in the storage spaces. Modern computational facilities, featuring ever faster computational platforms and increasingly large and complex storage capabilities, produce rapidly growing data volumes. Managing these volumes, especially the extraordinarily large data sets held in federated distributed centers, requires sophisticated Storage Resource Managers (SRMs) [SRM07, Shoshani02, Shoshani03], and software components whose function is to provide dynamic space allocation and file management on shared storage systems. An SRM invokes data-transport services to bring files into its data space transparently, providing effective sharing of files.

Designed to request multiple files concurrently, taking advantage of the available bandwidth, SRMs are based on a common specification, developed by the LBNL team members, which has evolved over time into a standard specification under an international collaboration. Having an open-standard specification that can be used by individual institutions to adapt to their own storage systems has proven to be a remarkable success. SRMs have

been used by several facilities, providing uniform interfaces to various storage systems at multiple sites both in the US and in Europe [SRM09, OGF09].

In ESG-CET, LBNL's implementation of an SRM, called the Berkeley Storage Manager (BeStMan) [MSS2007, BESTMAN, SSDBM04] provides a front-end for various disk-based systems and mass-storage systems such as HPSS (High-Performance Storage System) at NCAR, ORNL and LBNL/NERSC. This arrangement makes it possible for the gateway site to access all the remote-data sources through a uniform interface. Furthermore, the gateway uses BeStMan to manage its disk pool for providing file access to users. Because an ESG-CET gateway's disk is used to stage files temporarily for access by users, managing the shared disk on the gateway is necessary. This functionality includes managing space quotas for users, setting lifetimes for the files, and cleaning of the disk space after files are downloaded to the users if the space is needed (termed "garbage collection").

BeStMan also provides support for "streaming" files through the BeStMan disk-cache space. This capability permits a client to make a request for files that requires more space than can be allocated at one time by the gateway. Once a request is made, BeStMan brings as many files as will fit into the client's quota space. If some files are already in BeStMan's disk cache (for instance, they were brought in for another client previously), they are made available immediately. The client can now get the available files and "release" them when finished. For every file released BeStMan brings in another file not yet accessed, thus "streaming" the files to the client.

In addition, an adaptation of BeStMan was developed for locally developed MSS system at NCAR, making it accessible on the grid. The BeStMan provides secure access to the local MSS system by customization to their local security policies and through a plugin interface and uses community certificates to provide uniform access on the grid, as well as to invoke transfer services, such as GridFTP. Similarly, ORNL's BeStMan has a plug-in mechanism for two-layer site security, and the ORNL ESG team developed a plug-in that combines BeStMan in the external security zone and the HPSS data-fetching program in the internal security zone through a secure communication channel. This security configuration was successfully deployed and used over the last few years.

BeStMan servers for the three sites mentioned above also have been used by the ESG-CET Metadata Publishing tool for browsing all file information for files stored on deep storage (i.e. tape-based mass storage systems) to populate the metadata catalogues at the gateway.

The adaptation, deployment, and support of BeStMan in the ESG-CET project have been accomplished seamlessly. BeStMan is under an open source license, BSD with Grant-back provision (LBNL CR-2404), and available as sources and binary packages on https://codeforge.lbl.gov/projects/bestman/.

### 3.3.3   Addressing ESG Deep Archive Challenges with BeStMan Enhancements

The High Performance Storage System (HPSS) deployed at ORNL is capable of storing multi-petabytes of data for long term archival purposes, while Earth Grid System (ESG)'s online disk capacity at ORNL is limited to hundreds of terabytes.  Modeling activities at ORNL are projected to exceed the capacity of our disk cache and require the use of our HPSS archive.  While accessing deep archives from ESG usually only requires the use of the HPSS storage resource manager (SRM) within BestMan, ORNL's multi-level security environment made this standard approach impossible.

ORNL's HPSS archive is in a medium security enclave requiring two-factor authentication while the ESG gateway is in a low-security zone that does not require two-factor authentication. Systems in ORNL's medium security enclave are not allowed to have any ports open to the public Internet. Another challenge in this configuration was limiting the data that could be accessed via ESG from the ORNL HPSS archive so that only data published to ESG could be accessed by the ESG HPSS low-security data node.

In addition to the challenges imposed by ORNL's security environment, enhancements to the ESG data publication process were required to provide seamless access to data published to ESG that resided on ORNL's deep-archive.

Prior to ORNL's enhancements, data published to ESG that was served through the HPSS SRM captured and displayed only file name, file size, file location, and other trivial attributes when compared to data published on

local disk which included virtually all properties of the netCDF file (variable names, variable units, etc). ORNL's enhancements to the publication process bridged this gap, allowing data published via the HPSS SRM to harvest all properties of the underlying data set on the ESG gateway.

### 3.3.3.1  ESG HPSS Access: Design and Implementation

The "standard" ESG way of integrating with SRM is to have a data node with direct access to HPSS (or other mass storage). The data node runs BestMan (or some other SRM implementation) and the ESG Gateway knows the URL's of the files that BestMan can serve. The user selects the files from the list on the gateway, the gateway returns a list of URL's and the user uses SRM client software to retrieve those URL's from the data node.

We can't use the standard implementation because that would expose all of ORNL's HPSS archive to additional attack vectors from the public internet. **Figure**  illustrates how we solved this problem with a flexible client-server architecture: one low security data node server (L-DN) residing within the low security zone and another medium security data node server (M-DN) residing within the medium security enclave. The medium security server can communicate with HPSS, but is isolated from the public Internet. The low security server is on the public Internet and provides a place for ESG clients to connect to and download files; it cannot however, communicate directly to ORNL's HPSS infrastructure.
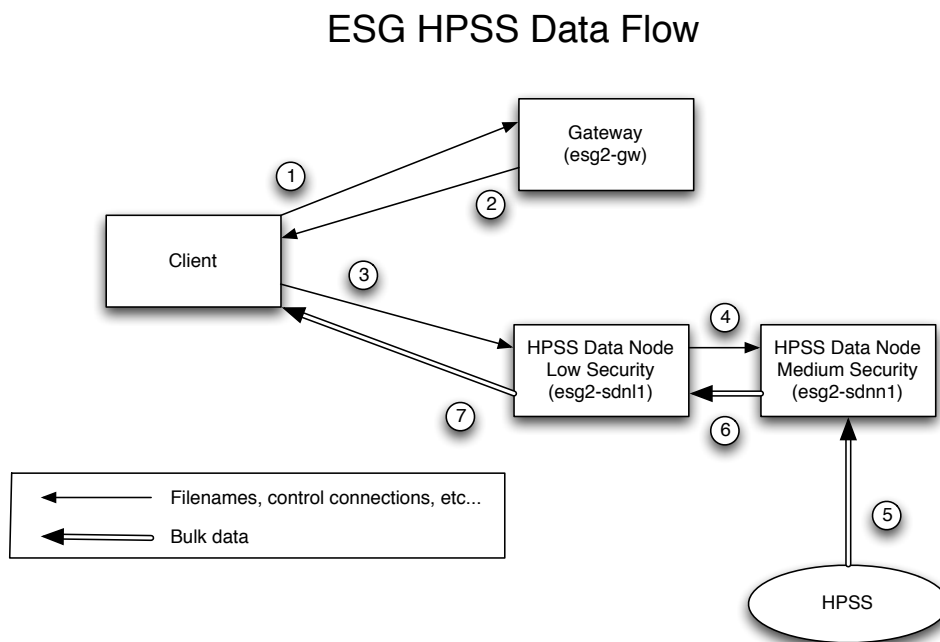
## ESG HPSS Data Flow



**Figure 18:** *ESG HPSS data flow at ORNL*

To enable this dual server design for HPSS access and enhanced data publishing for data set on deep-archive, we implemented the following components and customizations:
- New Lister program that can correctly parse the SRM output of NCCS.
- New plugin module that leverages the data node stack extension capability.
- New BestMan plugin for handling SRM copy request from the Gateway and interacting with the backend server
- New backend server that runs as daemon process on the medium security enclave, which handles SRM

    request from BestMan, interact with HPSS server for local data transfer, and invoke GridFTP client for remote data transfer.

- We also define a set of XML-RPC based API calls as the communication protocol between the BestMan plugin and backend server.
- Whitelist support on the backend server that verifies every request coming from the low security zone against an authorized published file list.

With these new components in place, ORNL has developed an end-to-end infrastructure that provides access to data sets within ORNL's HPSS archive while providing rich metadata similar to that of data sets published from local disk. **Figure** illustrated data sets available on ORNL's ESG gateway resident on ORNL's HPSS archive. These data sets were published using ORNL's enhanced data-publication mechanism. **Figure** illustrates the rich metadata that is made available to the user via this enhanced data-publication mechanism. Metadata for individual data sets show no difference for data resident on deep-archive to data resident on local disk.

Users may request a data set for download which generates a request to the L-DN for the data set. The M-DN queries the L-DN periodically for these requests and when a new request is discovered the M-DN begins processing the request for the data set. The M-DN checks the request against a whitelist and then transfers the data set to local disk-cache . The M-DN then invokes a GridFTP transfer of the data set from local disk-cache to the L-DN. The M-DN then notifies the L-DN of completion of the request. Once the data set is available on the L-DN the user is notified that the data set is available for direct download as illustrated in **Figure**

### File Download Selection

**ornl.arm.data.hpsstest**
2 File(s)                                                      Download ALL Selected File(s)

| | File | Size | Format | Location | Direct Download |
|---|---|---|---|---|---|
| ☐ | agnpp_C-LAMP_L1.nc | 21 MB | NetCDF | SRM | |
| ☐ | gpp_C-LAMP_L1.nc | 21 MB | NetCDF | SRM | |
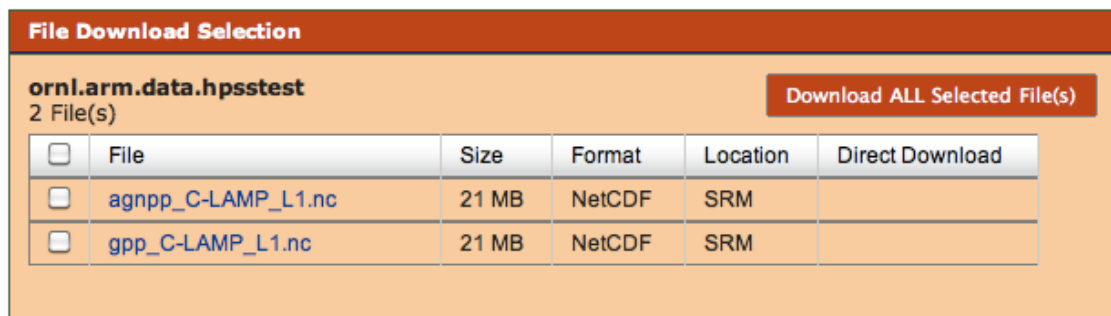
**Figure 19:** *Data sets published to ORNL's ESG Gateway from deep-archive*

**Figure 20:** *Full-metadata support for data sets published from deep-archive including descriptions for each variable in the netCDF file*



**Figure 21** *Data set on deep-archive available for download via ORNL's BestMan plugin and end-to-end infrastructure*

### 3.3.4   DataMover-Lite (DML)

DataMover-Lite (DML) is an ESGF-specific, file download tool with an intuitive graphical user interface that works with ESGF services from the gateways and ESGF data nodes and allows a user to easily request a large

collection of files for transfer to their system including files in deep storage. DML's features and functionality are as follows:

- DML is under an open-source license, BSD with Grant-back provision (LBNL CR-3004), and available as sources and binary packages on https://codeforge.lbl.gov/projects/dml/. A Webstart version is available on http://datagrid.lbl.gov/dml4/jnlp/dml.html. The DML webstart package is also available for any ESG gateway that wishes to host DML Webstart.

- Authentication and authorization: DML supports authorization using SAML (Security Assertion Markup Language) certificates and manages ESGF credentials during the file transfers using OpenID login.

- Data transfer and streaming: DML works with multiple transfer protocols such as HTTP, HTTPS and GridFTP supported by ESG data servers, including parallel streaming capability for both HTTP and GridFTP. It supports file-transfer monitoring and fault-tolerance capability in case of transient failures. DML also provides information to the gateway for file streaming support from BeStMan (the file steaming feature was explained in the BeStMan section), which prevents congestion in the gateway's BeStMan disk cache.

- Catalog access: DML includes ESG data set catalog browsing and search capability, which allows one-stop downloading convenience to users without their having to use multiple components such as a web browser and different downloading tools for specific transfer protocols. Figure 17 shows the DML data transfers with ESG data set catalog browsing and search window.
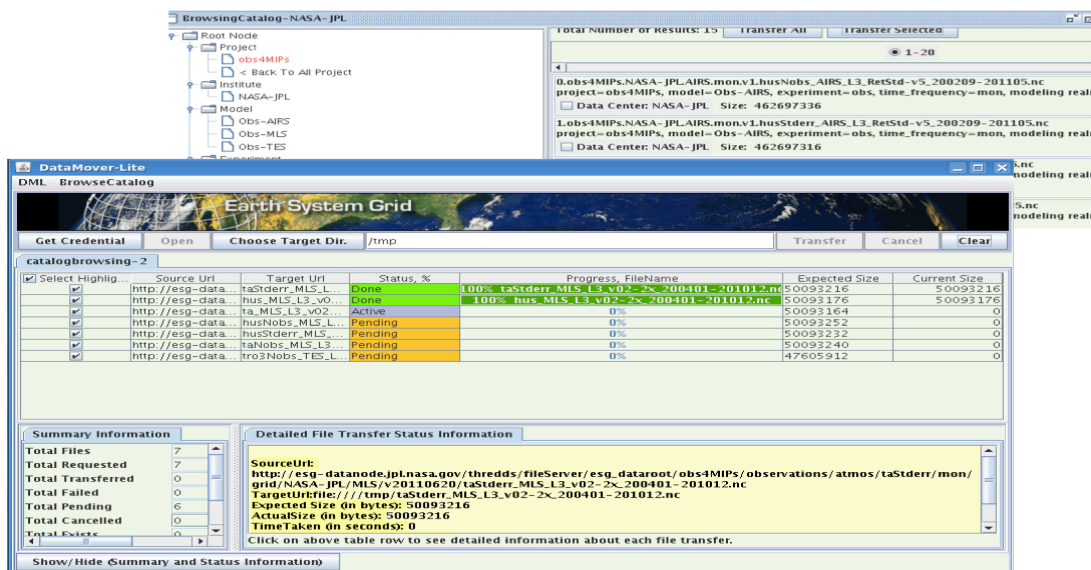


**Figure 22:** *DML's data catalog browsing and search capability and data downloading interface*

### *3.3.5* **HTTP Parallel Streaming in DML**

DML includes HTTP parallel streaming (http-ps) capability, similar to GridFTP parallel streaming, which makes http data downloads faster than usual http downloads. With the http-ps feature, DML downloads a single file by splitting it into multiple blocks and streaming through parallel HTTP connections to an ESGF data server, and each http connection streams partial blocks of files to compose the full file at the target. This capability permits DML to perform HTTP parallel streaming (http-ps) of multiple files from multiple ESGF data nodes when data replicas are available on the nodes. Figure 18 shows an example of how DML makes parallel HTTP connections to transfer blocks from a source file for streaming using the http-ps capability. The large source file is perceived as multiple small manageable blocks, and several available HTTP connections are opened with ESGF Data Transfer server to access the source file. Given a number of available connections, when the data transfer for one block is completed earlier than other blocks, DML initiates the next block transfer. For example, in Figure 18, block 6 (B6), was initiated as soon as block 1 (B1) completed. Similarly, the next transfer shifts to the block B7

after the block B4 completes. Note that the download progress for B7 in Figure 18 is faster than B6 because of possible changes in the network conditions. Thus, our algorithm is adaptive: it will initiate a new stream as soon as any other stream completes. The block size in http-ps can be as small as 1MB. After all blocks for the source file are streamed to the target site, the final downloaded file is created through a fast merging process.
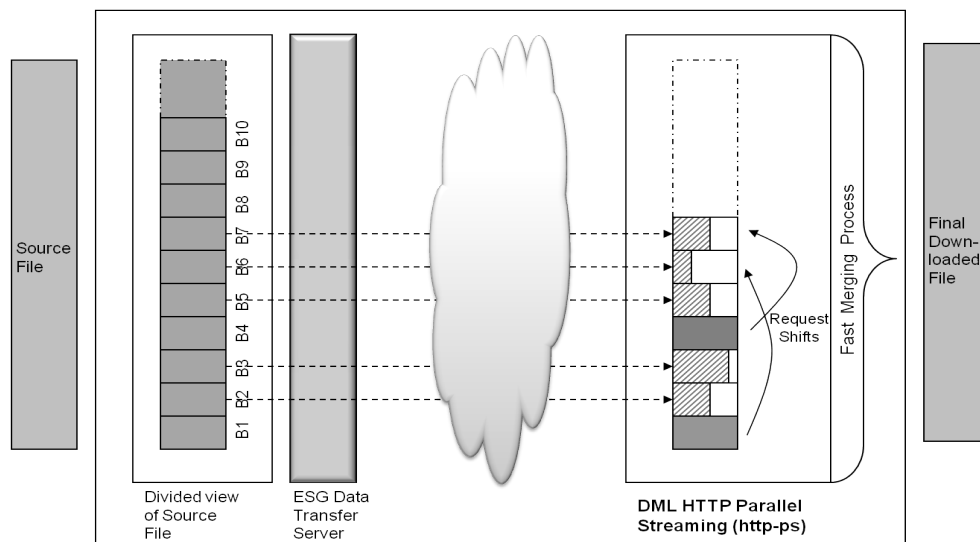


**Figure 23:** HTTP Parallel Streaming (http-ps) feature in DML

### *3.3.6 Bulk Data Mover (BDM) and Transfer Queue Management*

The Bulk Data Mover (BDM) [BDM, BDM10] was developed under ESG-CET project and can be used as a stand-alone file replication tool as well as integrated with ESG-CET Replication Service that also manages publication into metadata catalogues when data is replicated. BDM is a high-level data-transfer management component that manages file transfers using existing transfer services and tools and Adaptive Data Transfer (ADT) algorithms [Sim10]. BDM specializes in handling large variance in file sizes, including the case where a large number of the files are small, which is typical of climate data sets. BDM supports several underlying transfer protocols and services and provides optimized management of transfer requests adaptively based on the data set characteristics.
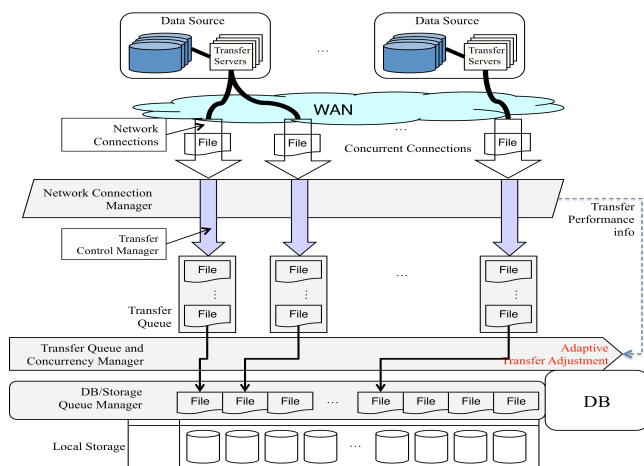


**Figure 24:** *Transfer and concurrency management in BDM, showing adaptive transfer adjustment*

It achieves high performance using various techniques, including multi-threaded concurrent transfer connections, data channel caching, load balancing over multiple transfer servers, and storage I/O pre-fetching. It can accept a request composed of multiple files, an entire directory, or multiple directories. When using the GridFTP transfer protocol, BDM supports transferring multiple files concurrently as well as using parallel TCP streams.

The optimal level of concurrency or parallel streams depends on the bandwidth capacity of the storage systems at both ends of the transfer as well as achievable bandwidth on the wide-area network. Setting up the optimal level of concurrency is important, especially for application with file-size variance. Concurrency that is too high becomes ineffective (high overheads and increased congestion), and concurrency that is too low does not take

advantage of the available bandwidth. A similar phenomenon was observed when selecting the level of parallel streams.

Transfer queue management and concurrent connection management contribute to higher transfer throughput, including both network and storage. When the data set has a large variance in the file sizes, continuous data flow from the storage into the network can be achieved by pre-fetching data from storage onto the transfer queue of each concurrent transfer connection. This overlapping of storage I/O with the network I/O helps improve the transfer performance. As shown in Figure 19, BDM manages a Data Base (DB) queue of the concurrent transfer connections and also manages the transfer queues for the concurrent file transfers. Each transfer queue checks the configurable threshold for the queued total files size and gets more files to transfer from the DB queue when the queued total files size goes below the configured threshold. Storage I/O pre-fetching includes inode creation for writing files at the destination. In many file system, many inode creations at the same time cause a significant overhead in file system performance, and this overhead affects the transfer performance. By creating inodes at the destination paths when files are being on the transfer queue, BDM achieves faster storage I/O during the transfers.

One of the climate data replication use cases is shown in Figure 20 from LLNL/PCMDI to LBNL/NERSC, and BDM maintained the transfer throughput rate throughout the entire transfer period.

BDM is under an open source license, BSD with Grant-back provision (LBNL CR-3005), and available as source and binary packages on https://codeforge.lbl.gov/projects/bdm/.
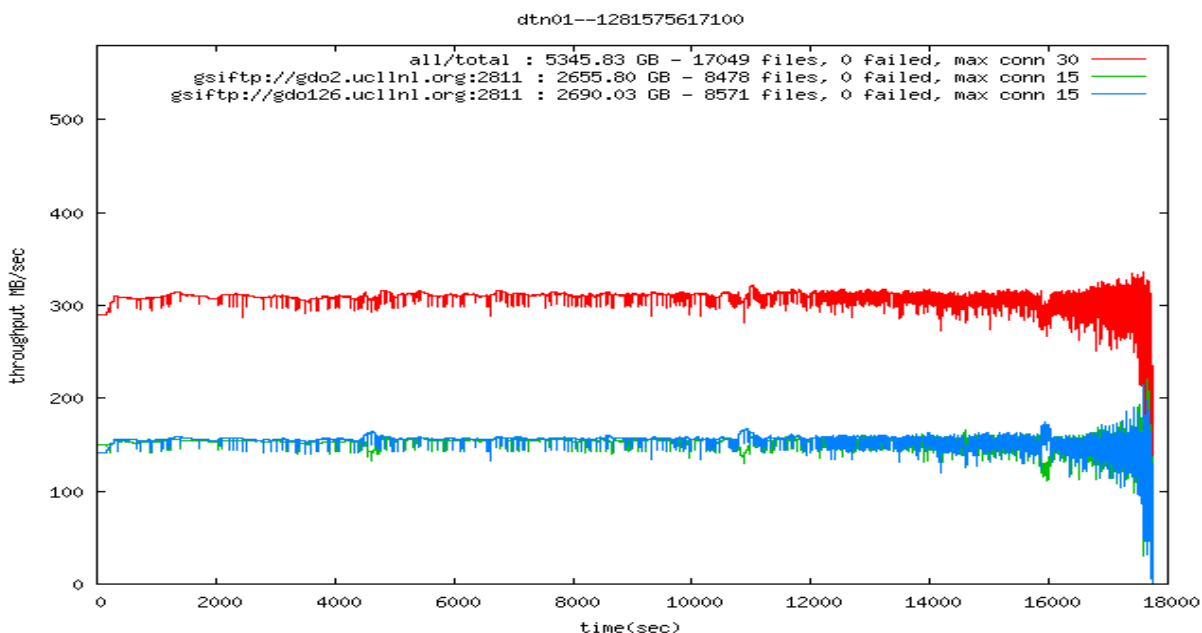


**Figure 25**: *Climate data replication from LLNL to NERSC over a shared network. Transfers of 17049 files totaling 45.3TB of climate data set from two sources at LLNL (shown in green and blue) to one destination at NERSC used 15 concurrent transfers for each data source, and achieved a high throughput (total shown in red) that was maintained over time.*

### 3.3.6.1   Dynamic adaptation algorithm for efficient data movement

The Adaptive Data Transfer (ADT) algorithm was designed to automatically adjust and maximize transfer throughput dynamically, for the movement of large climate data sets over WAN. Instead of making measurements with external profilers to set the level of concurrency, transfer parameters are calculated using information from current data transfer operations. Thus, we avoid the extra load on the network due to extraneous packet transfers and calculations for exact parameter settings. Observing the achieved application throughput for each transfer operation sets the number of multiple streams, and parameters are gradually adjusted according to the current

performance. The transfer time of each operation is measured and the total throughput is calculated. The best throughput for the current concurrency level is recorded. The actual throughput value of the data transfers is calculated, and the number of multiple streams is increased if the throughput value is larger than the best throughput seen so far. In this dynamic approach, we reach a near optimum value gradually, instead of finding the best parameters that achieve the highest throughput at once. This method focuses on application-level tuning rather than low-level network and server optimization. We adjust the number of multiple streams according to the dynamic environmental conditions, and also taking into consideration the possibility of other data transfer operations using the same network resources.

