Final Report

PMEL contributions to the collaboration:

SCALING THE EARTH SYSTEM GRID TO PETASCALE DATA

for the DOE SciDAC's Earth System Grid Center for Enabling Technologies

October 1, 2006 through September 30, 2011

PMEL Principal Investigator
Steve Hankin

NOAA / Pacific Marine Environmental Laboratory
7600 Sand Point Way NE
Seattle, WA, 98125

Collaborating principal investigators
Dean N. Williams, Ian T. Foster, and Don E. Middleton

with major contributions from

the Earth System Grid Center for Enabling Technologies Team

contract number: DE-AI02-06ER25771

1 Abstract

Drawing to a close after five years of funding from DOE's ASCR and BER program offices, the SciDAC-2 project called the Earth System Grid (ESG) Center for Enabling Technologies has successfully established a new capability for serving data from distributed centers. The system enables users to access, analyze, and visualize data using a globally federated collection of networks, computers and software. The ESG software—now known as the Earth System Grid Federation (ESGF)—has attracted a broad developer base and has been widely adopted so that it is now being utilized in serving the most comprehensive multi-model climate data sets in the world. The system is used to support international climate model intercomparison activities as well as high profile U.S. DOE, NOAA, NASA, and NSF projects. It currently provides more than 25,000 users access to more than half a petabyte of climate data (from models and from observations) and has enabled over a 1,000 scientific publications.

Contents

1 Abstract ii

2 Executive Summary 1

3 Overview 2

3.1 Problem Statement 2

3.2 Summary of Project History 3

3.3 Project Success and Accomplishments 5

3.4 Balanced Research Ecosystem: Overall Architecture Design 7
3.4.1 Peer-To-Peer Architecture 8
3.4.2 Gateway Architecture 9
3.4.3 Ecosystem: Both Systems Co-existing 9

4 PMEL Contributions to the ESG-CET 11

4.1 Product Services 11
4.1.1 Motivation 11
4.1.2 Data Preview 11
4.1.3 Comparisons 12
4.1.4 Server-side Analysis and Data Reduction 13
4.1.5 Federation 13
4.1.6 Data Aggregation 13
4.1.7 Analysis with Domain Specific Tools 13

5 Data and Overall Community Impact 14

6 Collaborations 16

7 References 17

Appendix A PMEL Outreach, Papers, Presentations, Poster, Books 18

A.1 NOAA-related outreach Activities 18
- October 1, 2010 through March 31, 2011 18
- April 1, 2010 through September 30, 2010 18
- October 1, 2009 through March 31, 2010 19
- April 1, 2009 through September 30, 2009 20
- April 1, 2008 through September 30, 2008 21
- October 1, 2007 through March 31, 2008 22
- April 1, 2007 through September 30, 2007 23
- October 1, 2006 through March 31, 2007 24

A.2 Papers with PMEL authorship 26
- April 1, 2011 through September, 30, 2011 26
- October 1, 2010 through March 31, 2011 26
- April 1, 2010 through September 30, 2010 27
- October 1, 2008 through March 31, 2009 27
- April 1, 2007 through September 30, 2007 27

A.3 Books with PMEL contributions 27
- ESM-Software book with ESGF contributions to the GRID Chapter 7 27
- Data Intensive Science 27

2 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 Data Access 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 is of course if funding continues at some level. This ongoing effort, though daunting in scope and complexity, would 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.

3 Overview

3.1 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 petabytes), 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.

3.2 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.*

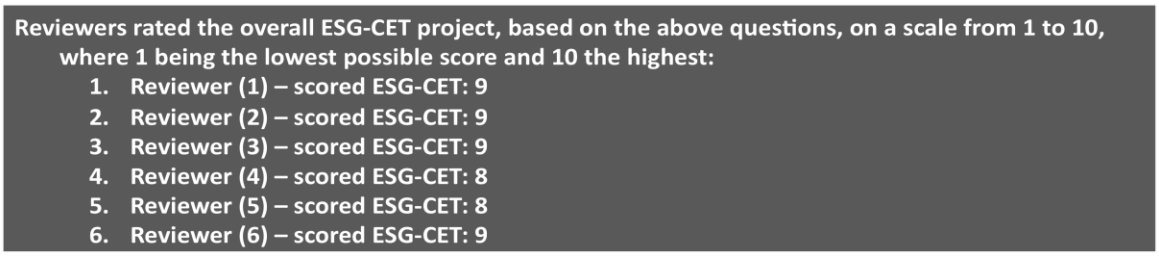


Figure 2: *Reviews 2009 mid-term score for the ESG-CET project.*

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.