## 3.4 Product Services

### 3.4.1 Motivation

The product services in ESGF provide users with custom visualization, subsetting, and basic analysis capabilities applied to the underlying ESGF data collection via a browser-based user interface. These services are essential because ESGF serves a diverse user community, including scientists in specialties not accustomed to working with complex model outputs and because climate scientists themselves need the ability to quickly discern which data are suitable for their needs. ESGF product services are built upon the Live Access Server (LAS), a server-side workflow engine developed by NOAA/PMEL.

### 3.4.2 Data Preview

The most fundamental capability of the product-services system is to provide visualizations of data: custom maps, time series and vertical profile line plots, and similar graphics along all orthogonal planes and axes from the 4-dimensional space-time coordinate system of the data set. The system selects default characteristics of the plots (contour levels, color palettes, continent maps, scales and annotations), which can be customized by data node managers and can be further refined by end users though UI widgets provided for that purpose.
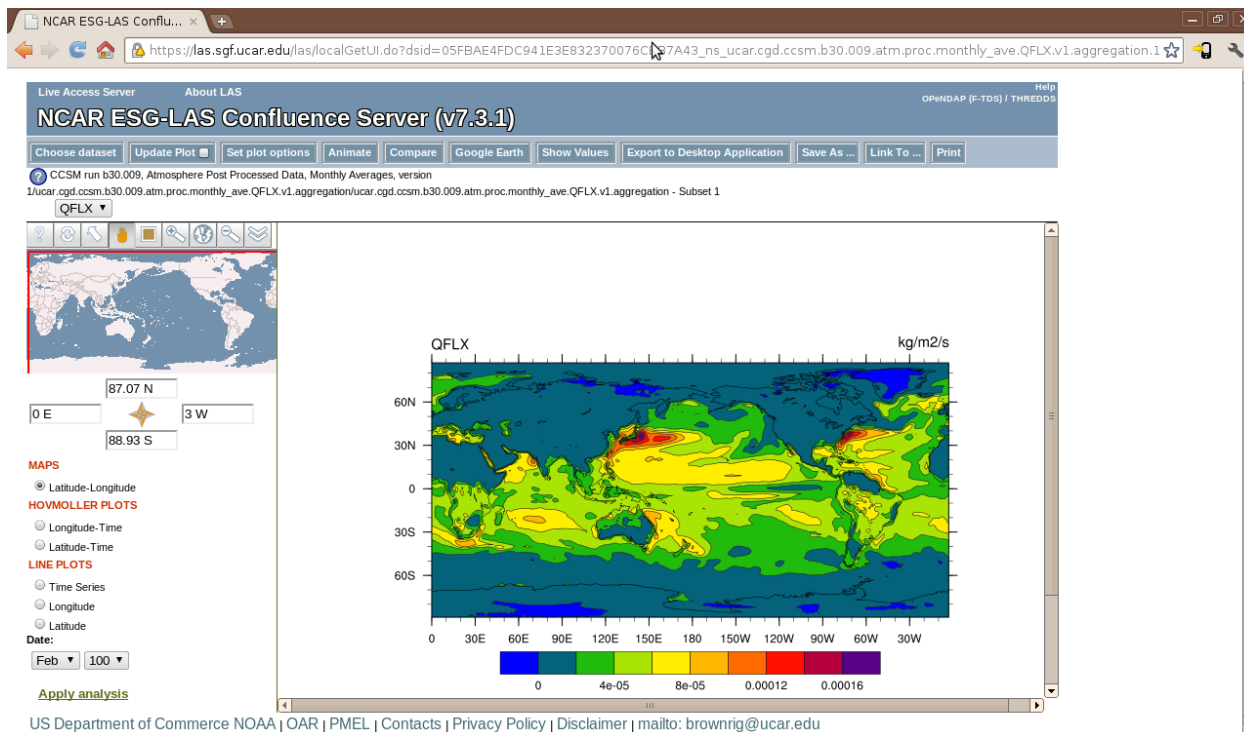


**Figure 26:** *Data preview using NCAR's NCL as the visualization engine for the product services*
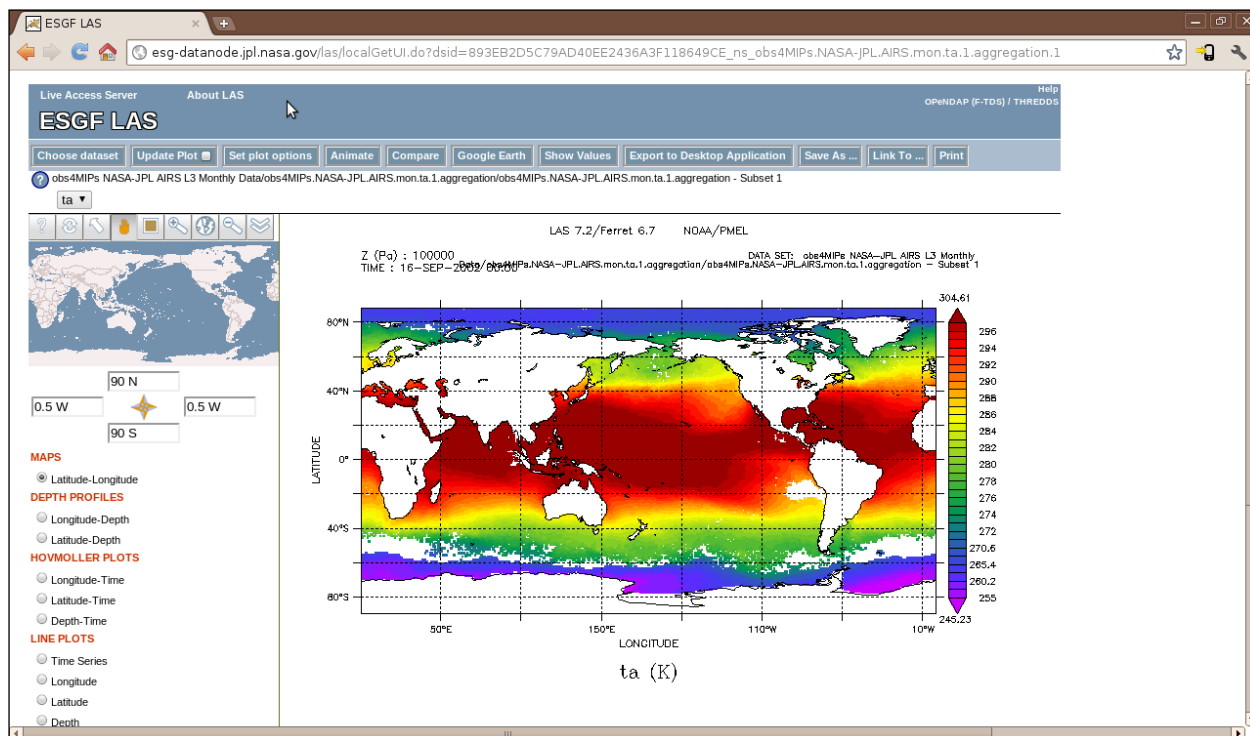
**Figure 27:** *Data preview using PMEL's Ferret as the visualization engine for product services from theNASA/ JPL data node.*

### 3.4.3   Comparisons

One of the greatest accomplishments of ESGF product services is the advent of a highly interactive and customizable comparison capability, the Scientific Visualization Gallery (vizGal). The ability to quickly intercompare model outputs and compare models to observations is essential for modern climate research.

The vizGal interface allows users to compare data qualitatively with side-by-side visualizations and quantitatively by computing and plotting difference fields along any plane of the 4D (space-time) data set. This interface also supports collaboration between scientists by providing URLs that capture the state of the user interface at the moment a given visualization was generated. Sharing such a link with a collaborator allows her to re-enter the ESGF product services at that precise point and pursue further investigation of the data sets.
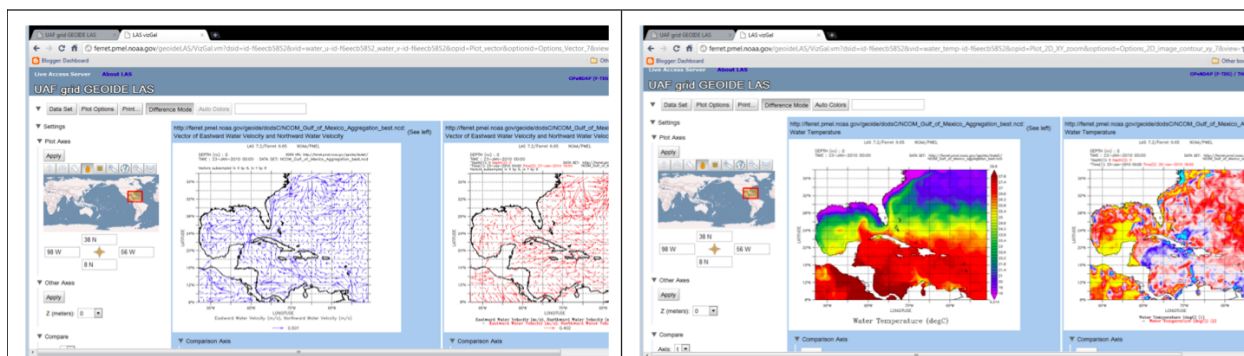


**Figure 28:** *The vizGal user interface client being used to compare vector and scalar data via differencing.*

### 3.4.4   Server-side Analysis and Data Reduction

Users of ESGF product services can perform basic analyses such as averages, sums, variances and extrema over time and area. Transformed data may be visualized with the same tools as untransformed data. The ESGF architecture was designed to whenever possible perform these computations on the server that hosts the data. Keeping large calculations "close to the data" greatly reduces the amount of data, which must be moved across the network.

ESGF developed the Ferret-THREDDS Data Server (F-TDS) to implement these server-side analysis capabilities. Taking advantage of the "pluggable" architecture of the TDS [Schweitzer, 2009], F-TDS harnesses the analysis capabilities of an independent application, Ferret [Hankin, 1996]. Arbitrary mathematical expressions are passed to Ferret through an extended version of the OPeNDAP request URL syntax. The delayed-mode character of analysis with Ferret enables F-TDS to perform on-demand computations only on the minimum subset of data needed to fulfill a particular request. The transformation pipeline is largely transparent to the visualization client software, which simply sees new (virtual) variables embedded in existing data sets. The LAS product services software handles the task of formulating the mathematical expression syntax.

### 3.4.5   Federation

The ESGF is a federation of data held by many different data nodes. Each of these data nodes can have associated with it an LAS Product Server, responsible for generating products from the local data. The data node registry service makes each local LAS aware of the URL end-points of the other LAS data nodes in the federation. The local data node is then responsible for sharing information describing its local data sets on-demand with other nodes. In this way, the LAS user interface from any individual node knows about the data collections held at all other nodes in the federation and can route product requests to the LAS responsible for that data.

The only exception to doing the server-side analysis at the data node where the data are held is in the case of computing a difference between data held at two different data nodes. In this case, the input data needed to compute the difference is moved from the second data node to the first where it is interpolated onto the grid of the first data set before the difference is calculated. All of the data transfer takes place via OPeNDAP client library calls to the remote OPeNDAP server.

### 3.4.6   Data Aggregation

When a user interacts with the ESGF Product Services layer, the interaction is with "data sets," rather than with individual files. This abstraction is created through the aggregation of related data (typically a time-series of files) using OPeNDAP servers such as the THREDDS Data Server and Hyrax. The product services layer, itself, has no awareness of individual files. See section 3.4.7 for more details.

### 3.4.7   Analysis with Domain Specific Tools

While we have worked hard to create a flexible, intuitive and highly interactive web-based interface to deliver Product Services in ESGF, there is no substitute for analysis on local computers using domain specific software tools. The product services system facilitates direct access to the underlying data by delivering the scripts need to access the data via OPeNDAP using many popular analysis tools. The ESGF software infrastructure handles the authorization and access through these analysis tools provided that they are linked with the security-enabled version of the OPeNDAP libraries. This strategy obviates the need to download massive multi-file data collections for an analysis that may require only (say) a small regional sub-set of the full data set.

## 3.5   Security

ESG-CET invested a substantial amount of time and resource into gathering use cases and requirements, analyzing, and selecting a security strategy along with core technology components. ESGF uses OpenID framework, a decentralized identity management system, for security. With OpenID's Single-Sign-On (SSO)

infrastructure, ESGF allows users at all sites to log into ESGF sites, with authorization layers providing the needed services for allowing access only to the permitted set of resources for each user. SSO systems are gaining popularity on the web for applications requiring resources located at different sites because a user can log but once and obtain access to resources housed at any of the remote federated sites that are trusted. OpenID was chosen for ESG-CET because it builds on existing Internet technologies such as HTTP, SSL, and Diffie-Hellman, and is lightweight in nature (thus not requiring much additional software to be installed at each gateway or ESGF Index Node sites). Like other web-based login systems, each user is presented with a login screen before accessing protected resources. The user then enters his or her unique, individually assigned OpenID, and is redirected to an ESGF-operated Identity Provider. This Identity Provider determines if the user is valid, based upon who the user claims to be and what the user is attempting to access; ultimately, the Identity Provider makes the decision whether to grant access.

For supporting single sign-on with non-browser clients, MyProxy Online CA is used to issue user credentials and to provision client machines. MyProxy servers have been deployed on the ESGF gateways and are primarily used by clients that transfer data and publish meta-data. Clients use simple clients to download credentials to use with ESGF services that require PKI authentication. Where necessary, user attributes are embedded in the issued credentials, which can be used to facilitate authorization based on attributes for non-web environments.

OpenID is a critical system that now ties ESGF together through identity management. While Single Sign-On (SSO) was one of the initial goals of our OpenID evaluation, it is used for many reasons including ease-of-use and installation, a small software footprint, and a proven security model. In particular, ESGF leverages the open source OpenID4Java project, a library written in Java that provides an API for using both client and server based functionalities as well as a method for the secure exchange of attributes. OpenID was originally designed to be an open and flexible web-based system that encourages sites to allow users authenticated from arbitrary Identity Providers access to their resources. Potentially, this allows an Identity Provider to be developed that will grant all users permissions to access any resource. For use in ESGF, this particular aspect is not acceptable, and so the OpenID4Java project had to be extended to support a few features that were not provided by default.

The first extension to the OpenID4Java project is what we've termed Identity Provider (IdP) Whitelisting. IdP Whitelisting is a software feature that will only allow authentications from a particular and known list of Identity Providers. The reason for this is that within ESGF, it is reasonable for a site to know who authorized the user to access a particular resource, in the interest of security. While OpenID will generally allow authorization from any IdP, the ESGF sites have a small acceptable list of providers that must be used. To meet this goal, we have added an API that allows a list of Identity Providers to be allowed during the discovery phase, while potentially eliminating others. In addition, a configuration based reference example has been provided.

The second extension is an API for interfacing with attribute providers. The ESGF system has determined that if a user is authenticated by an Identity Provider, attributes will need to be retrieved for the user that can be used at a later time (for example, in performing an attribute based authorization at the resource itself). While OpenID has a method of exchanging attributes, it does not provide a standard API for retrieving attributes. To meet this goal, we have added a flexible API to OpenID4Java that allows a module to be written that can use any method to retrieve attributes. A database module has been developed and provided as a reference example. This module demonstrates retrieving attributes from a relational database.

Both of these extensions have been added to the version of OpenID4Java used by ESGF and are already integrated into the ESG gateways, which now support the OpenID protocol. However, not all security aspects are web-based, as ESGF has clients that perform operations such as data transfer and metadata updates. For these clients, we have put a security system in place that leverages MyProxy.

MyProxy is a credential management system based on X509 Public Key Infrastructure (PKI). Once configured, it allows credentials to be issued to specific users, potentially for a fixed amount of time. This secure method of issuing credentials ensures a user has proper access to a given resource and service. ESGF uses MyProxy installations on the gateway or ESGF Index Node systems and the credentials are issued to non-web based clients

(e.g. GridFTP clients). Extensions to the MyProxy server have been added that allow custom attribute information to be embedded within the issued credentials on a per-user basis (see Figure 24).



**Figure 29:** *ESG's security services architecture*

## 3.6 Replication

ESG-CET researchers in the U.S. and international climate researchers formed ESGF, which deploys ESG-CET software at sites around the world to create an international federation of climate data nodes and gateways. Some member institutions of the ESGF act as mirror sites for climate data sets originally published by other institutions. A mirror site replicates existing data sets on its local data node and publishes metadata about the replicated data sets on its gateway so that those replicas may be discovered and accessed by researchers.

Widespread replication of climate data sets was long considered impractical because of their large size. However, with the growing importance of IPCC AR5 CMIP5 data sets, several sites around the world decided to host a replica or mirror of a key subset of climate data sets. The goals of these mirror sites are to provide reliable access to these data sets for local scientists; to reduce wide area data access latencies; and to improve fault tolerance of the distributed system by making data sets available at multiple sites. The first two ESGF mirror sites are the British Atmospheric Data Centre (BADC) and the German Climate Computing Center (DKRZ). Additional mirror sites are being deployed at the Australian National University and the University of Tokyo.

### 3.6.1  The Replication Client Software

The replication client is implemented as a Python egg archive. It requires the existing ESGF Publisher Python library developed at LLNL/PCMDI. Installed using the *easy_install* utility, the replication client becomes available in the command shell of the data node.

Replication in ESGF proceeds in three phases. Phase one queries metadata about a data set from an ESG gateway and prepares a data movement request as either a control file or a script for a data movement agent to process. The second phase is the actual data transfer by a data transfer agent. The final phase prepares the replica's metadata catalog and publishes its availability to the gateway for the mirroring organization.
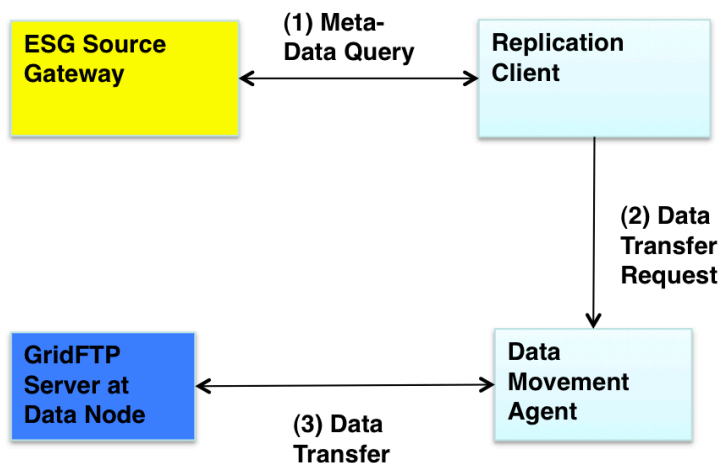


**Figure 30:** *Metadata query and data transfer in ESGF.*

#### 3.6.1.1  Metadata Query

The replication client does not directly transfer data. Instead, it queries a gateway metadata server and uses that metadata to prepare control files or scripts for use by one of the supported data movement agents. These steps are illustrated in Figure 25. To query the metadata, the replication client sends a data set identifier to the metadata server on the gateway. In response, the server returns an XML file describing the data set, including the names of the files, their sizes, what services are available for the file transfer, and the URL of the THREDDS metadata catalog on the data node hosting the data set. The replication client downloads this THREDDS metadata catalog from the data node and saves it for use in the third phase of replication. The client-side interaction with the gateway web service is managed using the Python implementation of the Hessian web services library found in the data node Python library.

#### 3.6.1.2  Creating a Data Movement Request

ESGF data nodes may allow data access using multiple transfer protocols. The replication client currently supports two data movement agents for performing data transfers. The client design is modular, and support for additional transfer agents can be added as necessary.

The default data movement agent is the Bulk Data Mover (BDM) developed at LBNL. BDM is a file transfer management system designed to maximize transfer throughput over networks of varying capacity. The replication client creates a transfer request for BDM by extracting relevant elements (file names, sizes, etc.) from the data set's metadata and writing appropriate commands to copy the files from their source location to the mirror site into an XML file that conforms to the BDM XML Schema. This file is used as the transfer control file for BDM.

An alternative data movement agent is the globus-url-copy utility from the Globus Toolkit. Globus-url-copy was developed at ANL. The utility is designed for effective transfers of data from a GridFTP server or between

GridFTP servers. When using globus-url-copy as its data movement agent, the replication client creates a plain text file of source/target pairs (one per line) that can be read by the command line globus-url-copy utility to perform the data transfers.

The replication client also supports the Globus Online (GO) service, a hosted data transfer service currently being developed at ANL. The replication client can produce the GO command format, which is similar to that of globus-url-copy. The GO service is hosted in the Amazon Cloud. Requests submitted to GO are managed reliably by the GO service, which is responsible for coordinating transfers between GridFTP servers at ESGF data nodes.

Once the command file for the data movement agent is constructed, an administrator at the mirror site initiates the replication of the data set by submitting the command file to that agent. The replication client does not automatically initiate the data transfers itself because mirror sites in the ESGF require the ability to control the timing of data- and network-intensive replication operations for large climate data sets.

### 3.6.1.3   Replica Publication

After the data movement agent transfers the files of the data set to the mirror site's data node, the replication client is used to prepare, check, and publish the data set as a replica on the mirror site's gateway node. These steps are illustrated in **Error! Reference source not found.**.

To prepare the data set for publication as a replica on the mirror site, the replication client calls on the ESG publication library to first scan all the files in the replicated data set and then use the results of the scan to create a new THREDDS metadata catalog, which is stored in an XML file. This catalog file will be used by the mirror site's data node to publish the data set to the scientific community.
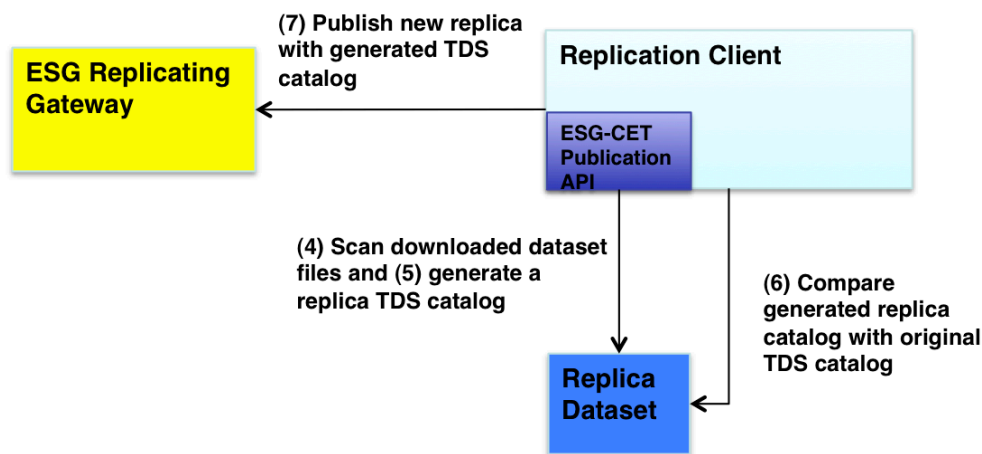


**Figure 31:** *The preparation, checking, and replica publication process in ESGF.*

The replication client compares the newly created THREDDS metadata catalog for the replicated data set to the THREDDS metadata catalog downloaded from the publication data node after the initial metadata query. ESGF data sets may have large numbers of files, which can create very large metadata catalogs. To limit the memory footprint required to compare two large XML files, the replication client uses a white list of properties to use for the comparison. The replication client parses each XML catalog file separately to create smaller Python objects. The class overrides the equals operator to provide flexibility and ease-of-use.

Properties of individual data set files are used for the comparison. The user is able to select which properties, from a predefined list, are to be checked and, of those properties, which ones are required to match and which ones are optional. "Optional" in terms of the replication client means that if one of the catalogs contains the property and the other does not, the comparison does not fail. If both catalogs have an optional property, those properties must match. If a property is mandatory, the comparison will fail if one catalog does not have that property.

The comparison of the THREDDS metadata catalog files is only intended as a check on the scan and replica catalog generation and is not a quality control check on the data files themselves. Users rely on the functionality of the transfer agents, such as the checksum matching in the Bulk Data Mover (BDM), to verify the correctness of data transfers. The original data publishers perform quality control for scientific purposes.

Once the replication client successfully compares the new THREDDS metadata catalog to the catalog from the site where the data set was originally published, the user may publish the replicated data set at the mirror site. Publishing makes the replicated data set's metadata visible on the mirror site's gateway node. After publication, this metadata will be shared with other gateway nodes in the distributed system. Climate scientists and other users around the world are then able to use any gateway to discover and access the newly replicated data set. The replication client uses the ESGF data publication library to perform the publication operation in similar fashion to the existing ESGF publication command line tools.

### 3.6.2   *Performance of Replication Client*

We measured the performance of the replication client on two data sets of size 5 GBytes and 50 GBytes. We downloaded these test data sets from the PCMDI ESG gateway and data node at LLNL/PCMDI. The performance of several sub-components of the replication client is shown in Table 6. The means are calculated for five or more runs of the replication client. These measurements include the metadata query done in phase 1 of the replication client's operation and the file scan, THREDDS metadata catalog generation, and THREDDS catalog comparison done in phase. The standard deviation of some of these measurements is high, resulting in, for example, the scan taking longer on average for the 5 GByte data set than for the 50 GByte data set.

The data transfer times are not included in the table. We downloaded each data set once using the globus-url-copy option for the replication client. The transfers took 827 seconds and 5854 seconds for the 5 GByte and 50GByte data sets, respectively.

**Table 6:** *Performance of replication client components. All times shown are in seconds.*

|  | Metadata 5GB | Metadata 50GB | Scan 5GB | Scan 50GB | Generate Catalog 5G | Generate Catalog 50G | Compare Catalogs 5GB | Compare Catalogs 50GB |
|---|---|---|---|---|---|---|---|---|
| Mean time (seconds) | 0.98 | 1.95 | 1.01 | 0.58 | 2.08 | 3.96 | 0.52 | 0.55 |
| Standard Deviation | 0.08 | 1.32 | 1.06 | 0.01 | 2.42 | 6.89 | 0.00 | 0.01 |

### 3.6.3   *Replication Client Testing*

The ESGF team supported the expanded use and testing of the ESG replication client by three sites: PCMDI at Lawrence Livermore National Laboratory, the British Atmospheric Data Centre (BADC), and the German Climate Computing Center (DKRZ). Personnel at these sites exercised and tested the features of the replication client and identified bugs and issues regarding the use of the client and how it functioned. The replication client was improved based on these bug reports and feature requests. Additional work includes packaging the software to make it easily deployable by participating sites and documenting the software to improve its ease of use by climate scientists.

## 3.7   DAP Services

THREDDS Data Servers (TDS) is at the heart of supplying OPeNDAP services across the ESG Federation. The data node publication process results in configuration catalogs that are consumed by the TDS server. The resulting services include access to individual files for bulk download and access to aggregated views of data variables such that time series that are split across multiple files can be accessed by client applications using a single

OPeNDAP URL. TDS is a pure-Java implementation that installs seamlessly with the other Java-based services (like the LAS product server), node management and authentication and authorization services. OPeNDAP Hyrax, from the developers of the Data Access Protocol (DAP) is alternative OPeNDAP server.

The Tetherless World Constellation (TWC) at Rensselaer Polytechnic Institute (RPI) was involved with efforts to provide OPeNDAP Hyrax services within the ESG Data Node. OPeNDAP Hyrax is a secure, high-performance access and transport mechanism that is highly configurable and extensible. The Hyrax solution incorporates a middle-tier lightweight front-end service called the OLFS (OPeNDAP Light-weight Front-end Service), which communicates with the BES (Back-End Server).

The OLFS is a tomcat servlet that handles DAP URLs. These DAP URLs are then translated into requests to the BES. The OLFS also provides support for reading THREDDS catalogs, including ones with embedded NCML Documents, and tying in information provided by the BES.

The BES is a C++ software framework that allows for the addition of dynamically loaded modules, making the BES highly configurable and extensible. Developers are able to add additional functionality to a BES installation, including but not limited to the reading of additional data sets (NetCDF, HDF4, HDF5, FreeForm, CDF, CEDAR, Fits, and more); the creation of new data products (movies, images, plots, graphs, specific ascii-formatted data products, and more); custom server-side functions available for data manipulation, constraining and projection; custom error handling; initialization and termination call-backs; reporting and metrics mechanisms; and more. Through our ESG work we have incorporated access to embedded NCML documents from THREDDS catalogs and aggregation features provided by our NCML Module. We also have the capability of utilizing the unique power of Ferret in our Ferret Module, allowing ferret to manipulate the data, run intensive calculations over the data, and produce plots, graphs and images of that data.

The work that TWC does might be specific to the ESG-CET project, but the results of these efforts will be incorporated back into the core of OPeNDAP Hyrax, making this open source software available to other data providers/users. And we are currently working to provide OPeNDAP Hyrax version 1.8 for ESGF data nodes with additional administrative features, secure connections, encrypted requests/responses, support for reading new NetCDF 4 data formats, support for HDFEOS, JGOFS format support, creation of provenance information from OPeNDAP objects, data product and services discovery responses, and more. This additional work is being lead by the OPeNDAP organization and supported by the Tetherless World Constellation.

# 4   Data and Overall Community Impact

There is a growing national and international interest in the benefits provided by ESGF to the communities for exchange of worldwide climate data in model simulation, observation, and reanalysis for a growing number of climate assessment reports. The United States (U.S.), United Kingdom (U.K.), Germany, Australia, Japan, and a number of other countries have implemented mandatory or voluntary community ESGF software-supporting standards defined by the community. Community benefits include a wide range of services and activities that ESGF provide to improve science in communities and increase access to climate data. Through the ESGF alliance, governed under the worldwide multi-agency Global Organization for Earth System Science Portals (GO-ESSP), the team has developed an operational system for serving climate data from multiple locations and sources. Model simulations, satellite observations, and reanalysis products will all be served from a distributed data archive. Researchers worldwide can now access ESGF data holdings through any of the ESG gateways or ESGF Index Nodes hosted by ESGF partners, including laboratories in the U.S. funded by the Department of Energy (DOE), the National Science Foundation (NSF), the National Aeronautics and Space Administration (NASA), and the National Oceanic and Atmospheric Administration (NOAA), and at laboratories elsewhere, for example at the Australian National University (ANU) National Computational Infrastructure (NCI), the British Atmospheric Data Center (BADC), the Max Planck Institute for Meteorology (MPI-M) German Climate Computing Centre (DKRZ).

In planning for CMIP5, the ESGF has built on the success of the earlier U.S. DOE funded Earth System Grid (ESG) project, which served CMIP3 model output led by LLNL/PCMDI. CMIP5 has driven all ESGF development work and has attracted the interest of others seeking to make their data widely available and easy to use (e.g., model simulation: CCSM, PCM, etc.; observation: ARM, Ameriflux, TRMM, AIRS, MLS, TES, CloudSat, etc.; reanalysis: MERRA, CERES, etc.). Currently, ESGF is impacting:

- Current CMIP5 activities, and prepare for future assessments (for example, the U.S. National Climate Assessment (NCA));
- Current CSSEF activities are pushing the design of ESGF to fit into a U.S. DOE lead testbed infrastructure for evaluating the uncertainty of climate models;
- Development of data and metadata facilities for inclusion of observations and reanalysis products in CMIP5 (called obs4MIPs);
- Enhancements and improvements for the current climate research infrastructure capabilities through involvement of the software development community and through adherence to sound software principles;
- Collaboration across national and international agencies and political boundaries;
- Integrating and interoperating with other software designed to meet the objectives of ESGF (e.g., software developed by NASA, NOAA, ESIP, USGCRP, the European ES-INES, Australian, and Japan);
- Software infrastructure and analysis tools that facilitate scientific advancements (e.g., DOE funded UV-CDAT, NSF funded NCL, and NOAA funded Ferret).

The software deployed in ESGF has been developed using an open-source approach, and all participants are encouraged to contribute to the ongoing development of the infrastructure.

The ESGF architecture is also making an impact in other scientific domains such as the DOE Office of Nuclear Energy's Large-scale Data Systems for Nuclear Energy project and NSF's Arctic data project.

The Table 7 below represents some of the national and international ESGF data archives that are impacting the community.

**Table 7:** *ESGF Data Archive*

**Priority of Data: H – High, M – Medium, L – Low**

| Institution | Data Set | P | Type | Description | Use | Status | Size |
|---|---|---|---|---|---|---|---|
| ANL | CAM-SE High-Res Gridded | L | Model | Gridded | Atmosphere | Published | 85 GB |
| BADC | TAMIP | M | Model | Gridded | Atmosphere | Collection | 17 TB |
| BNL | ARM CSAPR Rainfall | L | Observational | Gridded | Atmosphere | Collecting | 100 GB |
| BNL | NOAA NEXRAD MOSAIC Rainfall | L | Observational | Gridded | Atmosphere | Collecting | 50 GB |
| LANL | POP | H | Model | Gridded | Ocean | | |
| LLNL | CMIP5, CMIP3 | H | Model | Gridded | Atmosphere Land Ocean | Collecting | 5-10 PB |
| LLNL | NASA MERRA, CERES | M | Reanalysis | Satellite | Atmosphere | Published | 68 GB, 162 MB |
| LLNL/PNNL | NASA TRMM, | H | Observational | Satellite | Atmosphere | Collecting (TRMM) | 15 GB (TRMM_3842) |
| LLNL/PNNL | NASA AIRS, | H | Observational | Satellite | Atmosphere | Collecting | 30 GB |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | MLS, TES | | | | | | |
| LLNL/PNNL | CAM5 Sensitivity Runs | H | Model | Gridded | Atmosphere | Awaiting availability | N/A |
| MPI-M | CMIP5 | H | Model | Gridded | Atmosphere Land Ocean | Collecting | 50 TB |
| MPI-M | EUCLIPSE | H | Model | Gridded | Atmosphere | Will collect | TBD |
| MPI-M | LUCID | M | Model | Gridded | Atmosphere Land Ocean | Will collect | 6 TB |
| MPI-M | CORDEX | M | Model | Gridded | Atmosphere Land Ocean (regional) | Will collect | TBD |
| NCAR | NARCCAP | H | Model | Regional Gridded | Atmosphere | Collecting | |
| NCAR | PCM | M | Model | Gridded | Atmosphere Land Ocean | Published | |
| NCAR | CCSM | H | Model | Gridded | Atmosphere Land Ocean | Published | |
| ORNL | CLM Single Point | H | Model | Single Point | Land | Published | 54 GB |
| ORNL | CLM Gridded | H | Model | Gridded | Land | Published | 119 GB |
| ORNL/LLNL/ PNNL | ARMBE (a.k.a. CMBE) | M | Observational | Single Point Gridded | Atmosphere | Published 6 variables at ORNL | 82 MB |
| ORNL | C-LAMP | H | Model | Single Point Gridded | Land | Published | 153 GB |
| ORNL/PNNL | AmeriFlux | H | Observational | Gap-filled surface weather forcing data (U.S. sites) | Land | Published | 339 MB |
| ORNL | Fluxnet Canada | M | Observational | Gap-filled surface weather forcing data (Canadian sites) | Land | Awaiting Authorization | N/A |
| PNNL | USGS Basin Boundaries | L | Geospatial Reference | Vector | Land | Collected | N/A |
| PNNL | STATSGO, LAI, SAI (MODIS) | L | Reanalysis | Single Point Gridded | Land | Collected | <5 MB |
| PNNL | NLDAS2 | L | Observational Reanalysis | Single Point | Land | Collected | 3 GB |
| PNNL | DEM (MOPEX) | M | Geospatial Reference | Gridded | Land | Collected | 10 GB |
| PNNL | CLM4 Single Point, sample of parameters | H | Model | Single Point | Land | Collected | 300 GB |
| PNNL/ORNL | USGS Stream flow for MOPEX basins | M | Observational | Single Point Time Series | Land | Collected | 4 MB |

| PNNL | MTSAT | M | Observational | Satellite | Atmosphere | Collecting | 250 GB (twpmtsatX1.a1 version) |
|---|---|---|---|---|---|---|---|
| PNNL/LLNL | NASA CALIPSO, CloudSat | L | Observational | Satellite | Atmosphere | Collecting | 5 TB |
| SNL | Hydrobase3 | H | Observational climatologies | Gridded | Ocean | Awaiting availability | TBD |
| SNL | World Ocean Atlas 2009 | H | Observational climatologies | Gridded | Ocean | Will collect | TBD |
| SNL | LES runs | M | Model | Single Point Gridded | Ocean | Will collect | TBD |
| SNL | Florida Current Project | H | Observational | Single Point | Ocean | Will collect | TBD |
| SNL | RADARSAT-1 | H | Observational | Gridded | Sea ice | Will collect | TBD |
| SNL | ESA Cryosat Siral | M | Observational | Single Point Gridded | Sea ice Ice sheet | Determining availability | TBD |
| SNL | NSIDC, Hadley Ctr Passive microwave | H | Observational | Gridded | Sea ice Ice sheet | Will collect | TBD |
| SNL | ICESat | H | Observational | Gridded | Sea ice | Will collect | TBD |
| SNL | IceBridge | H | Observational | Gridded | Sea ice | Will collect | TBD |

Sections 4.1 and 4.2 list a few important portals and web links showing ESGF's impact on the climate communities.

## 4.1    Portals

- *ESG gateway portals: See Table 1 for list of portal URLs.*
- *ESGF P2P node portals: See Table 3 for list of portal URLs.*
- *ESGF data node THREDDS URLS: See Table 2 for list of THREDDS URLs.*

## 4.2    Project Web Links:

- *ESG-CET project website:*
  - *http://esg-pcmdi.llnl.gov/*
  - *https://wiki.ucar.edu/display/esgcet/Home*
- *ESGF international project website:*
  - *http://esgf.llnl.gov or http://esgf.org*
- Up to the minute CMIP5 archive status and summary:
  - http://esgf.llnl.gov/wiki/Cmip5Status/ArchiveView
  - http://cmip-pcmdi.llnl.gov/cmip5/esg_tables/esg_static_table.html
  - http://cmip-pcmdi.llnl.gov/cmip5/esg_tables/transpose_esg_static_table.html
- Data movement between core CMIP5 centers:

      o   http://esgf.llnl.gov/wiki/Cmip5Status/DataMovement

# 5   Collaborations

Partnerships and our intent to collaborate is reflected by a close relationships with a wide variety of data, science, and technology efforts. These relationships positioned ESG-CET to make a major impact on the progress of science in: (a) CMIP5, (b) CSSEF, (c) CCSM, and (d) other data intensive climate relative community projects as mentioned in section 4 *Data and Overall Impact*.

To effectively build the distributed infrastructure to accommodate the needed petascale data management and analysis ecosystem, the ESG-CET team established connections with researchers and scientists involved in other DOE Office of Science SciDAC-funded projects and international programs aimed at assisting in DOE's climate mission. These involved discussions—through workshops, conferences, and face-to-face meetings—with large numbers other SciDAC Centers and Institutes and the climate community with strong interests in collaborating. In most cases, these collaborative projects developed tools and technologies that were exclusively beneficial to the ESG-CET effort, and in some cases there were interests in generalizing and enhancing ESG-developed technology and disseminating to a larger audience in other scientific domains. For example, we will relied on the proposed *Center for Enabling Distributed Petascale Science* to enhance, support, and/or develop tools for distributed data movement and management (e.g., GridFTP and Globus Online). For another example, the *Visualization and Analytics Center for Enabling Technologies* relied on ESG-CET to integrate, package, and deliver their visualization and analysis products to the climate community.

For many on the collaboration list, our project was vital for their national and international programs and projects to reach their goals. For this reason, we positioned ESG-CET team members that overlap in services and institutions to represent our project as liaisons between many of the disparate organizations. See Appendix A *Enabling Technologies and Collaborations* for a more detailed list of our collaborations.

Over the time period, collaborations were developed with the following groups: TeraGrid Science Gateways, Earth System Curator, NOAA's GIP, METAFOR, World Meteorological Organization (WMO) Information System, Scientific Computing and Imaging Institute at the University of Utah, SciDAC Visualization and Analytic Center for Enabling Technologies (VACET), SciDAC Scientific Data Management Center (SDM), SciDAC Center for Enabling Distributed Petascale Science (CEDPS), Southern California Earthquake Center, Tech-X Corp., NASA Langley, NASA Goddard, GO-ESSP, Climate100, and many others.

# 6   The Future of ESGF

With CMIP5 data now being served, the ESGF federation is working to improve various aspects of the system by adding new capabilities that should better meet the needs of users.  Among the improvements expected over the next several months are:

1. A simpler scripting method for downloading files, involving Globus Online and direct access Data Mover-Lite;
2. An enhanced search capability, from the ESGF P2P node and from the ESG gateway;
3. An automatically updated table showing which simulations have been archived by each model, the start of these websites can be viewed under section 4.2 *Project Web Links*;
4. A notification service to advise users when errors are found in data sets via e-mail and/or Really Simple Syndication (RSS) feed;
5. A straight-forward method to report errors discovered in the data and to provide feedback to the modeling groups about their simulations;

6.  A list of publications based on CMIP5 model output, as recorded by users through a web form, the journal publication page will be located on the CMIP website (http://cmip-pcmdi.llnl.gov);

7.  General system enhancements related to scaling to millions of data sets and petabytes of data volume;

8.  An online visualization capability that will allow users quick inspection and comparison of data sets from multiple locations;

9.  An enhanced capability to perform server-side data reduction and calculations, which will reduce the volume of data transferred to the users via the Internet;

10. A dashboard capability showing the worldwide system metrics of users, downloads, and published data;

11. At some sites, there will be a crossover from using the ESG gateway to fully using the ESGF P2P node. This will better integrate U.S. climate data products and projects from other U.S. agencies (i.e., DOE, NASA, NOAA, etc.);

12. Address the ability to move large amounts of data using BDM and GO at high speeds among sites critical to CMIP5 goals (i.e., LLNL/PCMDI-U.S., BADC-U.K., DKRZ-Germany, ANU/NCI-Australia, DAIS-Japan).

The CSSEF project goal is to develop a new climate-model test bed infrastructure comprising of two major capabilities: (1) a calibration platform where uncertainty quantification (UQ) techniques are used to calibrate a model against regional observational data sets and (2) a validation platform where simulation quality is quantified against global observational data sets. This work is driven by requirements from both model components (atmosphere, land, and ocean) and the overall global climate model. The ESGF node will existing technologies, standards, and expertise to build a unique turnkey operation suitable for CSSEF planned studies, designed experiments, test beds, and large-scale data distribution. Over the coming year, if continued funding, we are expected to:

1.  Publish and maintain many of the data sets listed in Table 7;

2.  Provide a powerful platform for UQ;

3.  Develop analytical tools to include multivariate statistical analysis, standardized model evaluation metrics that can operate across multiple grid resolutions, regridding tools for ocean and atmosphere, and other data assimilation and summarization tools;

4.  Address major issues in data standards, data collection, data annotation, data storage, data dissemination, and data analysis;

5.  Address the ability to move large amounts of data using BDM and GO at high speeds among sites critical to CSSEF goals—high speed here means end-to-end (disk-to-disk) data movement at close to peak network speed over 10 Gbps or 100 Gbps networks);

6.  Develop ontology for describing and synthesizing ecological observational data into the ESGF node;

7.  Create a rapid development coupled model test bed with integrated access to customizable observational and derived data products for model verifications and uncertainty quantification.

The ROSES-ACCESS/NASA proposal is looking to include NASA remote sensing observations from Science Information Processing Systems (SIPS) and Science Computing Facilities (SCFs) such as the Atmospheric Infrared Sounder (AIRS), the Microwave Limb Sounder (MLS), and from NASA's Distributed Active Archive Centers (DAACs) including PO.DAAC, the GES DISC, and the Langley DAAC into the ESGF node infrastructure. These data sets will play an increasingly important role in the upcoming IPCC 5th Assessment Report (AR5) and the CSSEF test bed. If funded, the work will include:

1.  Automate the publishing of data from NASA DAACs and SIPS by providing a NASA ESGF P2P node stack that integrates with ESGF nodes;

2.  Provide a unified portal interface for intercomparing NASA observational data sets with climate model outputs;

3.  Coordination regarding publishing NASA observations to the ESGF;

    4.   Provide NASA mechanisms to remotely access ESGF node holdings;

The Nuclear Energy (NE) community is looking to emulate ESGF nodes for their data community under the Large-scale Data Systems for Nuclear Energy proposal. If funded, we are to develop a production environment for managing and accessing large-scale nuclear energy data sets and well as analysis and visualization tools for nuclear energy science and applications. Work will be based on existing capabilities already demonstrated for climate science community, including a large set of national and international data users and partners. The goal of this effort is to:

    1.   Provide the nuclear energy community with access to the data, information, model codes, analysis tools, and intercomparison capabilities required to deeply analyze nuclear energy simulations and data sets.

    2.   Successful development will require partnerships across the NE community.

ESGF project management strategy and the completion time-scales by which each of the above ESGF P2P node use has been evaluated and carefully planned and organized by the ESGF team and the projects themselves. The project management plans for each defines project goals and objectives; milestones and deliverables, specify tasks or how goals are to be achieved, what resources are needed, the roles of the participating institutions, and their leading individuals. This view of the project flow includes careful controls to ensure that each plan is being managed according to design. Moreover, the projects listed above are coordinated carefully together and are needed interoperable components to strengthen their counterpart projects.

We have good agreement that the high-level requirements that we articulated in all ESGF proposals are on target. However, in order to accommodate our final push for new releases, we restricted our scope in some specific areas.

*At this time, I would like to again thank our sponsors for the opportunity to serve the communities; our international partners for their partnering and service, guidance and help; stakeholders for their patience and understanding while coming online; and to the many team members that help make ESG-CET a success. I hope to continue our fruitful collaborations in the future.*

# 7   References

[SRM07] "The Storage Resource Manager Interface Specification Version 2.2", Editors: A. Sim, A. Shoshani. GFD.129, Open Grid Forum, Document in Full Recommendation, Feb. 2008, http://ogf.org/documents/GFD.129.pdf

[Shoshani02] "Storage Resource Managers: Middleware Components for Grid Storage", Arie Shoshani, Alex Sim, Junmin Gu, Nineteenth IEEE Symposium on Mass Storage Systems, 2002

[Shoshani03] "Storage Resource Managers: Essential Components for the Grid, in Grid Resource Management: State of the Art and Future Trends", Arie Shoshani, Alexander Sim, and Junmin Gu, Edited by Jarek Nabrzyski, Jennifer M. Schopf, Jan weglarz, Kluwer Academic Publishers, 2003

[SRM09] "Practical Grid Storage Interoperation", J Jensen, R Downing, D Ross, A Sim, J. Grid Comp, Vol 7, Issue 3 (2009), p. 309. DOI: 10.1007/s10723-009-9127-2

[OGF09] "Storage Resource Manager Interface Specification V2.2 Implementations Experience Report", A. Sim, A. Shoshani, F. Donno, J. Jensen, GFD.154, Open Grid Forum, Aug. 2009. http://ogf.org/documents/GFD.154.pdf

[MSS2007] "Storage Resource Managers: Recent International Experience on Requirements and Multiple Co-Operating Implementations", A. Shoshani, et al, Proceedings of the 24th IEEE Conference on Mass Storage Systems and Technologies, 2007

[BESTMAN] http://sdm.lbl.gov/bestman/

[SSDBM04] "DataMover: Robust Terabytes-Scale Multi-file Replication over Wide-Area Networks", Alex Sim, Junmin Gu, Arie Shoshani, Vijaya Natarajan. Proceedings of the 16th International Conference on Scientific and Statistical Database Management (SSDBM 2004), Greece, 2004

[BDM] http://sdm.lbl.gov/bdm/

[BDM10] "Efficient Bulk Data Replication for the Earth System Grid", A. Sim, D. Gunter, V. Natarajan, A. Shoshani, D. Williams, J. Long, J. Hick, J. Lee, E. Dart, Proceedings of International Symposium on Grid Computing, Data Driven e-Science: Use Cases and Successful Applications of Distributed Computing Infrastructures (ISGC 2010), 2010

[Sim10] "Adaptive Transfer Adjustment in Efficient Bulk Data Transfer Management for Climate Datasets", A. Sim, M. Balman, D. Williams, A. Shoshani, V. Natarajan, Proceedings of the 22nd International Conference on Parallel and Distributed Computing and Systems (PDCS 2010), 2010

[Clim100] http://sdm.lbl.gov/climate100/

[Schweitzer, 2009] **Roland Schweitzer**, Weathertop Consulting, LLC, College Station, TX; and K. M. O'Brien, J. Li, A. Manke, J. Malczyk, and S. Hankin (2009) **A General Purpose System for Server-side Analysis of Earth Science Data,** 25th Conference on International Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology, The 89th American Meteorological Society Annual Meeting (Phoenix, AZ).

[Hankin, 1996] Hankin,S., D.E.Harrison, J.Osborne, J.Davison and K. O'Brien (1996) A Strategy and a Tool, FERRET, for Closely integrated visualization and analysis. J. Visualization and Computer Animation, 7, 149-157.

## Appendix A   Enabling Technologies and Collaborations

- ***April 1, 2011 through September, 30, 2011***

  - *ESG Gateway and Data Node Deployment at NERSC*

    We collaborated with NERSC Science gateway team to replicate the IPCC AR4 CMIP-3 data sets (~35TB) from LLNL to NERSC. The data sets are accessible to the ESGF climate community by deploying an "ESGF data node" and publishing climate data over an "ESG gateway" at NERSC. We have used the recent ESG distribution, which includes the latest software releases, and successfully configured and deployed ESG gateway (esg.nersc.gov) and data node (esg-datanode.nersc.gov) at NERSC. We have developed and implemented the necessary scripts for production-level services, and prepared a deployment guide for ESG Data node components based on our experience. In this guide, we described how to create a data set catalog, list files in the THREDDS server and publish on the gateway, set up security filters, re-initialize catalog information, and create the initial ESG configuration. We also provided details about authentication and communication with the ESG gateway, explained all required components and listed how to configure the ESG software stack. We have documented our entire deployment experience, and shared this information at our project page (https://sdm.lbl.gov/wiki/Projects/EarthSystemGrid/). In order to join the ESG gateway federation, we have configured the required security components and generated certificates for the ESG federation. We also have updated and released gateway certificates to be updated in other gateway systems for the ESG federation.

    ESG-NERSC is one of the gateway nodes listed in the ESG federation. When IPCC AR5 CMIP5 data sets are available, those data sets will be available on the NERSC ESG gateway and data node.

  - *Collaboration with a Network Project (Climate100)*

    Climate100 (Clim100) is an Advanced Networking Initiative project to research, develop, and test end-to-end capabilities of the next generation networks in collaboration with the climate community. One of the main goals is to enhance data transport technology based on application requirements to ensure that the climate community is ready for the next-generation high-bandwidth networks.

    This project ensures that the software, services, and applications associated with massive climate data sets can be scaled to the 100-Gbps infrastructure as shown in Figure 24 providing high-performance data movement for distributed networking. The ESG community plans to deploy similar systems on multiple 10- or 100-Gbps connections to move massive data sets between its major data-node sites in the U.S. With demonstrated success at the U.S. sites, the ESG community will then extend high-performance data transfers to other data nodes in the U.S. and foreign partner sites. The scaled testing of the data transfers and network connections is important because climate researchers worldwide use many different network service providers. Continued success will enable a production-scale system that will empower scientists to try new and exciting data exchanges that could lead to breakthroughs.

    We have actively collaborated to exchange specific needs and data management requirements of climate research over 100-Gbps networks by using real-life data and real use cases for testing and also experimenting with the underlying network infrastructure.
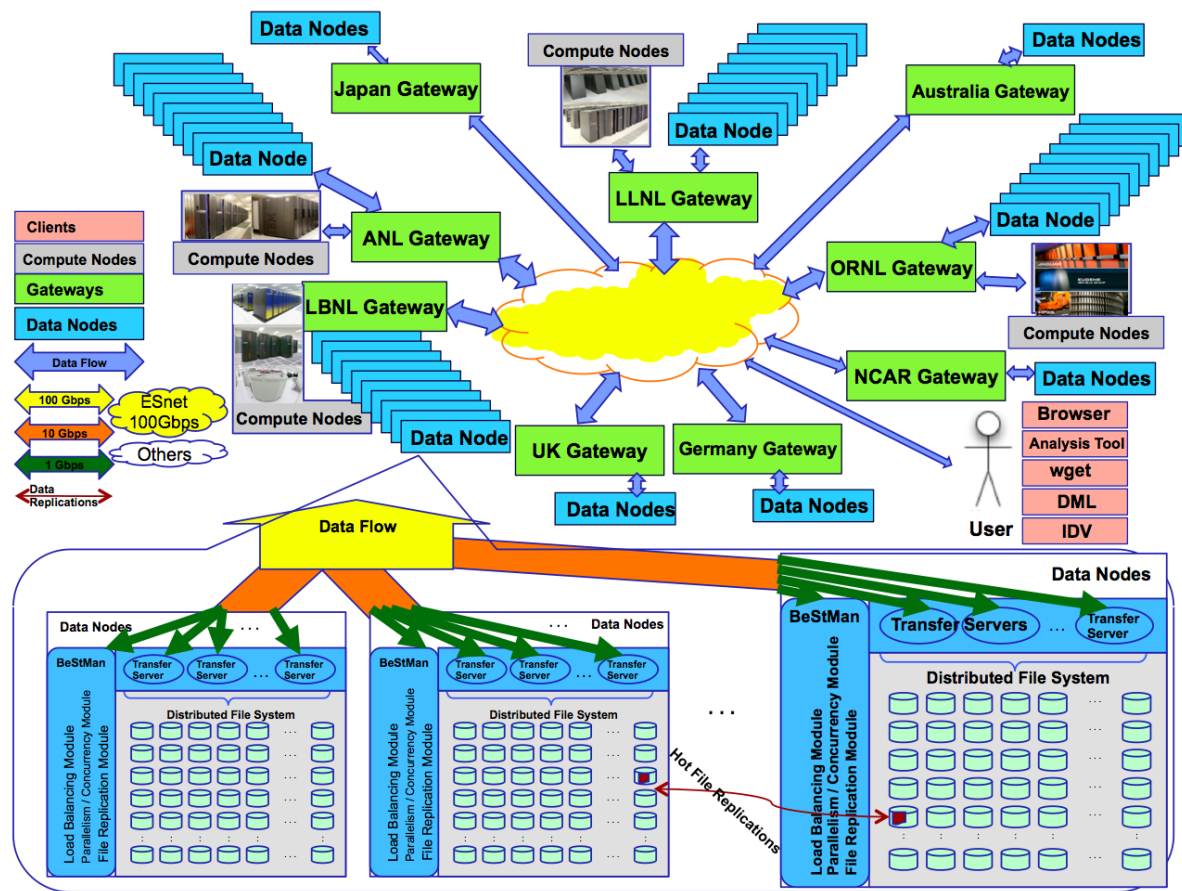
**The envisioned topology of the ESGF community uses 100-Gbps ESnet network connections to provide a network of geographically distributed gateways and data nodes in a globally federated, built-to-share scientific discovery infrastructure. In this way, independent data warehouses can deliver seamless access to vast data archives so that scientists can deploy specialized client applications for extreme analysis. Experts (e.g., model developers, climate researchers) and non-experts alike will benefit from rich data exploration and manipulation available with such fault-tolerant, end-to-end system integration.**

- *October 1, 2010 through March 31, 2011*

  - *LLNL Computation Directorate External Review*

    Dean N. Williams presented ESG-CET at the LLNL Computation Directorate External Review. The presentation entitled, "Enabling Worldwide Access to Petascale Climate Data" reported on ESG and discussed outstanding issues, future needs, and scientific directions for the climate community. The talk highlighted ESG as the cornerstone for distributing climate data to the community.

  - *DOE ASCR 2011 Workshop on Exascale Data Management, Analysis, and Visualization*

    Dean N. Williams and Gary Strand attended the DOE ASCR (Advanced Scientific Computing Research) 2011 Workshop on Exascale Data Management, Analysis, and Visualization held in Houston, Texas. The two-day workshop focused on scientific data analysis and visualization for charting the research roadmap necessary to support scientific discovery on exascale systems. Strand presented information on the needs of the climate community, which included the need for ESG.

  - *Terabit Networks for Extreme Scale Science*

    Dean N. Williams, Ian Foster, Alex Sim, and Galen Shipman represented ESG-CET at the DOE ASCR Terabit Networks for Extreme Scale Science. The workshop, held in Rockville, Maryland, focused on

advanced network services and middleware that integrate users and science applications to the underlying terabit network infrastructure. Williams gave a talk explaining ESG and how its served applications are data and network intensive. He also suggested that if the network can provide some new functionalities and services, it might be able to serve a broader range of disciplines.

▪ *American Meteorological Society Python Short Course*

Dean N. Williams presented ESG-related topics at the American Meteorological Society conference in a two-day short course symposium entitled, "Using Python in Climate and Meteorology: Advanced Methods." The short course, held in January in Seattle, Washington, was designed in modules covering various applications for which Python is used, such as Input/Output, visualization, diagnosis, languages coupling, etc.). The course included projects such as CDAT and ESG, showing how Python is used to manipulate and access petabyte data sets.

▪ *National Climate Assessment Teleconference*

Dean N. Williams presented for ESG-CET at the National Climate Assessment (NCA) Teleconference held December 13, 2010. NCA aims to incorporate advances in the understanding of climate science into larger social, ecological, and policy systems, with the goal of providing integrated analyses of impacts and vulnerability. ESG was discussed as a possible infrastructure for their effort.

▪ *DAARWG Workshop*

Dean N. Williams attended the NOAA Advisory's Board Data Archive and Access Requirements Working Group (DAARWG) in early December 2010—an important working group of NOAA's Science Advisory Board (SAB). DAARWG evaluates data archiving and access requirements from all NOAA observing systems and computational models, as well as from relevant non-NOAA sources. Its charter is to provide scientific advice and broad direction regarding the wide range of data, information, and products that NOAA archives, and ways in which the agency can best provide access to these resources. ESG was mentioned throughout the workshop as part of their end-to-end data environmental data management lifecycle.

▪ *NOAA Global Interoperability Program*

NOAA's GIP promotes coordination of software infrastructure development across agencies, the weather and climate communities, modeling and data services, and research and operational centers.

The "Standardized Analysis Workflows for Climate Models" project continues under GIP. The project addresses the need for Web-accessible tools via ESG that are suitable for the analysis and intercomparison of climate model outputs. Special focus has been on collections of model outputs that comprise CMIP5 multi-model ensembles—models that do not necessarily share the same grid coordinate structure and may include multi-tile "gridspec" coordinate systems.

▪ *National Science Foundation TeraGrid Science Gateways Program*

ESG-CET, NOAA GIP (Curator), and Purdue University staff collaborated on a proposal to develop a prototype Environmental Science Gateway and were granted support from TeraGrid's Science Gateway program for two years. The basic objectives of this project include developing a gateway-based model execution capability, a supporting workflow system that can operate on the TeraGrid gateway, and data publishing and harvesting capabilities that can interoperate with an integrated rule-oriented data system– (iRODS-) based archive as well as disk-hosted model output. During this reporting period, the ESG-CET team made substantial progress developing an Atom-based ingests capability, new portal job submission capabilities, and provisions for Earth System Modeling Framework (ESMF)-based models to produce CIM-compliant metadata for ingestion into the ESG gateway.

▪ *The North American Regional Climate Change Prediction Program*

Since its inception, the NARCCAP program has been using ESG infrastructure to successfully publish, manage, and share regional climate model data with the broader climate science community. NARCCAP staff continues to publish new data sets into ESG, even as NARCCAP heads into its final project year. Middleton attended the annual NARCCAP Principal Investigator (PI) meeting in 2010 at NCAR in Boulder, Colorado.

- ***April 1, 2010 through September 30, 2010***

  - *2009 Global Organization for Earth System Science Portal Workshop*

    Dean N. Williams, Steve Hankin, and Don Middleton are members of the GO-ESSP steering committee responsible for organizing the 2011 workshop, which will be held in Asheville, North Carolina, USA, and hosted by NCDC.

    The GO-ESSP workshop facilitates the organization and implementation of an infrastructure for full data sharing among a consortium spanning continents, countries, and intergovernmental agencies. This consortium envisions an environment that allows users open access to petabytes of model-generated, satellite, and in situ data including physical, biogeochemical, and ecosystem content. All initial ESG Federation partners attended the 2010 workshop (i.e., LLNL, NCAR, GFDL, BADC, DKRZ, and the University of Tokyo).

  - *NOAA Global Interoperability Program*

    NOAA's GIP program promotes coordination of software infrastructure development across agencies, the weather and climate communities, modeling and data services, and research and operational centers.

    The Curator project, now hosted under GIP, continues to be a highly productive collaboration with ESG-CET. Curator serves as an active liaison between the EU-based METAFOR project, which is developing a CIM for CMIP5/IPCC, and ESG-CET, which will be delivering data to the global community. We have jointly demonstrated new "model trackback" capabilities, which effectively combine the METAFOR questionnaire with the query and browse functions of ESG.

  - *WMO Information System*

    The WMO Information System (WIS) Program is developing a next-generation, globally federated data system to serve all of the WMO areas (such as climate, weather, and hydrology). Don Middleton contributes to several WIS committees and regularly provides briefings that encompass ESG-CET and our progress toward CMIP5/IPCC objectives and climate research in general. Our Gateway technology will provide federation with WIS while also serving as a WIS validation platform.

  - *NSF TeraGrid Science Gateways Program*

    ESG-CET, NOAA GIP (Curator), and Purdue University staff collaborated on a proposal to develop a prototype Environmental Science Gateway and was granted support from TeraGrid's Science Gateway program for two years. The basic objectives of this project include developing a Gateway-based model execution capability, a supporting workflow system that can operate on the TeraGrid Gateway, and data publishing and harvesting capabilities that can interoperate with an integrated rule-oriented data system–(iRODS-) based archive as well as disk-hosted model output. Work on this effort continued throughout this reporting period.

- ***October 1, 2009 through March 31, 2010***

- *2009 Global Organization for Earth System Science Portal (GO-ESSP) Workshop*

  Dean N. Williams, Steve Hankin, and Don Middleton are three of seven GO-ESSP steering committee members who coordinated the eighth annual GO-ESSP workshop held October 6 – 8 at the Institute for Pharmacy in Hamburg, Germany. In addition, Steve, Don, and Dean chaired workshop sessions.

The GO-ESSP workshop focuses on facilitating the organization and implementation of an infrastructure for full data sharing among a consortium spanning continents, countries, and intergovernmental agencies. This GO-ESSP consortium envisions an environment that allows users open access to PB of model-generated, satellite, and in-situ data including physical, biogeochemical and ecosystem content. All initial ESG Federation partners were present (i.e., LLNL, NCAR, GFDL, BADC, German Climate Computing Centre (DKRZ) and the University of Tokyo). The workshop, in part, covered data security, versioning, and replication concerns and addressed issues of collaboration. By 2011, this organization envisions allowing users open access to PB of multi-model generated data, as well as in-situ, satellite, biogeochemistry, and ecosystems data.

- *NOAA Global Interoperability Program (GIP)*

  Dean N. Williams, Don Middleton, Steve Hankin, and Luca Cinquini attended the NOAA-funded GIP kickoff meeting held at the Geophysical Fluid Dynamics Laboratory in Princeton, NJ on November 5–6, 2009. The GIP program promotes coordination of software infrastructure development across agencies, the weather and climate communities, modeling and data services, and research and operational centers.

  The Curator project is now hosted under the NOAA GIP program, and it continued a highly productive collaboration with ESG-CET. Curator continued to serve as an active liaison between the EU-based METAFOR project, which is developing a Common Information Model (CIM) for CMIP5/IPCC, and ESG-CET, which will be delivering data to the global community. We have jointly demonstrated new "model trackback" capabilities, which effectively combine the METAFOR questionnaire with the query and browse functions of ESG.

- *Hybrid Coordinate Ocean Model (HyCOM) consortium (NOAA, Navy, et. al.)*

  NOAA/PMEL (Steve Hankin, ESG co-PI) is a partner in the Hybrid Coordinate Ocean Model (HyCOM) consortium [http://hycom.rsmas.miami.edu/]. The HyCOM Consortium has developed a high resolution (1/12 degree) operational, global ocean modeling capability under cooperative U.S. Navy and NOAA funding. The HyCOM model presents unique technical challenges through the complicated coordinate system that it employs and its large data volumes, but the needs of HYCOM overlap in many respects with the ocean components of the climate models to be utilized in IPCC AR5. There is a significant and productive two-way technology transfer of technical capabilities developed in support of ESG and technical capabilities developed in support of HyCOM.

- *NOAA Geophysical Fluid Dynamics Laboratory*

  The NOAA GFDL Fluid Dynamics Laboratory is an active contributor to AR5 and an active participant in the ESG SciDAC. V. Balaji (Head, GFDL Modeling Systems Group) is a frequent participant and active contributor in ESG telephone conferences and meetings leading to a vigorous bi-directional exchange of ideas and technology. NOAA/PMEL (Steve Hankin, ESG site-PI) shares an Memorandum of Understanding (MOU) with GFDL for the development of the LAS and analysis tools, also leading to an active two-way technology transfer between NOAA and ESG.

- *NOAA Office of Climate Observations (OCO)*

  PMEL is the developer of the OSMC on behalf of NOAA/OCO and manages interactive access to the international SOCAT for quality control analysis. Through the PMEL membership in the ESG SciDAC a number of useful collaborative benefits are being explored and are likely to be realized in time for IPCC AR5 work. OSMC and SOCAT are both sources of integrated ocean-climate observations that are potentially useful to IPCC scientists in the evaluation of climate model outputs. PMEL will be helping to bring these collections of observations into the ESG framework for the benefit of IPCC scientists and others.

- *Unidata and the Climate and Forecast Conventions (CF)*

Several ESG members play key roles in the development of the CF conventions—the emerging standard for climate model outputs stored in netCDF. ESG is forging a strong collaborative relationship with Unidata, the development organization for netCDF.

- *US Integrated Ocean Observing System (IOOS)*

  PMEL is a member of the US Integrated Ocean Observing System (IOOS) Integrated Products Team (IPT). IOOS is a potential source of integrated ocean observations that are potentially useful to IPCC scientists in the evaluation of climate model outputs. PMEL will be collaborating with IOOS to locate climate-relevant U.S. coastal observations and bring them into the IPCC framework.

- *Global Earth Observation Integrated Data Environment (GEO-IDE)*

  The PMEL TMAP group put forward at the GEO-IDE meeting that it would lead a small community in the creation of a distributed THREDDS catalog of NOAA gridded data sets. We have already populated this publicly available catalog with data sets from across several different NOAA line offices, including Oceanic and Atmospheric Research (OAR), National Marine Fisheries Service (NMFS), and National Environmental Satellite, Data, and Information Service (NESDIS). This collection provides a rich set of data for inter-comparison and verification with the main ESG-CET collections.

- *ESG-CET Collaborates with CDIAC and ARM Data Centers at ORNL*

  The Carbon Dioxide Information Analysis Center (CDIAC) and the Atmospheric Radiation Measurement Program and the ESG-CET team at ORNL have begun a collaborative pilot project to evaluate the feasibility of integrating high-value observational data sets into ESG. This work has resulted in the successful publication of a number of observational data sets, identification of architectural enhancements within ESG to support observational data sets, and the identification of metadata enhancements that will be required for data discovery.

- *NCDC visit to ORNL to discuss ESG-CET and Observational Data*

  Thomas Karl, Scott Hausman, John Bates, Eileen Shea, and Russell Vose (National Climatic Data Center) visited ORNL to discuss multiple areas of collaboration. John Bates was particularly interested in ESG-CET to enable derivative data products that fuse observational data sets with climate model data. Further collaborations with Helen Frederick and others at NCDC are ongoing with the goal of providing seamless access to data sets at NCDC and ORNL.

- *WMO Information System (WIS)*

  The WMO WIS Program is developing a next-generation globally federated data system that is intended to serve all of the WMO areas (e.g. climate, weather, hydrology, etc.). Don Middleton contributes to several WIS committees and regularly provides briefings that encompass ESG-CET and our progress towards CMIP5/IPCC objectives and climate research in general. Our Gateway technology will provide federation with WIS while also serving as a WIS validation platform.

- *NSF TeraGrid Science Gateways Program*

  ESG-CET, NOAA GIP, and Purdue University staff collaborated to propose the development of a prototype Environmental Science Gateway, and was granted support from TeraGrid's Science Gateway program for two years. The basic objectives of this project include developing a Gateway-based model execution capability, a supporting workflow system that can operate on the TeraGrid, and data publishing and harvesting capabilities that can interoperate with an iRODS-based archive as well as disk-hosted model output.

- ***April 1, 2009 through September 30, 2009***

  - List of collaborations and visualization

To effectively build an infrastructure capable of dealing with petascale data management and analysis, we established connections with other DOE Office of Science SciDAC funded projects and programs at various meetings and workshops, such as the SciDAC Outreach Center Workshop held in San Francisco, California. In particular, collaborations were established with the following groups: TeraGrid Science Gateways, Earth System Curator, NOAA's Global Interoperability Program (GIP), MetaFor, World Meteorological Organization (WMO) - WMO Information System (WIS), Scientific Computing and Imaging (SCI) Institute at the University of Utah, SciDAC VACET, SciDAC SDM, SciDAC CEDPS, Southern California Earthquake Center (SCEC), Tech-X Corp., NASA JPL, NASA Goddard, GO-ESSP, and a whole host of others.
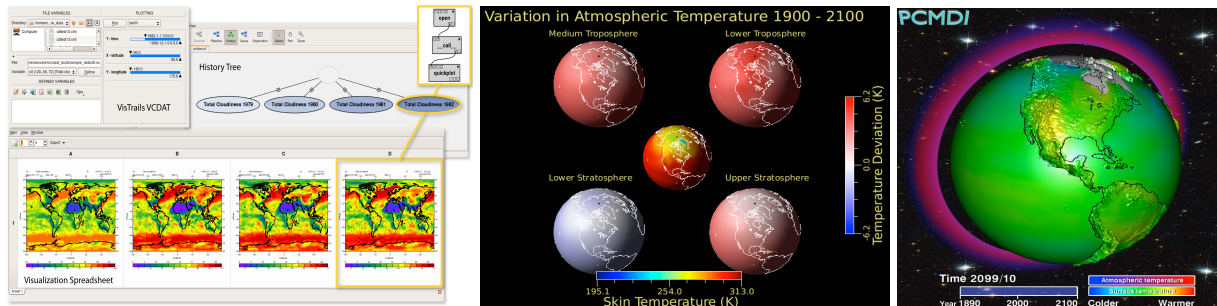


**Example of CDAT workflow built inside of VisTrails, and visualization experiments on how best to show the variations of Microwave Sounding Units (MSU) in a three-dimensional display.**

- *October 1, 2008 through March 31, 2009*

  - *External Support of Grid-Enabled OPeNDAP*

    Prof. Giovanni Aloisio and Dr. Sandro Fiore of Scientific Computing & Operations, CMCC-Euromediterranean Centre for Climate Change, who are working on the Climate-G project, also are collaborating with the Data and Transport ESG-CET team. OPeNDAP efforts at NCAR, focusing on data infrastructure and GridFTP developments, are testing and incorporating into their system.

    Dr. Monique Petitdiddier, Dr. Horst Schwichtenberg, Dr. Wim Som de Cerff (and collaborators) from the EGEE - Enabling Grids for E-sciencE project have been collaborating on numerous aspect of data transport within Grid systems with the ESG-CET team.

  - *University of Utah Visit to Discuss VisTrails, ESG, and CDAT Collaboration*

    As part of the SciDAC VACET collaboration effort, ESG is working closely with the University of Utah to integrate one of the ESG analysis package (i.e., The Climate Data Analysis Tools (CDAT)) into Vistrails (**Error! Reference source not found.**). (VisTrails is an emerging paradigm for capturing complex analysis processes at various levels of details and provides provenance information necessary for reproducible results in a shared environment.)
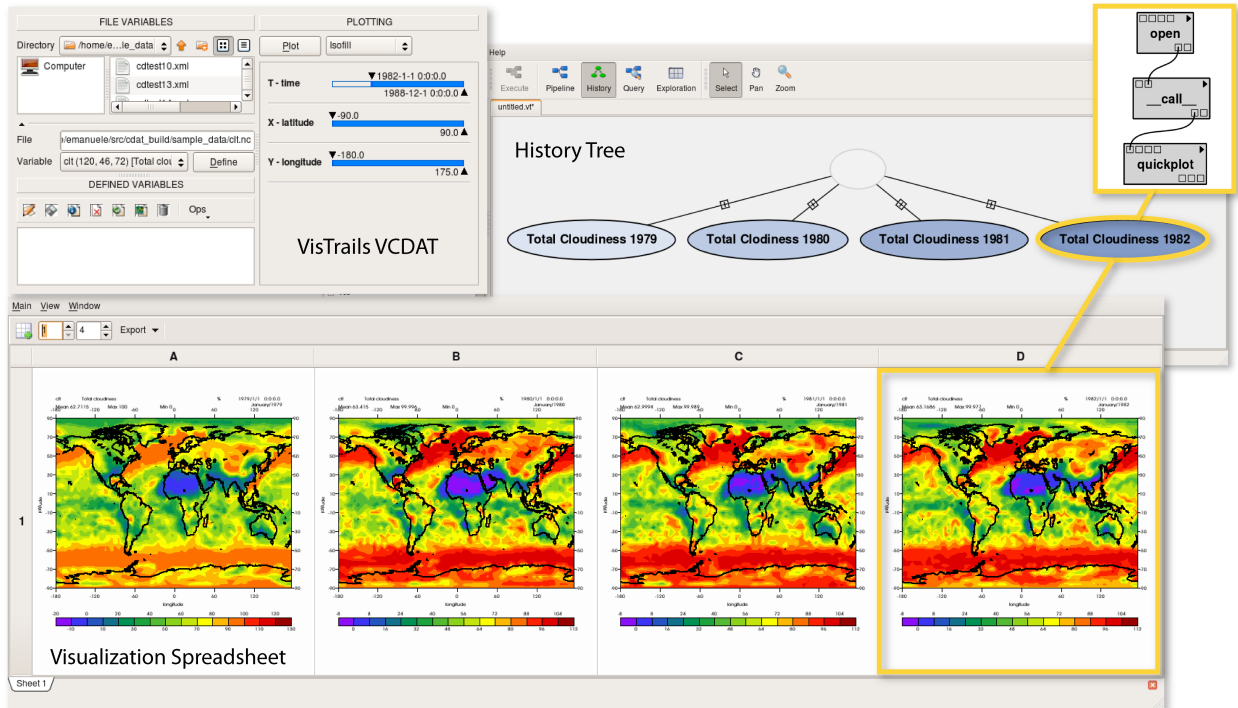
**CDAT VisTrails package usage. Users interact with the VisTrails VCDAT window to manipulate the climate data sets and visualize them in the Visualization Spreadsheet. Provenance is automatically captured as a tree, where each node corresponds to a workflow that was also automatically generated based on the actions performed on the Vistrails VCDAT window.**

- *Collaboration with the SciDAC VACET Visualization Team*

  The ViSUS 2.0 framework has been released as part of latest CDAT (i.e., version 5.0). This result follows from a substantial amount of software engineering effort over the past 18 months. This result is significant because all CDAT users now have access to a solid set of fundamental 3D visualization capabilities, of which will be eventually accessed via ESG (Figure 11 and 13).
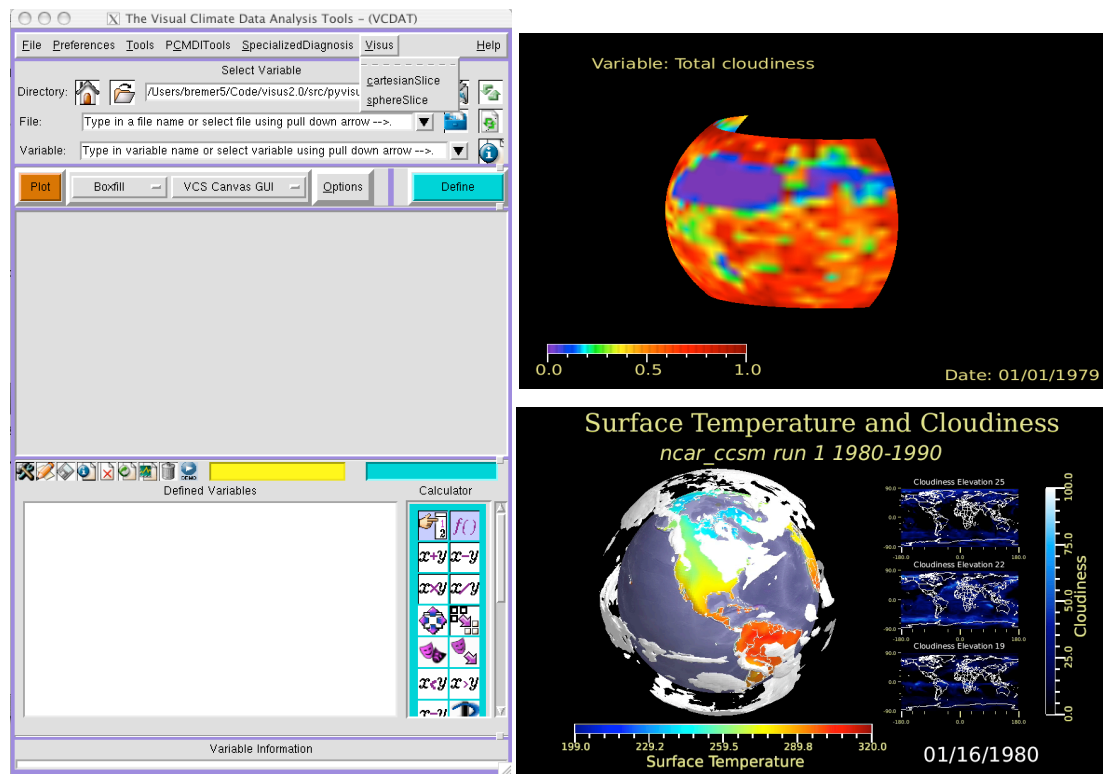
**Custom VCDAT interface showing the new ViSUS capabilities. Two images showing the partial mapping of cloudiness and a scene showing 3D iso-contouring, slicing, and multiple 2D plots.**

- *Global Organization for Earth System Science Portal (GO-ESSP) Workshop*

   Since its conception in 2002, ESG-CET leaders Steve Hankin (2008 workshop host), Don Middleton, and Dean N. Williams (2006 workshop host) have participated in the GO-ESSP steering committee. The GO-ESSP workshop focuses on facilitating the organization and implementation of an infrastructure for full-data sharing among a consortium spanning continents, countries, and intergovernmental agencies.

- *Hybrid Coordinate Ocean Model (HyCOM) consortium (NOAA, Navy, et. al.)*

   NOAA/PMEL (Steve Hankin, ESG co-PI) is a partner in the Hybrid Coordinate Ocean Model (HyCOM) consortium [http://hycom.rsmas.miami.edu/]. The HyCOM Consortium has developed a high resolution (1/12 degree) operational, global ocean modeling capability under cooperative US Navy and NOAA funding. The HyCOM model presents unique technical challenges, through the complicated coordinate system that it employs and its large data volumes, but the needs of HYCOM overlap in many respects with the ocean components of the climate models to be utilized in IPCC AR5. There is a significant and productive two-way technology transfer of technical capabilities developed in support of ESG and technical capabilities developed in support of HyCOM

- *NOAA Geophysical Fluid Dynamics Laboratory*

   The NOAA GFDL Fluid Dynamics Laboratory is an active contributor to AR5 and an active participant in the ESG SciDac. V. Balaji [Head, GFDL Modeling Systems Group] is a frequent participant and active contributor in ESG telcons and meetings leading to a vigorous bi-directional exchange of ideas and technology. NOAA/PMEL (Steve Hankin, ESG co-PI) shares an MOU with GFDL for the development of the Laboratory's data portal, also leading to an active two-way technology transfer between NOAA and ESG.

- *NOAA Office of Climate Observations (OCO)*

  PMEL is the developer of the ocean Observing System Monitoring Center (OSMC) on behalf of NOAA/OCO and manages interactive access to the international Surface Ocean Carbon ATlas (SOCAT) for quality control analysis. Through the PMEL membership in the ESG SciDAC a number of useful collaborative benefits are being explored and are likely to be realized in time for IPCC/AR5 work. OSMC and SOCAT are both sources of integrated ocean-climate observations that are potentially useful to IPCC scientists in the evaluation of climate model outputs. PMEL will be helping to bring these collections of observations into the ESG framework for the benefit of IPCC scientists and others.

- *Unidata and the Climate and Forecast Conventions (CF)*

  Several ESG members play key roles in the development of the CF conventions – the emerging standard for climate model outputs stored in netCDF. ESG is forging a strong collaborative relationship with Unidata, the development organization for netCDF.

- *US Integrated Ocean Observing System (IOOS)*

  PMEL is a member of the US Integrated Ocean Observing System (IOOS) Integrated Products Team (IPT). IOOS is a potential source of integrated ocean observations that are potentially useful to IPCC scientists in the evaluation of climate model outputs. PMEL will be collaborating with IOOS to locate climate-relevant US coastal observations and bring them into the IPCC framework.

- ***April 1, 2008 through September 30, 2008***

  - *University of Utah Visit to Discuss VisTrails, ESG, and CDAT Collaboration*

    As part of the SciDAC VACET collaboration effort, Dean Williams visited the Scientific Computing and Imaging (SCI) Institute at the University of Utah to discuss the possibilities of VisTrails in the ESG framework. (VisTrails is an emerging paradigm for capturing complex analysis processes at various levels of details and provides provenance information necessary for reproducible results in a shared environment.) The LLNL group was able to install VisTrails and implemented a small subset of CDAT commands callable from within the VisTrail visual environment. Although CDAT has been developed with climate applications in mind, most of the concepts of the system can be used to develop a more general data analysis tool by merging the functionalities of CDAT and VisTrails (**Error! Reference source not found.**).
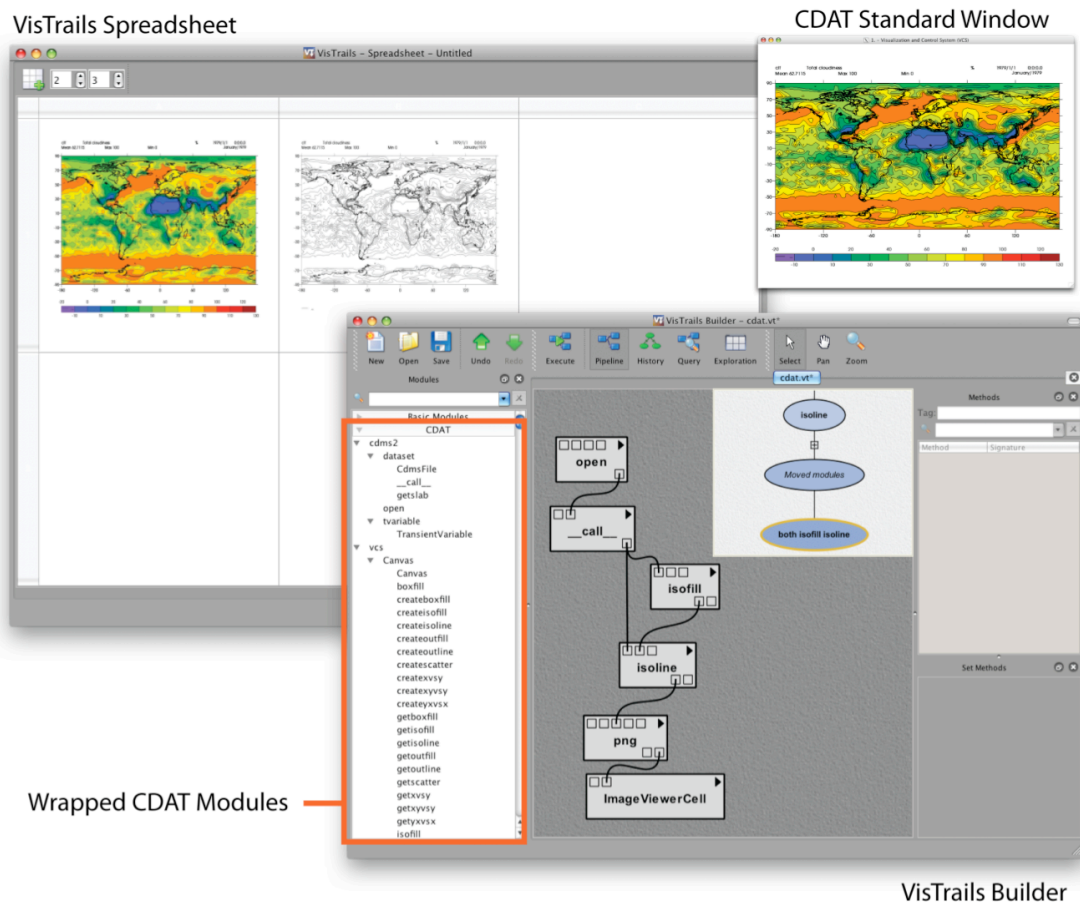
**Example of CDAT workflow built inside VisTrails**

- *Collaboration with the SciDAC VACET Visualization Team*

  During the 2008 SciDAC conference the VACET team demonstrated the new 3D visualization system add-on to the 5.0 beta release of the CDAT package to John Drake and Phil Jones. Both were enthusiastic about the 3D capabilities (Figure 13), their tight integration into the existing pipeline, and especially the flexible annotations offered by the new system. John and Phil re-iterated their concerns about the high spatial resolutions of upcoming simulations and the inability of existing tools to adequately handle these large data sets. All parties agreed that the ViSUS package and its ability to process tera-scale data sets in a streaming and/or out-of-core fashion is an important and welcome addition to the CDAT package.

  Going forward, John and Phil suggested two climate-specific extensions for the current tool chain: Mapped z-elevations and mapped logical grids. Mapped z-elevations are important for climate visualization since most common climate simulation codes operate on slices of constant air pressure rather than constant elevation. While both measurements are related, they are not identical and their mapping must be taken into account for an accurate visualization. The current version of the ViSUS system already has the capability to display height fields, as mapped 2D planes, and this capability will be extended to support general elevation mappings. While older simulation codes used polar grids with discontinuities at both geographic poles, newer packages use warped polar grids that move the singularities to less significant locations (e.g., the middle of Canada), or more generally mapped grids.

  While these structures logically remain regular grids geometrically they can be mapped almost arbitrarily onto the earth depending on the specific needs of a given simulation. Multiple options to support mapped grids in the ViSUS system were discussed with higher dimensional coordinates being the clear favorite.

This concept entails storing the geometric locations of samples along with the standard physical variables (pressure, temperature, etc.). Since the grid logically remains regular, the "enhanced" data set can easily be handle by the existing streaming infrastructure. The mapping can now be seen as a special case of multi-modal visualization where, for example, temperature is one field and x,y,z, positions are a second field. Custom rendering code will then combine both fields to create a mapped temperature field.
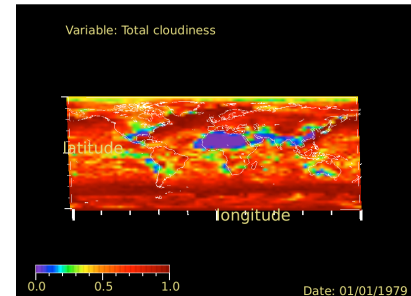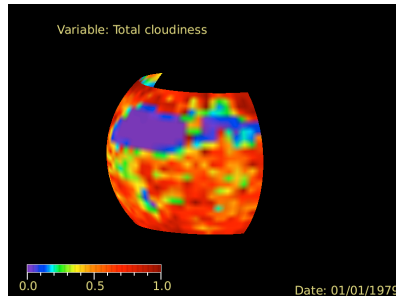


**(a) Earth display using satellite imaging.**  **(b) Partial mapping of cloudiness.**  **(c) Continents rendered using new poly-line display capability.**

**Example 3D Capabilities Provided by the VACET Team**

One of the first new features that the VACET team will introduce is the potential management of multiple meshes with different domains and resolutions that may be incompatible. In particular, the VACET team will work on the visualization of atmospheric and ocean data that are generated on independent domains.

▪ *Collaborations with the Southern California Earthquake Center (SCEC)*

Ann Chervenak initiated discussions on possible collaborations regarding metadata schemas and catalogs with the ESG-CET and the Southern California Earthquake Center (SCEC) project. She also led phone meetings between the two projects. Participants from SCEC included Thomas Jordan (PI), Phil Maechling, David Okawa and Tran Huynh – all from the University of Southern California's Information Sciences Institute (USC/ISI).

▪ *Scientific Data Management (SDM) Center for Enabling Technology (SciDAC CET)*

Based on their experience with DataMover-Lite, Arie Shoshani and his team developed a new client version of an SRM (called "SRM-Lite") in order to move files to and from sites that have highly secure systems by using one-time-passwords. This tool is similar in design to DML, but runs as a client version of SRM to pull/push files from behind a firewall. Unlike DML, SRM-lite cannot use a GUI remotely, and therefore has only a command-line interface. For example, Chi-Fan Shih of NCAR's Data Support Section used this tool to move ESG data from a disk system inside the firewall to one outside the firewall. SRM-Lite, based on high-performance scp (hpn-scp) for higher throughput, is being used successfully to move the large volumes of data.

▪ *Center for Enabling Distributed Petascale Science (CEDPS) (SciDAC CET)*

Collecting statistics on data movement and storage usage is important to ESG. In order to collect such data from BeStMan, Arie Shoshani's team collaborated with members of the Center for Enabling Distributed Petascale Science (CEDPS) project to extract such data automatically from the SRM logs. They have worked with Dan Gunter and Keith Beattie at LBNL for CEDPS troubleshooting to collect logging information into DB, display with NetLogger and web-frontend. They also plan to use their statistics collection tools to automatically generate summary statistics and help with troubleshooting.

ESG also collaborates with CEDPS in the data transfer area, working to define new requirements for GridFTP and for data replication services.

▪ *SBIR CDAT-MODAVE – Collaboration with Tech-X*

Dean Williams is working with Alex Pletzer (from Tech-X) and V. Balaji (from the Geophysical Fluid Dynamics Laboratory, GFDL) on the Mosaic Data Analysis and Visualization Extension (MoDAVE). An SBIR-funded project, MoDAVE started in July 2008 with the aim of extending CDAT to handle the large multi-block data sets that are increasingly becoming the norm, as atmospheric models move away from longitude/latitude based grids (Figure 14). Since its inception, MoDAVE has focused on developing a cubed-sphere netCDF reader using CDAT's Climate Data Management System 2 (cdms2) object to infer the connectivity between mosaic tiles. Knowledge of the tile connectivity is necessary in order to accurately visualize cell-centered data on the cubed-sphere (finite-volume-based discretization leads to cell-centered scalar fields). A general API (mvViz3d) was defined to allow multiple 3D visualization engines to be plugged into CDAT. At present mvViz3d supports two implementations: one based on the Visualization Toolkit (VTK) and the other on VisIt (www.llnl.gov/visit/); both can be used to render 3D iso-surfaces on the sphere. In addition, we will continue to explore ways to parallelize data analysis and visualization so as to be able to handle high-resolution (< 10km) data sets. CDAT is one of several standard analysis tools that must operate on ESG's "Product Services" back-end.
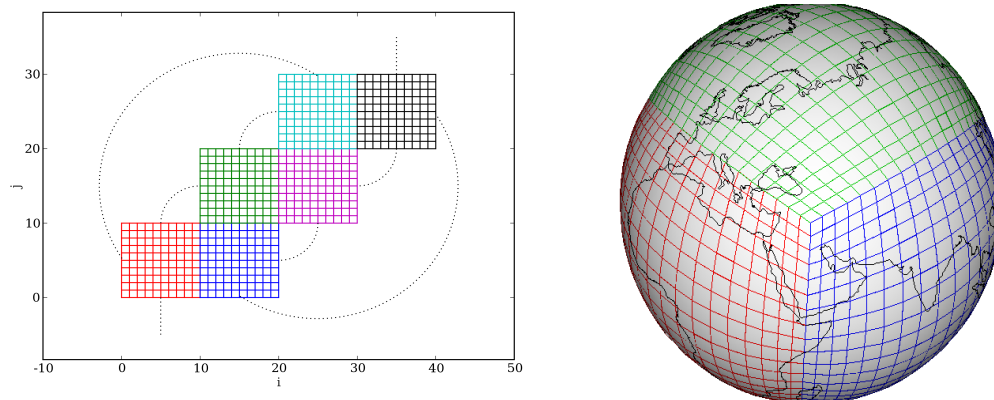


**Example of a Mosaic Grid: Cubed Sphere Connectivity**

- *Global Organization for Earth System Science Portals (GO-ESSP) Workshop*

  Steve Hankin (this year's workshop host), Don Middleton, and Dean N. Williams are three of seven GO-ESSP steering committee members who coordinated the seventh annual GO-ESSP workshop held September 16 – 19 at the Seattle Washington Public Library. In addition, Steve, Don, and Dean chaired six of the seven workshop sessions. The GO-ESSP workshop focuses on facilitating the organization and implementation of an infrastructure for full data sharing among a consortium spanning continents, countries, and intergovernmental agencies. All ESG-CET testbed partners (i.e., LLNL, NCAR, GFDL, BADC, DKRZ, and the University of Tokyo) were present. The workshop, in part, covered testbed security concerns and addressed issues of collaboration. By 2011, this organization envisions allowing users open access to petabytes of multi-model generated data, as well as in-situ, satellite, biogeochemistry, and ecosystems data.

- *Hybrid Coordinate Ocean Model (HyCOM) consortium (NOAA, Navy, et. al.)*

  NOAA/PMEL (Steve Hankin) is a partner in the Hybrid Coordinate Ocean Model (HyCOM) consortium [http://hycom.rsmas.miami.edu/]. The HyCOM Consortium, which has developed a high-resolution (1/12 degree) operational, global ocean modeling capability under cooperative US Navy and NOAA funding. The HyCOM model presents unique technical challenges, such as the complicated coordinate system that it employs and its large data volumes, but the needs of HYCOM overlap in many respects with the ocean components of the climate models to be utilized in CMIP5 (IPCC AR5). A significant and productive two-way technology transfer has been developed in support of ESG-CET, and of HyCOM in particular.

- *NOAA Geophysical Fluid Dynamics Laboratory*

  The NOAA GFDL Fluid Dynamics Laboratory is an active contributor to CMIP5 and an active participant in the ESG-CET. V. Balaji [Head, GFDL Modeling Systems Group] is a frequent participant and active contributor in ESG-CET telcons and meetings leading to a vigorous bi-directional exchange of ideas and technology. NOAA/PMEL (Steve Hankin) shares a Memorandum of Understanding (MOU) with GFDL for the development of the Laboratory's data portal, also leading to an active two-way technology transfer between NOAA and ESG-CET.

- *NOAA Integrated Ocean Observing System (IOOS) and NOAA Office of Climate Observations (OCO)*

  NOAA has been designated as the lead agency in the development of the U.S Integrated Ocean Observing System (IOOS). PMEL is a member of the IOOS Integrated Products Team (IPT) and is also the developer of the ocean Observing System Monitoring Center (OSMC) on behalf of NOAA/OCO. Through the PMEL membership in the ESG-CET a number of useful collaborations benefits are being explored and are likely to be realized in time for the CMIP5 (IPCC AR5) work. IOOS and OSMC are both sources of integrated ocean observations that are potentially useful to IPCC scientists in the evaluation of climate model outputs. PMEL will be helping to bring these collections of observations into the ESG-CET framework for the benefit of IPCC scientists and others.

- *NASA JPL visit to LLNL to discuss ESG-CET and Satellite Data*

  Dan Crichton and Amy Braverman from NASA's Jet Propulsion Laboratory (JPL) are actively pursuing initiatives in climate modeling, and so are interested in establishing an active relationship with ESG. Their recent visit (along with other JPL scientists) to LLNL explored how to enable the climate community to gain access to JPL satellite data, and allow modelers at JPL to gain access ESG data. JPL is now engaged in a pilot activity that demonstrates access between JPL and LLNL and, if successful, will encourage deeper collaborations between NASA and ESG-CET.

- *External Support of Grid-Enabled OPeNDAP*

  Karen Schichardt and Jeff Daily of PNNL, who are working with Dr. David Randall (CSU) on an NSF STC for extreme weather modeling, also are collaborating with the Data and Transport ESG-CET team. OPeNDAP efforts at NCAR, security infrastructure, and GridFTP protocol developments now are helping this group integrate their efforts into the ESG-CET environment.

- **October 1, 2007 through March 31, 2008**

  - *Hybrid Coordinate Ocean Model (HyCOM) consortium (NOAA, Navy, et. al.)*

    NOAA/PMEL (Steve Hankin, ESG co-PI) is a partner in the Hybrid Coordinate Ocean Model (HyCOM) consortium [http://hycom.rsmas.miami.edu/]. The HyCOM Consortium has developed a high- resolution (1/12 degree) operational, global ocean modeling capability under cooperative US Navy and NOAA funding. The HyCOM model presents unique technical challenges, owing to its complicated coordinate system and large data volumes, but the needs of HYCOM overlap in many respects with those of the ocean components of the IPCC AR5 climate models. There is thus a significant and productive two-way transfer of technical capabilities developed in support of ESG and HyCOM.

    The ESG-CET collaboration has worked towards enabling support, within the current ESG operational system, for publishing and distributing NARCCAP (North America Climate Regional Climate Change Project) data. An extensive data management plan was developed that involves distributed data access from the ESG portal at NCAR to data resources stored at both NCAR and PCMDI. The existing user registration system was extended to allow a separate community of NARCCAP users vetted by specific administrators, and the first test users were approved for access.

  - *NOAA Geophysical Fluid Dynamics Laboratory*

The NOAA GFDL Fluid Dynamics Laboratory is an active contributor to AR5 and an active participant in the ESG SciDac. V. Balaji [Head, GFDL Modeling Systems Group] is a frequent participant and active contributor in ESG teleconferences and meetings, resulting in a vigorous bi-directional exchange of ideas and technology. NOAA/PMEL (Steve Hankin, ESG co-PI) shares an MOU with GFDL for the development of the Laboratory's data portal, thereby also implementing an active two-way technology transfer between NOAA and ESG.

- *Global Organization for Earth System Science Portal (GO-ESSP)*

  The GO-ESSP is a collaboration designed to develop a new generation of software infrastructure that will provide distributed access to observed and simulated data from the climate and weather communities. Of the seven members of the GO-ESSP steering committee, three are members of the ESG-CET team: Steve Hankin, Don Middleton, and Dean N. Williams.

- *Earth System Curator (ESC)*

  The ESG-CET and the Earth System Curator (ESC) are working together to develop prototype ontology, user interface, and relational databases to include additional information on the model configurations that produce data sets. ESG-CET team members Luca Cinquini and Don Middleton (as a co-PI) are working closely with this group. Other ESG-CET team members may be involved as work progresses.

- *Scientific Data Management (SDM) Center for Enabling Technology (SciDAC CS CET)*

  Based on the experience with DataMover-Lite, a new client version of an SRM (known as "SRM-Lite") was developed in order to invoke it to move files to and from sites that have one-time-password (OTP) security or other highly secure systems. SRM-Lite is similar in design to DML, but has only a command-line interface. We plan to use this tool in a workflow system that the SDM center uses, called Kepler, as well as other application projects.

- *VACET: VisTrails*

  VisTrails is a new scientific workflow management system. The LLNL team is currently working to modify the Climate Data Analysis Tools (CDAT) XML output to allow the use of VisTrails. If successful, this workflow may be implemented in the next generation of ESG architecture.

- *Institute for Ultrascale Visualization*

  Jian Huang (U. Tennessee), a member of the SciDAC Institute for Ultrascale Visualization, is developing tools for web-based collaborative visualization of climate data, with initial application to data from the ORNL C-LAMP portal. ORNL plans to deploy Huang's visualization server on the ORNL C-LAMP computer, and will work towards integrating it as a "product server" in the next-generation ESG architecture. Huang has been invited to present a paper on this work at the May 2008 International Symposium on Collaborative Technologies and Systems (CTS 2008).

- **April 1, 2007 through September 30, 2007**

  - *North American Regional Climate Change Assessment Program (NARCCAP)*

    The ESG-CET collaboration has worked towards enabling support, within the current ESG operational system, for publishing and distributing NARCCAP (North America Climate Regional Climate Change Project) data. An extensive data management plan was developed that involves distributed data access from the ESG portal at NCAR to data resources stored both at NCAR and PCMDI. The existing user registration system was extended to allow a separate community of NARCCAP users vetted by specific administrators, and the first test users were approved for access.

  - *GO-ESSP Collaboration: Semantic Technologies*

During the past few months, considerable effort was spent in investigating the use of emerging semantic technologies (RDF, OWL, Sesame) to develop the next generation of ESG-CET services for search and discovery of scientific data. Prototype search services and interfaces were set up against the current IPCC, CCSM and PCM metadata holdings in order to test the performance, flexibility, and scalability of this approach. Although the first results in this area are encouraging, work is still underway.

More recently, discussions have taken place with the Earth System Curator (ESC) collaboration, which has decided to leverage this prototype ESG-CET infrastructure to provide powerful detailed search capabilities for climate models and their components, as described by the extensive ESC metadata schema. The plan is for ESC to reuse the existing ESG-CET semantic service and persistence layers, collaborating to extend the current ESG-CET ontology with additional classes and properties, while at the same time adding custom functionality for compatibility checking among model components. A meeting will be held at GFDL in mid-October 2007 to assess progress and to plan for the next phases of the collaboration between the two projects.

- *IO Strategies and Data Services for Petascale Data Sets from a Global Cloud Resolving Mode Collaboration*

  The ESG executive committee has met with Karen Schuchardt (the SAP PI on Global Cloud Resolving Models) on numerous occasions, outlining the strategy for working together as a team. More recently at the Climate Change Prediction Program (CCPP) conference in Indianapolis, Karen and Dean discussed working more closely at the PI level. The general agreement is to include Karen, once a month, on ESG executive committee meetings (starting in October). This will keep her abreast of ESG activities and help ESG leverage work completed by her team. We also discussed pairing members of her group with working groups already established in ESG: the Metadata Work Group (i.e., working with Bob and Luca on metadata schemas, RDF, etc.), and the User Interface Working Group (i.e., working with Jens and other doing ESG user interface development). Also planned is providing help for the LLNL team to extend CDAT to support a geodesic grid, which also involves Geophysical Fluid Dynamics Laboratory (GFDL) gridspec work. (The results of the gridspec effort, led by V. Balaji at GFDL, will be implemented in the netCDF Climate and Forecast (CF) convention.) In addition, LLNL team members will also discuss the Climate Model Output Rewriter (CMOR) and how to improve processing data for model intercomparisons such as CMIP3 (IPCC AR4).

- *Atmospheric Radiation Measurement (ARM) Collaboration*

  The team at Argonne has started collaborating with Environment Science Division at ANL, specifically to work with scientists at Climate Research Station on the Data Domain to Model Domain Conversion Package (DMCP) (see URL: http://www.atmos.anl.gov/DMCP/). This recently initiated effort has been exploring ways to publish subsets of ARM data with mechanisms to support useful parameter-based server-side processing of data. The collaboration also will investigate options to allow publishing the resulting data as an independent data set.

  A test installation of Live Access Server (LAS) has been set up and work is ongoing to evaluate the upload, visualization and processing of a sample subset of ARM data. The results from the evaluation of the prototype will be used in the design and implementation of server-side processing on ESG systems. (See section 2.6.)

- *Hybrid Coordinate Ocean Model (HyCOM) consortium (NOAA, Navy, et. al.)*

  NOAA/PMEL (Steve Hankin, ESG co-PI) is a partner in the Hybrid Coordinate Ocean Model (HyCOM) consortium (see URL: http://hycom.rsmas.miami.edu/). The HyCOM Consortium is developing a high resolution (1/12 degree) operational global ocean modeling capability under cooperative US Navy and NOAA funding. The HyCOM model presents unique technical challenges, through the complicated vertical coordinate system that it employs, but its needs overlap in many respects with the ocean components of the climate models to be utilized in CMIP4 (IPCC AR5). There is a significant and

productive two-way transfer of technical capabilities developed in support of ESG and of HyCOM. (See Figure 4, showing the HyCOM model intecomparison.)



**LAS Slide Sorter output showing the HyCOM model intercomparison**

- *NOAA Geophysical Fluid Dynamics Laboratory*

   The NOAA Geophysical Fluid Dynamics Laboratory (GFDL) is an active contributor to CMIP4 (IPCC AR5) and an active participant in the ESG-CET.  V. Balaji (Head, Modeling Systems Group at GFDL) is a frequent participant and active contributor in ESG telcons and meetings, resulting in a vigorous bi-directional exchange of ideas and technology.  NOAA/PMEL (Steve Hankin, ESG co-PI) shares an MOU with GFDL for the development of the Laboratory's data portal, also effecting an active two-way technology transfer between NOAA and ESG.

- *Scientific Data Management (SDM) Center for Enabling Technology (SciDAC CS CET)*

   Similar to the DataMover-Lite (DML) client component in ESG, the SDM center has identified a need for moving files to and from sites that have one-time-password (OTP) security or other highly secure systems.  The intention is to have an SRM client program at the secure sites that communicate commands and data through SSH.  The SDM center has developed a prototype version of this client program, called SRM-Lite, and is planning to use this technology for a combustion project in the near future.

- *VACET Collaboration: VisTrails*

   VisTrails is a new scientific workflow management system.  While originally (and solely) developed by researchers at the University of Utah to provide support for data exploration and visualization, VisTrails now is being applied to climate data analysis and visualization as part of the SciDAC-2 Visualization and Analytics Center for Enabling Technology (VACET) collaboration. The image below shows the use of the visual workflow interface to connect CDAT module boxes to perform calculations and a related plot.
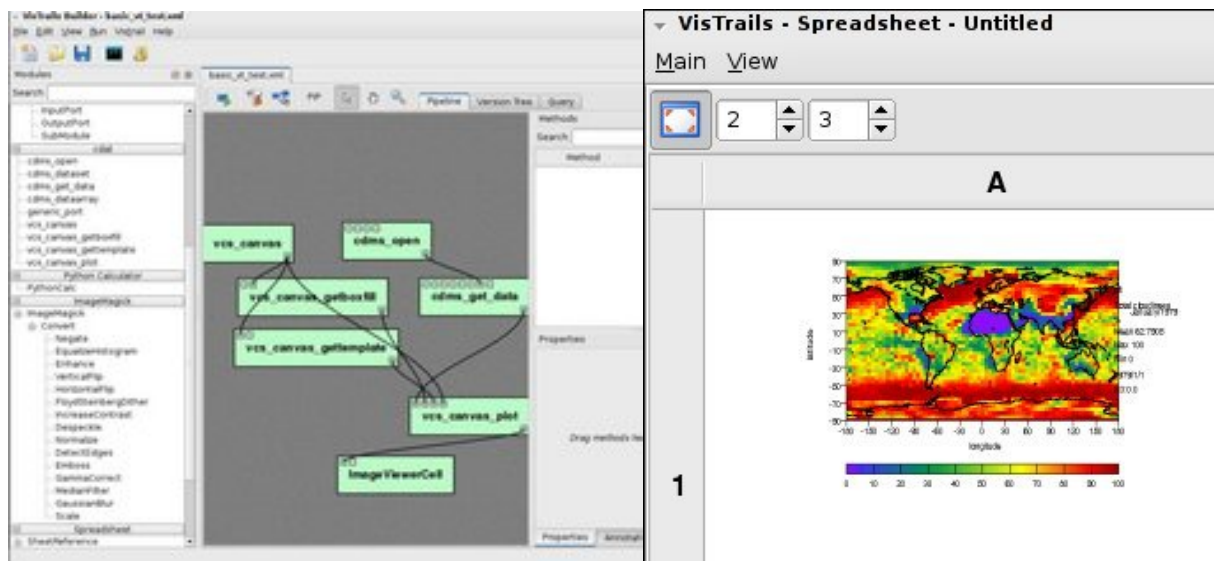
**The result of the CDAT run viewed in VisTrails showing results in a spreadsheet application**

Work on this new GUI application interface for climate data analysis and exploration continues in collaboration with the VACET team.

▪ *VACET Collaboration: 3D Visualization*

In its collaboration with VACET, the ESG-CET team has worked to produce several compelling, high-quality, 3D images that will be reproducible by any scientists who have access to ESG-CET's computational resources to do ground-breaking 3D visualization and computing. Initially, these images would lend themselves to the creation of "glitzy" movies used for general public consumption. In the future, we aim for scientists to produce these images in pursuit of understanding key climate science questions. The visualization appearing below shows surface temperature, atmospheric temperature, and sea ice and cloud coverage on an elevated Earth model In this example, the data (e.g., surface temperature) represent the combined average influence of an ensemble of all the climate models that are available in the CMIP3 (IPCC AR4) data archive. The animation over time shows an upward climate temperature trend, indicative of global warming. This visualization/animation example was computed on 200 processors in about 15 minutes using custom visualization software that will be integrated into climate analysis tools.
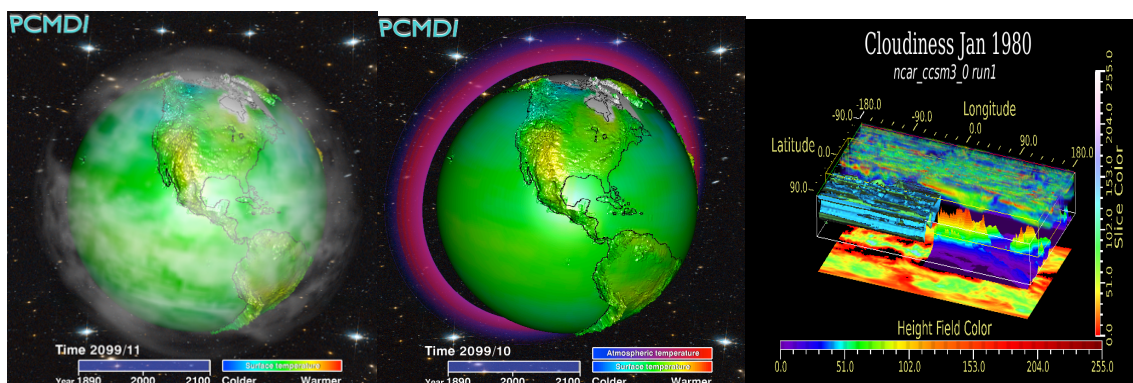


**Working with VACET developers to make 3D graphics accessible to the climate community**

- ***October 1, 2006 through March 31, 2007***

  - *A Scalable and Extensible Earth System Model for Climate Change Science (SciDAC Climate Application)*

    Collaborating with the SciDAC science application project A Scalable and Extensible Earth System Model for Climate Change Science (SEESM), ESG-CET has agreed to address the challenges of enabling data management, discovery, access, and advanced data analysis for their community. In constant contact with the SEESM PI, John Drake, and co-PI, Phil Jones, ESG-CET will act as the liaison for this group with other SciDAC projects (such as VACET and IUSV).

  - *Intergovernmental Panel on Climate Change Fifth Assessment Report (IPCC AR5, United Nations Environment Program)*

    Dean Williams serves as the liaison between the current AR4 Working Group I committee leaders, led by Jerry Meehl, and the future AR5 Working Group I committee, to be led by Ron Stouffer. In October, Dean met with Ron and others to discuss proposed scenarios/experiments for AR5 and to work out strict deadlines for the project.

  - *Computational Climate End Station (ORNL Leadership Computing Facility)*

    David Bernholdt (ORNL) serves as liaison with the Computational Climate End Station at the Leadership Computing Facility at ORNL.  He is in regular contact with the CCES Chief Scientist John Drake (ORNL) and the leader of the C-LAMP project, Forrest Hoffman (ORNL).  As noted above, a portal has been established at ORNL for the C-LAMP project, and it is expected to being receiving data in April 2007.

  - *Open Science Grid (OSG, SciDAC Physics SAP) and the Center for Enabling Distributed Petascale Science (CEDPS, SciDAC CS CET)*

    The three-way group of OSG, CEDPS, and ESG has been working together to define areas of interaction and complementary work. Don Middleton (NCAR) met with Thomas Ndousse and colleagues from OSG and CEDPS in a March half-day meeting at the San Diego Supercomputing Center (SDSC) in order to advance this activity.

  - *Visualization and Analytics Center for Enabling Technologies (VACET, SciDAC Visualization CET)*

    LANL team member (Phil Jones) and LLNL team members (Bob Drach, and Dean Williams) participated in the VACET project-wide meeting held in Utah. Science projects supported by ESG and the SciDAC awarded climate application were discussed and plans for collaborations are in progress. Integration of the VACET product visTrails and an ESG supported client application is now under way. VisTrails utilizes a visualization pipeline architecture that provides infrastructure to combine existing client applications into one environment. A particularly powerful use of visTrails is parameter exploration.

  - *Workshop (DOE Cybersecurity R&D Challenges for Open Science: Developing a Roadmap and Vision)*

    Representing ANL, ESG, and NCAR, and the science/data community in general, Middleton & Ananthakrishnan attended the DOE-sponsored Cybersecurity Workshop, which was held in January 2007 at AGU headquarters in Washington, D.C. Middleton provided a short position paper along with concerns and scenarios. Rachana presented the whitepaper "Trust-Root Provisioning and Validation Facilities" by Frank Siebenlist, Michael Helm, Rachana Ananthakrishnan, and Ian Foster, which discusses deployment concerns for DOE projects like ESG.

  - *TeraGrid (NSF)*

    Don Middleton continued to participate in TeraGrid meetings and activities, including the extension of ESG to function as a science gateway utilizing TG resources. Middleton attended the January 2007 TG Data Workshop in San Diego, CA.

- *Community Access to Global Cloud Resolving Model Data And Analyses  (SciDAC Climate SAP)*

   Don Middleton (NCAR) attended the NSF CMMAP (Center for Multi-scale Modeling of Atmospheric Processes) in Kauai, Hawaii, meeting with colleagues Karen Schuchardt (PNNL) and John Helley (SDSC). This group worked on plans to develop an overall program plan that was complimentary relative to our specific project goals across SciDAC-2 and NSF Center for Enabling Technology goals.

- *A Data Domain to Model Domain Conversion Package (DMCP) for Sparse Climate Related Process Measurements (SciDAC Climate SAP)*

   ANL and ORNL team members (Dan Fraser and David Bernholdt) met with members of the DMCP project in late February and early March to explore possibilities for collaboration.  Since much of what was discussed would take us beyond the current scope and funding of ESG-CET, we are currently deciding how to proceed.

- *Scientific Data Management Center for Enabling Technology (SciDAC CS CET)*

   The experience with the DataMover-Lite (DML) client component in ESG, led us to consider its use in the SDM center for moving files from sites that have one-time-password (OTP) security or other highly secure systems.  Rather than "pulling" files into the client's site as done in ESG, the SDM center will use a DML for "pushing" files out of highly secure sites to their destinations.  The SDM center is planning to use this technology for a combustion project as well as for part of the workflow in a fusion project.

- *North American Regional Climate Change Assessment Program (NARCCAP, multi-agency)*

   Don Middleton (NCAR) continued to serve in the co-PI role for this project which will host its data products via the ESG system. During this reporting period, we refined the project's data management plan and began to develop the required software and system infrastructure.

## Appendix B        Outreach, Papers, Presentations, Poster, Books

To effectively build an infrastructure capable of dealing with petascale data management and analysis, we had established connections with other funded DOE Office of Science SciDAC projects and national and international programs and agencies at various meetings, workshops and conferences. This section describes many of the substantive interactions we've had during the life of the ESG-CET funded project.

### B.1   Outreach Activities

- ***April 1, 2011 through September, 30, 2011***

- [Add latest posters]

- ***October 1, 2010 through March 31, 2011***

- ***April 1, 2010 through September 30, 2010***

  - *BER Network Requirement Workshop*

    Dean N. Williams, Gary Strand, and Galen Shipman, "Climate Science: Enabling Worldwide Access to Petascale Climate Data," Proceedings of the Office of Biological and Environmental Research Network Requirements Workshop, April 29–30, 2010. According to the report, the biggest BER user of ESnet in the future will likely be the climate community. Through ESG, LLNL and ORNL will host petabyte archives that will be accessible worldwide.

- ***October 1, 2009 through March 31, 2010***

  - *SC-09 High Performance GridFTP Transport of ESG Data*

    The Supercomputing 2009 (SC'09) Bandwidth Challenge entry titled, "High Performance GridFTP Transport of Earth System Grid (ESG) Data", demonstrated high performance GridFTP transport of climate data from multiple Department of Energy laboratory locations to the targeted SC showroom floor. The transferred multi-terabyte data consisted of a small portion of the multi-model Coupled Model Intercomparison Project, Phase 3 (CMIP-3) data set used in the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (AR4). SC'09 conference proceedings.

  - *Lessons learned from moving Earth System Grid data sets over a 20 Gbps wide-area network*

    Raj Kettimuthu, Alex Sim, Dan Gunter, Bill Allcock, Peer-Timo Bremer, John Bresnahan, Andrew Cherry, Lisa Childers, Eli Dart, Ian Foster, Kevin Harms, Jason Hick, Jason Lee, Michael Link, Jeff Long, Keith Miller, Vijaya Natarajan, Valerio Pascucci, Ken Raffenetti, David, Ressman, Dean Williams, Loren Wilson, and Linda Winkler, "Lessons learned from moving Earth System Grid data sets over a 20 Gbps wide-area network", http://esg-pcmdi.llnl.gov/publications_and_documents/HPDC10_BWC_Final.pdf/view.

  - *Experience Report on the Supercomputing 2009 Bandwidth Challenge from the Perspective of the Earth System Grid*

    Alex Sim and Dean N. Williams, "Experience Report on the Supercomputing 2009 Bandwidth Challenge from the Perspective of the Earth System Grid", Supercomputing Conference '09 Bandwidth Challenge, December 2009, http://esg-pcmdi.llnl.gov/publications_and_documents/ExpReports-ESG-SC09_final.pdf/view.

- ### *April 1, 2009 through September 30, 2009*

  - *WCRP/WGCM Meeting*

    Dean Williams and Bob Drach attending the 13th Session of the Working Group on Coupled Modeling (WGCM) meeting in San Francisco, California. The meeting included preparations for CMIP5 model experiments and the deliverance of the data to the climate community. Dean gave an update of ESG-CET to the WCRP/WGCM organizing committee.

  - *Unidata and the Climate and Forecast Conventions (CF)*

    Several ESG members play key roles in the development of the CF conventions – the emerging standard for climate model outputs stored in netCDF. ESG is forging a strong collaborative relationship with Unidata, the development organization for netCDF.

  - *SciDAC 2009 Conference*

    *Don* Middleton and David Bernholdt attended the annual SciDAC 2009 conference held June 14 – 18 2009 in San Diego, California. They met with other ESG collaborators and discussed development possibilities with stakeholders.

  - *Supercomputing 2009 (SC09) Bandwidth Challenge*

    Since July, members of the ESG-CET team, including Dean Williams, Alex Sim, Rachana Ananthakrishnan, and Ian Foster, have been participating in the SC09 bandwidth challenge titled, "High Performance GridFTP Transport of Earth System Grid (ESG) Data." The goal is to demonstrate high performance and reliable GridFTP transport of Earth System Grid data between multiple source locations and the showroom floor. The objective is to move 10 TB of CMIP3 data in under two hours. This activity is in collaboration with ESnet, SDM, Globus, and VACET.

  - *UK e-Science Envoy Visit to Chicago and NCAR*

    Ian Foster hosted a visit from Dr. Malcolm Atkinson and Dr. David de Roure to Chicago, and Don Middleton hosted a visit at NCAR, who were touring the U.S. in pursuit of information regarding scientific research data and use. Foster represented ESG in Chicago; in Boulder, Middleton and Strand represented ESG and McGinnis represented the affiliated NARCCAP project.

  - *SciDAC Outreach Center*

    Dean Williams attended the 3rd Workshop on HPC Best Practices: HPC Center Software Lifecycles hosted by the SciDAC Outreach Center. The workshop was held September 28-29 in San Francisco, California. Dean was a key speaker, presenting "The Earth System Grid: Lessons Learned."

  - *GFDL Meeting on Climate Services*

    Don Middleton attended this one-day meeting representing ESG-CET. Due to a major fire at the center, the group had to abandon the facility and was not able to work through goals and plans; this will presumably happen later.

  - *USGS Visit to NCAR*

    Luca Cinquini gave a presentation on ESG-CET, which was well received by USGS representatives visiting NCAR to explore data services alternatives and potential collaborations. The visit led to a follow-up conference call with ESG, NCAR, and USGS representatives to further explore possibilities for substantial collaboration as USGS broadens and strengthens its emphasis on climate-related activities.

  - *Computing in the Atmospheric Sciences, Annecy, France, September 2009:*

    While there was not a dedicated presentation from ESG-CET at this event, ESG was prominently featured in a talk on the TeraGrid Science Gateway Program given by Nancy Wilkins-Diere of the San Diego

Supercomputing Center (SDSC). ESG-CET was also called out in Warren Washington's presentation on climate change.

- *NOAA Geophysical Fluid Dynamics Laboratory*

    The NOAA GFDL Fluid Dynamics Laboratory is an active contributor to AR5 and an active participant in the ESG SciDac. V. Balaji (Head, GFDL Modeling Systems Group) is a frequent participant and active contributor in ESG telcons and meetings leading to a vigorous bi-directional exchange of ideas and technology. NOAA/PMEL (Steve Hankin, ESG co-PI) shares an MOU with GFDL for the development of the Laboratory's data portal, also leading to an active two-way technology transfer between NOAA and ESG.

- *NOAA Office of Climate Observations (OCO)*

    PMEL is the developer of the ocean Observing System Monitoring Center (OSMC) on behalf of NOAA/OCO and manages interactive access to the international Surface Ocean Carbon ATlas (SOCAT) for quality control analysis. Through the PMEL membership in the ESG SciDAC a number of useful collaborative benefits are being explored and are likely to be realized in time for IPCC/AR5 work. OSMC and SOCAT are both sources of integrated ocean-climate observations that are potentially useful to IPCC scientists in the evaluation of climate model outputs. PMEL will be helping to bring these collections of observations into the ESG framework for the benefit of IPCC scientists and others.

- *Hybrid Coordinate Ocean Model (HyCOM) consortium (NOAA, Navy, et. al.)*

    NOAA/PMEL (Steve Hankin, ESG co-PI) is a partner in the Hybrid Coordinate Ocean Model (HyCOM) consortium [http://hycom.rsmas.miami.edu/]. The HyCOM Consortium has developed a highresolution (1/12 degree) operational, global ocean modeling capability under cooperative US Navy and NOAA funding. The HyCOM model presents unique technical challenges, through the complicated coordinate system that it employs and its large data volumes, but the needs of HYCOM overlap in many respects with the ocean components of the climate models to be utilized in IPCC AR5. There is a significant and productive two-way technology transfer of technical capabilities developed in support of ESG and technical capabilities developed in support of HyCOM.

- *US Integrated Ocean Observing System (IOOS)*

    PMEL is a member of the US Integrated Ocean Observing System (IOOS) Integrated Products Team (IPT). IOOS is a potential source of integrated ocean observations that are potentially useful to IPCC scientists in the evaluation of climate model outputs. PMEL will be collaborating with IOOS to locate climate-relevant US coastal observations and bring them into the IPCC framework.

- *Global Earth Observation Integrated Data Environment (GEO-IDE)*

    The PMEL TMAP group put forward at the GEO-IDE meeting that it would lead a small community in the creation of a distributed THREDDS catalog of NOAA gridded data sets. We have already populated this publicly available catalog with data sets from across several different NOAA line offices, including OAR, NMFS and NESDIS. This collection provides are rich set of data for inter-comparison and verification with the main ESG-CET collections.

▪ ***October 1, 2008 through March 31, 2009***

- *Advanced Scientific Computing Research (ASCR) Networking Requirements Workshop, Washington D.C., April 15-16, 2009*

    Dean N. Williams represented the Earth System Grid Center for Enabling Technologies (ESG-CET), where ESG-CET was one of the case studies examined at the workshop. The workshop goal was to accurately characterize the networking requirements of science funded by the ASCR Program Office.

- *Visit to CMCC*

  Ian Foster visited the Italian Euro-Mediterranean Center for Climate Change (CMCC) to participate in their scientific advisory board. He discussed opportunities for collaboration in distributed data management.

- *Institute Simon Pierre Laplace, Paris, France, March 27, 2009 and ORAP Workshop, Lille, France, March 26, 2009*

  Don Middleton gave an invited presentation at the bi-annual ORAP workshop; a France-centered activity aimed at advocacy in the high-performance and large data area, and with goals of becoming a more general E.U. initiative. Middleton's presentation was entitled *"Building a Global Data Federation for Climate Change Science: The Earth System Grid and International Partners"*. After this event in Lille, France, Middleton traveled to Paris and gave the presentation again to a climate modeling and IT group with the Institute Simon Pierre Laplace (IPSL) at the University of Paris Jussieu campus. This meeting also included discussion of the upcoming CMIP-5 process, and how IPSL would collaborate with ESG to function as a data provider.

- *Open Science Grid Workshop, Baton Rouge, Louisiana, March 3-5, 2009*

  Don Middleton gave an invited presentation at the OSG Workshop entitled *"The Earth System Grid and Global Federation of Climate Model Data"*. The presentation combined an overview of the science domain (OSG participants were mostly in the physics area), ESG's goals, and potential for collaboration with OSG in software delivery.

- *DAARWG Workshop in Saint Louis, Mississippi, February 3-5, 2009*

  Dean N. Williams is a member of the NOAA Advisory's Board Data Archive and Access Requirements Working Group (DAARWG), an important working group of NOAA's Science Advisory Board (SAB). DAARWG evaluates data archiving and access requirements from all of NOAA's observing systems and computational models, as well as from relevant non-NOAA sources. Its charter is to provide scientific advice and broad direction regarding the wide range of data, information, and products that NOAA should archive, and ways in which this agency can best provide access to them.

- *NASA Goddard Institute for Space Studies (GISS) Workshop on Advancing the Scientific Value of the CMIP/IPCC Process, New York City, NY, November 17-19, 2009*

  Don Middleton attended this meeting representing ESG and other projects, and led a discussion session on data management and access, including a briefing to the attendees on ESG-CET.

- ### April 1, 2008 through September 30, 2008

  - *Cyberinfrastructure Summer Institute for Geoscientists (CSIG), San Diego, CA., August 11-15, 2008*

    Don Middleton gave an invited presentation entitled *"Cyberinfrastructure for Global Earth System Data Federation"*. The presentation had a primary focus on ESG, along with additional information on related projects and data analysis and visualization capabilities.

  - *World Meteorological Organization (WMO) Information System Intercommission Coordination Group (ICG-WIS), Brazilia, Brazil, July 14-17, 2008*

    Don Middleton serves as an expert advisor to this high-level coordination group for the development of next generation globally federated data systems for climate, weather, hydrology, oceanography, and other U.N./WMO program elements. Middleton attended a meeting of this group in August in Brazilia, Brazil and verbally briefed the council on ESG-CET activities and implications related to the next IPCC process and related federated data systems.

- *DAARWG Workshop in Ashville, NC*

  Dean Williams accepted the invitation to join the NOAA Advisory's Board Data Archive and Access Requirements Working Group (DAARWG), an important working group of NOAA's Science Advisory Board (SAB). DAARWG evaluates data archiving and access requirements from all of NOAA's observing systems and computational models, as well as from relevant non-NOAA sources. Its charter is to provide scientific advice and broad direction regarding the wide range of data, information, and products that NOAA should archive, and ways in which this agency can best provide access to them.

- *Global Earth Observing System of Systems (GEOSS)*

  GEOSS is a primary thrust of the Group for Earth Observations (GEO), which is a United Nations program coordinated under the auspices of the World Meteorological Organization (WMO). Don Middleton attended several days of GEOSS meetings in Boulder, CO. in September 2008. Middleton also gave an invited presentation entitled "Global Federated Climate and Weather Data Systems" and focused on ESG-CET in the area of climate and the THORPEX Interactive Grand Global Ensemble (TIGGE) in the area of weather.

- *Steering Committee for HPDG08*

  Peter Fox is a member of the Steering and Programme Committee for HPDG08 (High Performance Data Grid 2008) to be held in Dunedin, New Zealand Dec 3-4, 2008.

- *EGU Session Convener*

  Peter Fox convened a European Geophysical Union meeting session on "The Grid for Geosciences and Geoscience Applications for the Grid". Fox was co-convener of the Geosciences applications on Grid session also at EGU 2009.

- *ESIn Editor*

  Peter Fox is an *Earth Science Informatics* (ESIn) editor for a special issue on "Grid computing for Geosciences, Earth Science Informatics".

- *World Meteorological Organization Expert Team on GISC's and DCPC's Meeting, Darmstadt Germany, June 6-12, 2008*

  Don Middleton serves on this WMO expert team, and gave a presentation at the annual meeting that included a status update on ESG.

- *Ocean Observatory Initiative*

  Ian Foster accepted an invitation to join the Advisory Committee for the NSF Ocean Observatory Initiative. He attended his first meeting in Dallas Texas and was able to discuss opportunities for joint work with ESG-CET.

- ***October 1, 2007 through March 31, 2008***

  - *ASCAC-BERAC Charge Committee*

    Ian Foster and Dean Williams served on the Joint Advanced Scientific Computing Advisory Committee (ASCAC) and the Biological and Environmental Research Advisory Committee (BERAC) Subcommittee to investigate barriers and bottlenecks to achieving successful outcomes of complementary investments—specifically in climate modeling by ASCR and BER. Ian and Dean contributed to a 19-page report that was transmitted to Ray Orbach, Director of the DOE Office of Science.

- *Global Organization for Earth System Science Portal (GO-ESSP) Mini-Workshop*

  In mid-October, Steve Hankin, Don Middleton, and Dean N. Williams attended the GO-ESSP mini-workshop which focused on establishing open access to distributed climate data archives.

- *SDM Center All-hands Meeting*

  Hosted by Arie Shoshani, Don Middleton and Dean Williams attended the SDM Center All-hands Meeting where climate analysis and workflow use cases where discussed. Don and Dean represented ESG-CET and made technology collaboration contacts (e.g., for Kepler and VisTrails).

## *April 1, 2007 through September 30, 2007*

- *Presentation: Co-Chair of the GO-ESSP Workshop in Paris, France*

  As Principal Investigators and members of the organizing committee, Dean Williams, Don Middleton, and Steve Hankin attended the 6th Annual Global Organization for Earth System Science Portal (GO-ESSP) Workshop promoting this effort's goals and objectives. The GO-ESSP is a collaboration designed to develop a new generation of software infrastructure that will provide distributed access to observed and simulated data for the climate and weather communities. GO-ESSP will achieve this goal by developing individual software components and by building a federation of frameworks that can work together using standards agreed upon by its participants. The GO-ESSP portal frameworks will provide efficient mechanisms for data discovery, access, and analysis of the data. Participants shared their progress in developing software infrastructure that facilitated discovery, acquisition, and analysis of climate date. Particular interest was expressed on current and future integration activities that facilitate community analysis of widely distributed climate data archives (e.g., CMIP3 (IPCC AR4) and CMIP4 (IPCC AR5)).

- *SciDAC 2007 Organizing Committee*

  Ian Foster and Dean Williams served on the SciDAC 2007 organizing committee, which selected topics that represent state-of-the-art for a given scientific area and suggested appropriate speakers on each topic. Ian was the committee organizer for "Grids/Networking", and Dean served as both the committee organizer for the "Climate Community" and as a "Session Chair" at the conference. The OC also suggested topics and presenters for invited poster sessions. For each topic area, the respective OC member was responsible for peer-review presenter abstracts before the conference, and of proceedings papers immediately after the conference.

## *October 1, 2006 through March 31, 2007*

- *WMO Information System (WIS)*

  Don Middleton (NCAR) serves on multiple WMO WIS committees, and provided updates and overviews on ESG-CET at an intercommission coordination group in Beijing, China, and again at the international "Extraordinary Assembly" of WMO, a technical conference for information technology, held in Seoul, Korea November 2007.

## B.2   Papers

## *April 1, 2011 through September, 30, 2011*

[Add latest posters]

- ***October 1, 2010 through March 31, 2011***

  - *IEEE International Conference on Cloud Computing Technology and Science, 2010*

    "*Finding Tropical Cyclones on a Cloud Computing Cluster: Using Parallel Virtualization for Large-Scale Climate Simulation Analysis,*" D. Hasenkamp, A. Sim, M. Wehner, K. Wu, Proceedings of the 2nd IEEE International Conference on Cloud Computing Technology and Science, 2010.

  - *ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis (SC'10), 2010*

    "*Finding Tropical Cyclones on Clouds,*" D. Hasenkamp, under the supervision of A. Sim, M. Wehner, K. Wu, ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis (SC'10), 2010.

  - *22nd International Conference on Parallel and Distributed Computing and Systems*

    "*Adaptive Transfer Adjustment in Efficient Bulk Data Transfer Management for Climate Datasets,*" A. Sim, M. Balma[n,] D. Williams, A. Shoshani, V. Natarajan, Proceedings of the 22nd International Conference on Parallel and Distributed Computing and Systems (PDCS 2010), 2010.

  - *27th Conference on Interactive Information Processing Systems (IIPS)*

    "*The Interactive Earth Science Data Visualization Gallery (vizGal),*" Roland Schweitzer, Weathertop Consulting, LLC, College Station, TX; and S. Hankin, J. Malczyk, A. Manke, and K. M. O'Brien, 27th Conference on Interactive Information Processing Systems (IIPS), AMS Annual Meeting, Seattle, Washington (2011).

- ***April 1, 2010 through September 30, 2010***

  - *BER Network Requirement Workshop*

    Dean N. Williams, Gary Strand, and Galen Shipman, "Climate Science: Enabling Worldwide Access to Petascale Climate Data," *Proceedings of the Office of Biological and Environmental Research Network Requirements Workshop*, April 29–30, 2010. According to the report, the biggest BER user of ESnet in the future will likely be the climate community. Through ESG, LLNL and ORNL will host petabyte archives that will be accessible worldwide.

  - *SciDAC '10 Conference Proceedings*

    D. N. Williams, J. Ahrens, R. Ananthakrishnan, M. Balman, G. Bell, S. Bharathi, D. Brown, M. Chen, A. L. Chervenak, L. Cinquini, R. Drach, I. T. Foster, P. Fox, S. Hankin, D. Harper, N. Hook, P. Jones, D. E. Middleton, N. Miller, E. Nienhouse, R. Schweitzer, G. Shipman, A. Shoshani, F. Siebenlist, A. Sim, W. G. Strand, F. Wang, C. Ward, P. West, H. Wilcox, N. Wilhelmi, and S. Zednik, "Earth System Grid Center for Enabling Technologies: Building a Global Infrastructure for Climate Change Research," *Journal of Physics: Conference Series, SciDAC '10 Conference Proceedings*.

  - *Proceedings of the International Symposium on Grid Computing*

    Sim, D. Gunter, V. Natarajan, A. Shoshani, D. Williams, J. Long, J. Hick, J. Lee, and E. Dart, "Efficient Bulk Data Replication for the Earth System Grid," *Proceedings of the International Symposium on Grid Computing, Data Driven e-Science: Use Cases and Successful Applications of Distributed Computing Infrastructures* (ISGC 2010).

- ***October 1, 2009 through March 31, 2010***

  - *SC-09 High Performance GridFTP Transport of ESG Data*

    The Supercomputing 2009 (SC'09) Bandwidth Challenge entry titled, "High Performance GridFTP Transport of Earth System Grid (ESG) Data", demonstrated high performance GridFTP transport of climate data from multiple Department of Energy laboratory locations to the targeted SC showroom floor. The transferred multi-terabyte data consisted of a small portion of the multi-model Coupled Model Intercomparison Project, Phase 3 (CMIP-3) data set used in the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (AR4). SC'09 conference proceedings.

  - *Lessons learned from moving Earth System Grid data sets over a 20 Gbps wide-area network*

    Raj Kettimuthu, *Alex Sim*, Dan Gunter, Bill Allcock, Peer-Timo Bremer, John Bresnahan, Andrew Cherry, Lisa Childers, Eli Dart, *Ian Foster*, Kevin Harms, Jason Hick, Jason Lee, Michael Link, Jeff Long, Keith Miller, Vijaya Natarajan, Valerio Pascucci, Ken Raffenetti, David, Ressman, *Dean Williams*, Loren Wilson, and Linda Winkler, "Lessons learned from moving Earth System Grid data sets over a 20 Gbps wide-area network", http://esg-pcmdi.llnl.gov/publications_and_documents/HPDC10_BWC_Final.pdf/view.

  - *Experience Report on the Supercomputing 2009 Bandwidth Challenge from the Perspective of the Earth System Grid*

    Alex Sim and Dean N. Williams, "Experience Report on the Supercomputing 2009 Bandwidth Challenge from the Perspective of the Earth System Grid", Supercomputing Conference '09 Bandwidth Challenge, December 2009, http://esg-pcmdi.llnl.gov/publications_and_documents/ExpReports-ESG-SC09_final.pdf/view.

- ***April 1, 2009 through September 30, 2009***

  - *Grid Computing Environments Workshop Paper*

    F. Siebenlist, R. Ananthakrishnan, D. E. Bernholdt, L. Cinquini, I. T. Foster, D. E. Middleton, N. Miller, D. N. Williams, "Enhancing the Earth System Grid Authentication Infrastructure through Single Sign-On and Auto-provisioning," Grid Computing Environments Workshop 2009, Nov 2009 (submitted Sep 2009).

  - *TeraGrid Conference Paper*

    F. Siebenlist, R. Ananthakrishnan, D. E. Bernholdt, L. Cinquini, I. T. Foster, D. E. Middleton, N. Miller, D. N. Williams, "The Earth System Grid Authentication Infrastructure: Integrating Local Authentication, OpenID and PKI," 2009 TeraGrid Conference (TG09), June 2009.

  - *IEEE Computer Graphics and Application Paper*

    Potter, Kristin; Wilson, Andrew; Bremer, Peer-Timo; Williams, Dean; Doutriaux, Charles; Pascucci, Valerio, "Ultrascale Data Visualization using Ensemble-Vis," IEEE Computer Graphics and Applications, CG&ASI-2009-09-0118.

  - *Eos, Transactions, American Geophysical Union Paper*

    Charles Doutriaux, Robert Drach, Renata McCoy, Velimir Mlaker, and Dean Williams, "Climate Data Analysis Tools Facilitate Scientific Investigations", Eos, Transactions, American Geophysical Union, Volume 90, Number 35, September 1, 2009.

- *IEEE International Conference on Data Mining Paper*

  Dean Williams, Charles Doutriaux, Robert Drach, Renata McCoy, "The Flexible Climate Data Analysis Tools (CDAT) for Multi-model Climate Simulation Data", IEEE International Conference on Data Mining, December, 2009.

- *IEEE International Conference on Data Mining Paper*

  Chris Mattmann, Daniel Crichton, Amy Braverman, Dean Williams, Michael Gunson, David Woollard, Sean Kelly, and Michael Cayanan, "Climate Data eXchange (CDX) System", IEEE International Conference on Data Mining, December, 2009.

- **October 1, 2008 through March 31, 2009**

  - *Earth System Modelling (ESM) Software, Tools, and Environments Book*

    Robert Drach, Steve Hankin, Don Middleton, and Dean N. Williams, provided written input for chapter 5 (i.e., entitled, "*IO and Post-processing*") of the book titled, "*Earth System Modelling (ESM) Software, Tools, and Environments*". They authored the sub-sections: "*Data Representation*" and "*Data Analysis and Visualization*". More written input highlighting "*The Earth System Grid: Distributed and Uniform Access to ESM Data*" will be provide for chapter 7 (i.e., entitled, "ESM-*Data Archives in times of the GRID*").

  - *SciDAC Review Article*

    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. "*The Planet at Their Fingertips: Climate Modeling Data Heats Up*", Spring 2009.

    The ESG-CET team completed the SciDAC Review Article entitled, "The Planet at Their Fingertips: Climate Modeling Data Heats Up". The article talks about the increasing importance of climate modeling and the tremendous need for the Earth System Grid to allow fast and accurate access to hundreds of petabytes.

    (URL: http://www.scidacreview.org/0902/html/esg.html )

  - *Paper in the Bulletin of the American Meteorological Society (BAMS)*

    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, "*The Earth System Grid: Enabling Access to Multi-Model Climate Simulation Data*", in the Bulletin of the American Meteorological Society, February 2009.

    (URL: http://ams.allenpress.com/perlserv/?request=get-abstract&doi=10.1175/2008BAMS2459.1)

    This article, by the ESG-CET team, follows the Meehl et al., 2007 *BAMS* article on *"The WCRP CMIP3 Multi-Model Dataset: A New Era in Climate Change Research"*.

  - *Scientific Grand Challenges: Challenges in Climate Change Science and The Role of Computing at the Extreme Scale Report*

    Dean N. Williams, Don Middleton, et al. co-authored the section, "Data, Visualization, and Computing Productivity" in the Scientific Grand Challenges: Challenges in Climate Change Science and The Role of Computing at the Extreme Scale Report, November 6-7, 2008.

- *Developing Service-Oriented Applications in a Grid*

  Garcia, J., Fox, P., West, P. and Zednik, S. 2009, Developing Service-Oriented Applications in a Grid Environment. Experiences Using the OPeNDAP Back-End-Server Presentations, Earth Science Informatics, in press.

- **April 1, 2008 through September 30, 2008**

  - *Earth Science Informatics Article*

    Developing Service-Oriented Applications in a Grid Environment: Experiences Using the OPeNDAP Back-End-Server, 2008, J. Garcia, P. Fox, P. West. S. Zednik.

    Submitted to Earth Science Informatics as part of the special issue on "Grid Computing for Geosciences".

  - *Federation*

    Robert Schuler and Ann Chervenak, wrote a federation document for the ESG-CET team entitled, "Search Metadata: Storage and Sharing Considerations". In addition, they gave a talk on "Federated Metadata" at the ESG All-Hands Meeting in April 2008.

  - *SciDAC Review Highlights*

    The ESG-CET team prepared an article for the SciDAC Review highlighting ESG-CET. Several of the SciDAC projects that reported at the SciDAC 2008 conference were asked to present a one-page summary of their project, together with an image and a caption.

- **October 1, 2007 through March 31, 2008**

  - *Paper in the Cyberinfrastructure Technology Watch (CTWatch)*

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

  - *International Symposium on Collaborative Technologies and Systems*

    W. Kendall, M. Glatter, J. Huang, F. Hoffman, and D. E. Bernholdt, 2008: Web enabled collaborative climate visualization in the Earth System Grid. Invited paper for the 2008 International Symposium on Collaborative Technologies and Systems (CTS 2008), to appear.

  - *LLNL 2007 Computation Directorate Annual Report*

    A.A. Mirin, and D.N. Williams, 2008: Climate modeling – Conducting research and providing enabling technology. *LLNL 2007 Computation Directorate Annual Report*.

- **April 1, 2007 through September 30, 2007**

  - *Poster and Paper: SciDAC '07 Conference*

    The ESG team presented a peer-reviewed paper in the SciDAC 2007conference proceedings. The complete citation is: 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, K Halliday, S Hankin, P Jones, C Kesselman, D E Middleton, J Schwidder, R Schweitzer, R Schuler, A Shoshani, F Siebenlist, A Sim, W G Strand, N. Wilhelmi, M Su, and D N Williams, "Building a Global Federation System for Climate Change Research: The Earth System Grid Center for Enabling Technologies (ESG-CET)", in the Journal of Physics: Conference Series, SciDAC '07 conference proceedings.

- ***October 1, 2006 through March 31, 2007***

  - *Paper and Presentation: eScience'06*

    Ann Chervenak (ISI/USC) presented a peer-reviewed paper at eScience 2006 in Amsterdam, Netherlands. The complete citation is: Ann Chervenak, Jennifer M. Schopf, Laura Pearlman, Mei-Hui Su, Shishir Bharathi, Luca Cinquini, Mike D'Arcy, Neill Miller, and David Bernholdt, Monitoring the Earth System Grid with MDS4, in *Second IEEE International Conference on e-Science and Grid Computing (e-Science'06)*, page 69, Los Alamitos, CA, USA, 2006, IEEE Computer Society.

  - *Position Paper: DOE Cybersecurity Workshop*

    Representing ESG, NCAR, and the science/data community in general, Don Middleton (NCAR) presented a short position paper at the DOE-sponsored Cybersecurity Workshop, which was held in January 2007 at AGU headquarters in Washington, D.C.

## B.3 Presentations

- ***April 1, 2011 through September, 30, 2011***

  [Add latest posters]

- ***October 1, 2010 through March 31, 2011***

  - *DOE Network Workshop 2010*
    "*ESG and Cloud Computing with an Experience from Exploring Cloud for Parallelizing Tropical Storm Tracking,*" A. Sim, Expedition Workshop, Seeing Through the Clouds: Exploring Early Communities and Markets Streamlined by Open Government Principles, October 19, 2010.

  - *NARCCAP PI Meeting*

    Don Middleton attended the October 19, 2010, Fall NARCCAP PI meeting help at NCAR in Boulder Colorado, and provided a brief update on ESG status.

  - *Dynamic, Distributed, Data-Intensive Programming Abstractions and Systems*

    Don Middleton presented "Cyberinfrastructure and the Global Environmental Data Challenge" at this event, also known as "3DPAS," in February 2011 in Edinburgh, Scotland, UK. 3DPAS is a joint venture of the US National Science Foundation, the UK e-Science Institute, and UK government research programs.

  - *NASA-IPCC Data System Workshop*

    ESG team members presented multiple talks at the NASA-IPCC Data System Workshop held at NASA Goddard Space Flight Center on November 9–10, 2010. The purpose of this workshop was to discuss a NASA technical implementation plan to share observations for CMIP5 and create an effective partnering relationship between NASA and other federal agencies. This meeting followed a successful workshop at PCMDI on October 12–13, 2010, that identified potential observational data sets to be shared for the study.

    o Dean N. Williams, Luca Cinquini, Dan Crichton, and Galen Shipman presented "Partnership for 21st Century Earth System Science."

    o Luca Cinquini, Amy Braverman, Dan Crichton, Chris Mattmann, and Dean Williams presented "Requirements and Process for Publishing Observational Datasets into the global CMIP5 Archive."

    o Gary Strand presented "CESM, CMIP5, ESG, and Data Management."

    o   Galen M. Shipman, Sig Christensen, Luca Cinquini, John Harney, Misha Krassovski, Ross Miller, Giri Palanisamy, and Feiyi Wang presented "ESG – Observation Data Set Integration."

- *PCMDI – NASA Meeting*

    Dean N. Williams, Luca Cinquini, Gavin Bell, Daniel Crichton, Bob Drach represented ESG-CET at PCMDI—a NASA meeting sponsored by NASA and DOE. The meeting was held at LLNL in Livermore, California on October 12-13, 2010. As part of an ongoing effort to get NASA and NOAA observational data into the CMIP5 archive, this group is looking to include observational data sets to help score models when evaluating ensemble predictions and to improve the understanding of physical processes.

- *Greenhouse-Gas Information Systems*

    Dean N. Williams represented ESG at the Greenhouse-Gas Information Systems requirement meeting held at LLNL in early October 2010. Continued discussions are planned for the 2011 fiscal year regarding the potential role of ESG within the proposed mission operation center.

▪ ***April 1, 2010 through September 30, 2010***

- *National Climate Assessment*

    Dean N. Williams represented ESG-CET at the National Climate Assessment meeting sponsored by NSF, NOAA, and the U.S. Geological Survey (USGS), which was held at the USGS National Center Auditorium on September 21, 2010. As part of a more sustainable and ongoing effort, this group is looking to tie into the ESG infrastructure. A National Climate Assessment is required every four years, and the next report due in June 2013.

- *Fall 2010 Climate Visualization 10-05 Architecture Meeting*

    Bob Drach and Dean N. Williams represented ESG at the Fall 2010 Climate Visualization 10-05 Architecture Meeting. This meeting brings together technologists and scientists interested in leveraging emerging technologies and solutions to support, enable, and enhance climate research. Meeting attendees were particularly interested in the ESG work on infrastructures and tools to support analysis and sharing of climate model output and observations.

- *Greenhouse-GAS Information Systems*

    Dean N. Williams represented ESG at the Greenhouse-GAS Information Systems Requirement Meetings held at ORNL and Sandia National Laboratories in August and September. Further discussions are planned for FY11 regarding the potential role of ESG within the proposed mission operation center.

- *2010 Federation of Earth Science Information Partner (ESIP) Meeting*

    Luca Cinquini, Dan Crichton, Cecelia DeLuca, Dean Williams presented, "The Earth System Grid Federation" at the 2010 ESIP Federation Meeting held at the University of Tennessee's Conference Center in Knoxville, TN July 20-23. The Federation of Earth Science Information Partners is a unique consortium of 121 organizations that collect, interpret and develop applications for remotely sensed Earth observation information. Luca represented the ESG team at the meeting.

- *Climate Science for a Sustainable Energy Future*

    Dean N. Williams and Galen Shipman represented ESG at the Climate Science for a Sustainable Energy Future (CSSEF) Project meeting held at ORNL on May 24–26. Their talk, "Federation and Integration of Data from Multiple Sources," covered critical data functionalities, interagency partnerships, and requirements for CSSEF.

- *DOE ASCR–BER Network Requirements Workshop*

  Dean N. Williams, Galen Shipman, and Gary Strand discussed OBER-related climate network needs at the ASCR–BER Network Requirements Workshop. Dean presented "Enabling Worldwide Access to Petascale Climate Data" and wrote a 16-page paper based on this talk for the workshop report. The workshop report discusses outstanding network issues, future needs, and scientific directions for the climate community. The planned network will have ESG as the cornerstone for distributing climate data to the community. In addition, the report notes that the climate research community is likely to be the biggest BER user of ESnet in the future. Climate sites (e.g., LLNL, ORNL) will host huge data sets that will be accessed by users worldwide. By 2020, the climate data repositories at these sites may exceed hundreds of exabytes in size. Several key sites for distributing these data are the CMIP5 Data Centers at LLNL, BADC, DKRZ, the Australian National University , and the Japan Agency for Marine-Earth Science and Technology. NCAR and LBNL will distribute data from the CCSM model, and ORNL will distribute observational and other data. Data will be replicated across these and other locations, and bandwidth requirements will likely reach 100 Gbps in the next few years.

- *DOE SciDAC UltraViz Institute All-Hands Meeting*

  Don Middleton gave an invited talk at the 2010 DOE SciDAC UltraViz Institute All-Hands Meeting, held at the University of California at Davis in April 2010. His talk, "Challenges in Climate Data Analysis and Visualization," included ESG-CET as a primary theme.

- *DOE BER Climate Change Modeling Program Science Team Meeting*

  Dean N. Williams presented "The Earth System Grid Center for Enabling Technologies (ESG-CET): Serving Climate Data to the World" at the DOE Climate Change Modeling Program Science Team Meeting in Gaithersburg, Maryland, March 29–April 2, 2010.

- *HPDC Challenges for Large Applications in Distributed Environments*

  Don Middleton presented "The Earth System Grid Federation: A Globally Distributed Environment for Climate Research," an invited talk, at the 2010 Conference for the High Performance Distributed Computing, Challenges of Large Applications in Distributed Environments, which was held in Chicago, Illinois.

- *ESG Federation Meeting*

  Eric Nienhouse presented "Enabling Community Driven Development in the ESG Gateway," at the international ESG Federation meeting held in Princeton, New Jersey.

  Gavin Bell presented "Peer-2-Peer Architecture" at the international ESG Federation meeting held in Princeton, New Jersey.

  Bob Drach presented "CMIP5 (IPCC AR5) System Requirements" ant the international ESG Federation meeting held in Princeton, New Jersey.

  Luca Cinquini presented "Solr and Federation Architecture" and the international ESG Federation meeting held in Princeton, New Jersey.

- ***October 1, 2009 through March 31, 2010***

  - *DOE BER Climate Change Modeling Program Science Team Meeting*

    Dean N. Williams presented, "The Earth System Grid Center For Enabling Technologies (ESG-CET): Serving Climate Data to the World" at the DOE Climate Change Modeling Program Science Team Meeting in Gaithersburg, MD, March 2010. The conference was held in Gaithersburg, MD, March 29–April 2, 2010.

- *NOAA Climate Observations and Models Workshop*

  Dean N. Williams participated in the presentation titled, "Observation for CMIP-5 Simulations" given at the Climate Observation and Models Workshop in Asheville, NC, March 24, 2010.

- *High Performance Computing in Weather and Climate Conference*

  Wang Feiyi presented, "Petascale Data Management in Support of DOE and NSF Communities" at the High Performance Computing in Weather and Climate in Beijing, China March 22, 2010.

- *WMO Information Systems Meeting, Inter-Asia Workshop on WIS*

  The Japanese Meteorological Agency (JMA) hosted an Inter-Asia and Middle East workshop on building the WMO WIS system March 9-12, 2010 in Tokyo, Japan. As an invited lecturer, Don Middleton presented an overview of NCAR's participation in WMO WIS, which included the federation of ESG-CET capabilities.

- *GlobusWorld 2010*

  Rachana Ananthakrishnan presented, "Earth System Grid: Community Update", at the GlobusWorld 2010 conference held at the Argonne Illinois Conference Center, March 2–4, 2010.

  Frank Siebenlist presented, "Earth System Grid Data Access Security", at the GlobusWorld 2010 conference held at the Argonne Illinois Conference Center, March 2–4, 2010.

- *Ocean Sciences Meeting*

  Steve Hankin et al. presented NOAA efforts to compliment and leverage the work of the ESG in a talk titled, "The Unified Access Framework for Gridded Data" at the 2010 Ocean Sciences Meeting in Portland, Oregon, February 22–26, 2010 (http://www.agu.org/meetings/os10/).

  Jeremy Malczyk and Steve Hankin presented ESG software components in a talk titled, "Using Software Version Control Tools for Data Management and Quality Control Collaboration in the Surface Ocean CO2 Atlas (SOCAT) Project" at the 2010 Ocean Sciences Meeting in Portland, Oregon, February 22–26, 2010 (http://tinyurl.com/SOCAT-AGU-OS-2010).

- *WMO Information Systems Meeting, Intercommission Coordination Group (WMO ICG-WIS)*

  Don Middleton represented NCAR and ESG at the WMO ICG-WIS meeting in Seoul, Korea February 22-26, 2010. His presentation was entitled "NCAR DCPC Update: With a Focus on CMIP5/IPCC, and on Behalf of an International Collaboration".

- *Oak Ridge National Laboratory Meetings*

  Galen Shipman presented, "The Earth System Grid at ORNL" to representative of the National Climatic Data Center, October 21, 2009.

  Galen Shipman presented, "The Earth System Grid" to representatives of the University of Idaho, February 1, 2010.

  Galen Shipman presented, "Facilitating a Data Intensive Computational Environment at the OLCF" to representatives of the Exxon Corporation, February 17, 2010.

- *LBNL Magellan and Advanced Network Initiative Meeting*

  Dean N. Williams and Alex Sim presented, "The Earth System Grid Center for Enabling Technologies: Network and Analysis Requirements" to the leads of ESnet and LBNL Magellan projects. The meeting took place at LBNL in Berkeley, California, November 10, 2009.

- *NOAA Global Interoperability Program Kickoff Meeting*

  Luca Cinquini, Steve Hankin, Don Middleton, and Dean N. Williams presented the "Earth System Grid Center for Enabling Technologies" at the NOAA funded Global Interoperability Program (GIP) kickoff meeting held at the Geophysical Fluid Dynamics Laboratory in Princeton, NJ on November 5–6, 2009.

- *2009 Global Organization for Earth System Science Portal (GO-ESSP) Workshop*

  At the GO-ESSP workshop, the following presentations were presented by ESG-CET team members: Luca Cinquini – "An Introduction to the ESG-CET CMIP-5 Federation", Dean N. Williams and Bob Drach – "ESG Data Node Configuration", Dean N. Williams and Karl Taylor – "CMIP-5 Overview", and "Climate Model Output Rewriter", Steve Hankin – "Report from OceanObs09: Opportunities and challenges for the emerging framework of netCDF-CF-THREDDS-DAP" and "Remote Processing Issues", Roland Schweitzer – "Model Intercomparison Using the LAS Interactive Earth Science Data Visualization Galary (vizGal)", and Gary Strand – "Distributing Climate Model Data to the World: A 10 year retrospective". The workshop was held October 6–8 at the Institute for Pharmacy in Hamburg, Germany.

- ### April 1, 2009 through September 30, 2009

  - *Dr. Steve Koonin's Visit to LLNL*

    Dean Williams briefed Dr. Steve Koonin, the Under Secretary for Science at the Department of Energy, on enabling the broader climate science community to analyze and evaluate models. This presentation included ESG-CET and its significances in cataloging and disseminating petascale data to the climate community.

  - *Dr. Steve Koonin's Visit to ANL*

    Ian Foster briefed Dr. Steve Koonin, the Under Secretary for Science at the Department of Energy, on the challenges and opportunities inherent in understanding regional impacts of, and human responses to, climate change. This presentation included ESG-CET and its importance for the climate community.

  - *Space Mission Challenges for Information Technology*

    Dean Williams presented, "The Earth System Grid Center for Enabling Technologies" at the 3rd IEEE International Conference on Space Mission Challenges for Information Technology. The conference was held in Pasadena, California, July 19 – 23, 2009.

  - *TeraGrid 2009*

    Frank Siebenlist presented, "The Earth System Grid Authentication Infrastructure: Integrating Local Authentication, OpenID and PKI" at the TeraGrid 2009 Tech Track, Session VI: Gateways and Gateway Authentication, in Arlington, Virginia, June 23, 2009.

  - *OceanObs 2009*

    Steve Hankin et al., "NetCDF-CF-OPeNDAP: Standards for Ocean Data Interoperability and Object Lessons for Community Data Standards Processes", Ocean Data Standards Processes, Feb 19, 2009, V1.1, presented at OceanObs'09, Venice, Italy, Sept, 25, 2009.

  - *Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology*

    Schweitzer, Roland and K. M. O'Brien, J. Li, A. Manke, J. Malczyk, and S. Hankin, "A General Purpose System for Server-side Analysis of Earth Science Data", 25th Conference on International Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology, Phoenix, AZ, Wednesday, 14 January 2009.

▪ *October 1, 2008 through March 31, 2009*

- *Collaboratories Workshop*

  Dean N. Williams et al., presented "Composing Collaboratories: Earth System Grid Center for Enabling Technologies (ESG-CET)" at the Collaboratories Workshop held in Chicago, IL, February 24-25, 2009.

- *Osni Marques Visit to LLNL*

  Dean N. Williams presented "The Earth System Grid Center for Enabling Technologies at LLNL" to Osni Marques, ASCR's Computer Science Program Manager for SciDAC. He is working with Dan Hitchcock on all aspects of the Computer Science Program. The talk was given on January 15, 2009.

- *Conference on International Interactive Information and Processing Systems (IIPS)*

  Schweitzer, Roland and K. M. O'Brien, J. Li, A. Manke, J. Malczyk, and S. Hankin, "A General Purpose System for Server-side Analysis of Earth Science Data", 25th Conference on International Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology Phoenix, AZ, Wednesday, 14 January 2009.

- *MAGIC Teleconference*

  Don Middleton and Dean N. Williams presented the Earth System Grid at the MAGIC Meeting Teleconference. The meeting was held on December 3, 2008.

- *Supercomputing 2008*

  Dean N. Williams gave a presentation on "Climate Change: Current Knowledge and Future Challenges" at the 2008 Supercomputing Conference for High Performance Computing, Network, Storage and Analysis, held in Austin, TX, November 15-21, 2008. Dean presented efforts to build up and disseminate climate data via the Earth System Grid.

- *Extreme Scale Computing Challenges in Climate Change Science Workshop*

  Don Middleton and Dean N. Williams co-chaired the Data, Visualization, and Computing Productivity session at the Extreme Scale Computing Challenges in Climate Change Science Workshop. Other responsibilities at the workshop included: organization, hosting breakout sessions, presentations, and providing written input for the final workshop report. The workshop was held in Washington D.C., November 6-7, 2008.

- *Presentation at GO-ESSP Community Meeting*

  Schweitzer, Roland and K. M. O'Brien, A. Manke, J. Malczyk, and S. Hankin, "*A General Purpose System for Server-side Analysis of Earth Science Dat*a", 6th GO-ESSP Community Meeting, Seattle WA, 19, September 2008.

- *Presentation at EGU 2009*

  The Climate-G testbed: towards a large scale data sharing environment for climate change by G. Aloisio et al.

▪ *April 1, 2008 through September 30, 2008*

- *Net@Edu Meeting*

  Don Middleton provided the keynote address at the annual Net@Edu meeting in Tempe, Arizona in February 2008 (this was not reported in the previous ESG-CET report for that time period). Entitled *"Cyberinfrastructure and Emerging Scientific Data and Knowledge Systems"*, Middleton's presentation highlighted ESG-CET as a primary focal point of the talk.

- *North American Regional Climate Change Assessment Program (NARCCAP) PI and Community Meeting*

  NARCCAP is a multi-agency (NSF, DOE, NOAA, EPA) regional climate modeling program that is closely affiliated with ESG-CET, and providing data services to the community using ESG cyberinfrastructure. Don Middleton contributed to the organization of this meeting and presented an opening talk entitled *"NARCCAP Data: The Earth System Grid (ESG), Data & Knowledge Systems, and a Few Useful Tools".* We also held a hand-on training session for community members on using the ESG-based NARCCAP data system, and it was well received and effective.

- *Virtual Climate Change Research Institute Workshop*

  Ann Chervenak gave a presentation on "The Earth System Grid: Turning Climate Datasets into Community Resources" at the Virtual Climate Change Research Institute Workshop held at USC on June 25, 2008. A group of paleoclimatologists is proposing a new NSF Science and Technology Center and hopes to collaborate with ESG-CET.

- *World Meteorological Organization (WMO) ET-WISC Meeting*

  In June 2008, Don Middleton attended a meeting in Darmstadt, Germany of the Expert Team on the World Meteorological Organization (WMO, a program of the United Nations) Information System (WIS) Global Federated Systems (ET-WISC). Middleton provided a presentation on federated scientific data systems included a briefing on ESG-CET for the WMO expert team.

- *Washington trip to meet with Sponsors*

  On June 27th, Dean Williams spent a day in Washington, D.C. with sponsors from the DOE Office of Science SciDAC and from the Office of Biological and Environmental Research (OBER). Dean presented information on ESG-CET and discussed the progress of this project. The meeting also provided an opportunity to introduce Dean and ESG-CET to Susan Turnbull, the new ESG-CET SciDAC Program Manager.

- *ESnet Climate meeting at LBNL with Dan Hitchcook*

  David Bernholdt contributed to a presentation by John Drake (lead PI for the SciDAC CCSM effort) discussing data volumes and networking needs anticipated for CMIP5 (IPCC AR5), presented at the BER-ESnet workshop on 4 August. Dean Williams was also present at this meeting and presented information on ESG-CET.

- *BER Program Manager Wanda Farrell's visit to ORNL*

  On August 6th, BER Program Manager Wanda Farrell visited ORNL. David Bernholdt contributed to a presentation by John Drake on modeling-related activities at ORNL, including ESG-CET.

- *Presented ESG-CET to Christine Chalk*

  On August 6th, BER Program Manger Christine Chalk visited LLNL. Dean Williams presented ESG-CET and discussions took place covering the use of ESG-CET for SciDAC's Biology & Environment scientific area.

- *Presented ESG-CET to David Skinner and Horst Simon*

  On August 13th, Dean Williams visited NERSC to discuss the ESG-CET and the possible need for more storage space on NERSC's long-term storage facility. In addition, future conversations were planned on NERSC as an ESG "Data Node" and as a compute server.

- *BER Program Manager Anjuli Bamzai visit to ORNL*

  On September 11th, BER Program Manager Anjuli Bamzai visited ORNL, where David Bernholdt gave a presentation on ESG.

- *Global Organization for Earth System Science Portal (GO-ESSP) Workshop*

  At the GO-ESSP workshop, the following presentations were presented by ESG-CET team members: Luca Cinquini - "The ESG-CET Gateway: A Distributed and Federated Architecture for Data Search, Access and Analysis", Nate Wilhelmi - "ESG-CET Demonstration", Eric Nienhouse - "The Challenges of Building Rich Web Clients Geo-Spatial Applications – Lessons Learned", and Roland Schweitzer - "Server-side OPeNDAP Analysis – Concrete steps toward a generalized framework via a reference implementation using F-TDS".

- *Presentation at EGU 2008, IN09, The Grid for Geosciences*

  Grid-enabled OPeNDAP Hyrax servers: parallel/multiple back-end services and remote netCDF invocation (RNI), by J. Garcia, S. Zednik, P. West and P. Fox.

- ### *April 1, 2007 through September 30, 2007*

  - *Presentation: Co-Chair of the IPCC WG1*

    Dean Williams and Robert Drach demonstrated ESG-CET to Dr. Susan Solomn prior to her April 2007 LLNL "Director Distinguished Lecturer" series presentation on the scientific findings of the IPCC Working Group I (WG1), which were recently published in its fourth comprehensive assessment report (AR4). Dr. Solomon is a senior scientist at the Aeronomy Laboratory (a National Oceanic and Atmospheric Administration facility) and has served as co-chair of the IPCC Working Group I (WG1).

  - *Presentation: Fusion Energy Science Community -- Dr. William Tang*

    Dean Williams (LLNL) gave a presentation on ESG-CET to Dr. William Tang, the Chief Scientist at the Princeton Plasma Physics Laboratory (PPPL), a national laboratory for fusion research. Dr. Tang played a prominent leadership role for the Department of Energy's development multi-disciplinary program in advanced computational science, (i.e., the Scientific Discovery through Advanced Computing (SciDAC)). We discussed ways in which ESG-CET might be used to assist the DOE's Fusion Energy science community. This collaboration also involves the use of LLNL's computing resources, such as the Green Data Oasis and the Green Linux Capacity Cluster (GLCC).

  - *PCMDI Program Review:*

    Dean Williams presented a PowerPoint presentation on ESG-CET, subtitled: "Data and Software: Turning Climate Datasets into Community Resources" to the PCMDI Program Review Committee on August 27, 2007 in Livermore, CA.

  - *Poster and Presentation: Climate Change Prediction Program (CCPP) '07 Conference*

    Representing ESG, Dave Bernholdt and Dean Williams presented the ESG-CET poster at the September 2007 Climate Change Prediction Program (CCPP) conference, which was held in Indianapolis, Indiana. The poster was entitled: "Building a Global Infrastructure for Climate Change Research". Dean also presented a PowerPoint presentation on ESG-CET, entitled: "Data and Software Infrastructure for the Global Climate Community".

  - *Presentation: World Meteorological Organization Information System (WMO-WIS) Intercommission Coordination Group*

    The World Meteorological Organization (WMO) is in the process of designing and building its next generation global information system, an effort known as WMO-WIS. While WMO has long had an operational network for meteorological observations and warnings, the new system is to provide data management and access across the various WMO directorates, thus encompassing weather, climate, oceans, and more. Don Middleton serves on the Expert Team chartered with architecting and designing the federation of national and international systems and also serves as an advisor for the high-level

Intercommission Coordination Group (ICG-WIS). Middleton gave a presentation at the group's recent September meeting in Reading, U.K. that included an update on ESG-CET, and outlined opportunities for collaboration and idea exchange in the areas of metadata, federation, and virtual organizations.

- ### *October 1, 2006 through March 31, 2007*

  - *Presentation: VACET All-hands Meeting*

    Bob Drach (LLNL) gave an ESG-CET presentation at the VACET All-hands meeting held at the University of Utah in Salt Lake City, Utah.

  - *Presentation: Boulder Data Group*

    Don Middleton (NCAR) presented a talk entitled, "The Earth System Grid: Delivering Petascale Climate Data to a Global Community" at the inaugural meeting of the new Boulder Data Group, held in Boulder, CO.

  - *Presentation: SC2006*

    ESG-CET activities were featured at booths sponsored by ANL, Global Grid Forum, LBNL, NCAR, and ORNL in the Exhibit Hall of SC2006, Tampa Bay, FL.

  - *Presentation: DOE LLNL Program Review*

    Dean Williams (LLNL) presented ESG-CET at the DOE PCMDI Program Review held in November 2007 at LLNL, CA.

  - *Presentation: Curator Workshop*

    *Dean Williams (LLNL) presented ESG-CET information at the Curator Workshop on Metadata held at GFDL in Princeton, NJ.*

## B.4  Posters

- ### *April 1, 2011 through September, 30, 2011*
  [Add latest posters]

- ### *October 1, 2010 through March 31, 2011*

  - No posters presented for the reporting period.

- ### *April 1, 2010 through September 30, 2010*

  - DOE SciDAC Annual PI Meeting

    Don Middleton, Dean N. Williams, and Ian Foster presented the poster, "The Earth System Grid Center for Enabling Technologies: A Petascale Data Service for the Global Community," at the annual SciDAC PI Meeting, held in Chattanooga, Tennessee, in July 2010.

- ### *October 1, 2009 through March 31, 2010*

  - DOE BER Climate Change Modeling Program Science Team Meeting

    Dean N. Williams and Galen Shipman presented a poster titled, "The Earth System Grid Center for Enabling Technologies (ESG-CET): Serving Climate Data to the World" at the DOE Climate Change Modeling Program Science Team Meeting in Gaithersburg, MD, March 2010.

  - SC'09 Conference

Brady, E., Chervenak, A. L., "Mirroring Earth System Grid Data Sets", Supercomputing (SC09) Conference, ACM Student Poster Session, Portland, Oregon, November 2009.

▪ ***April 1, 2009 through September 30, 2009***

- Supercomputing 2009 (SC09)

  Ann Chervenak and Erin Brady will present their poster titled, "Mirroring Earth System Grid Data Sets" at the SC09 International Conference for High Performance Computing, Networking, Storage and Analysis in Portland, Oregon, November 2009.

▪ ***October 1, 2008 through March 31, 2009***

- *Climate Change Prediction Program (CCPP) Meeting*

  Dean N. Williams, et al., presented the poster titled, "The Earth System Grid: Scientific Discovery Made Easy", at the 2009 CCPP meeting in Washington, D.C., 6-9 April 2009.

- *American Geophysical Union (AGU) Fall Meeting*

  Dean N. Williams, et al., presented the poster titled, "The Earth System Grid: Scientific Discovery Made Easy" at the AGU Fall Meeting in San Francisco, 15-19 December 2008.

- *Global Ocean Data Assimilation Experiment Symposium*

  Hankin Steven, et al., "Climate and Forecast (CF) Conventions for NetCDF – The Foundation of GODAE Data Interoperability", (poster) Global Ocean Data Assimilation Experiment final symposium -- Observing and Forecasting the Ocean, Nice, France, 10-15 November 2008.

▪ ***April 1, 2008 through September 30, 2008***

- *SciDAC '08 PI Meeting in Denver, CO*
  David Bernholdt, Ian Foster, Don Middleton, and Dean Williams (the ESG-CET executive team) attended the SciDAC PI meeting held in Denver, CO, where Don and Dean presented a poster on ESG-CET. Besides providing a great opportunity for the executive to intermingle with other Advanced Scientific Computing Research (ASCR) computer science efforts, this meeting provided an opportunity to discuss topics in areas that may assist future ESG development, including performance tools, application development frameworks, data management/analytics, and visualization.

- *European Geophysical Union Meeting*

  In April of 2008, Don Middleton attended the annual European Geophysical Union (EGU) meeting in Vienna, Austria. Middleton presented a poster on ESG-CET, which was developed in concert with the ESG-CET Executive team.

▪ ***April 1, 2007 through September 30, 2007***

- *Poster and Paper: SciDAC '07 Conference*

  Don Middleton presented a poster on ESG-CET at the SciDAC '07 conference held in Boston, MA. Also representing ESG at the conference were Ian Foster, Dave Bernholdt, and Dean Williams. (Taking advantage of the conference, The ESG executive committee held many face-to-face meetings.)

- ***October 1, 2006 through March 31, 2007***

  - *Poster: SciDAC'07 PI Workshop*

    Dean Williams (LLNL) and Don Middleton (NCAR) presented a poster on ESG-CET at the SciDAC '07 Workshop held in Atlanta, GA.

  - *Poster: 3ʳᵈ WGNE Workshop on Systematic Errors in Climate and NWP Models*

  - *Dean Williams (LLNL) presented a poster on ESG-CET at the 3ʳᵈ WGNE Workshop on Systematic Errors in Climate and NWP Models held in February 2007 in San Francisco, CA.*

## B.5  Books

- ***ESM-Software book with ESGF contributions to the GRID Chapter 7***

  The Williams et al. GRID chapter is special in that it is covering a very fluid subject, compared to many of the other chapters. First year college students are the intended audience of the book and it is expected to come out in early 2012. The publisher Springer will be offering smaller volumes of the book where each chapter would become its own mini-book. Please visit the following URL for more details: http://www.springer.com/authors/book+authors/springerbriefs.

- ***Data Intensive Science***

  Williams et al. contributed a chapter titled "*Earth System Grid Federation: Infrastructure to Support Climate Science Analysis as an International Collaboration*" to the forthcoming book, *"Data Intensive Science"*. The book is to be published by Chapman & Hall. We focused our chapter on ESGF and the data-driven activities in extreme-scale climate science. The intended audiences for this book are college level juniors and seniors. The book is scheduled to be out in late 2012.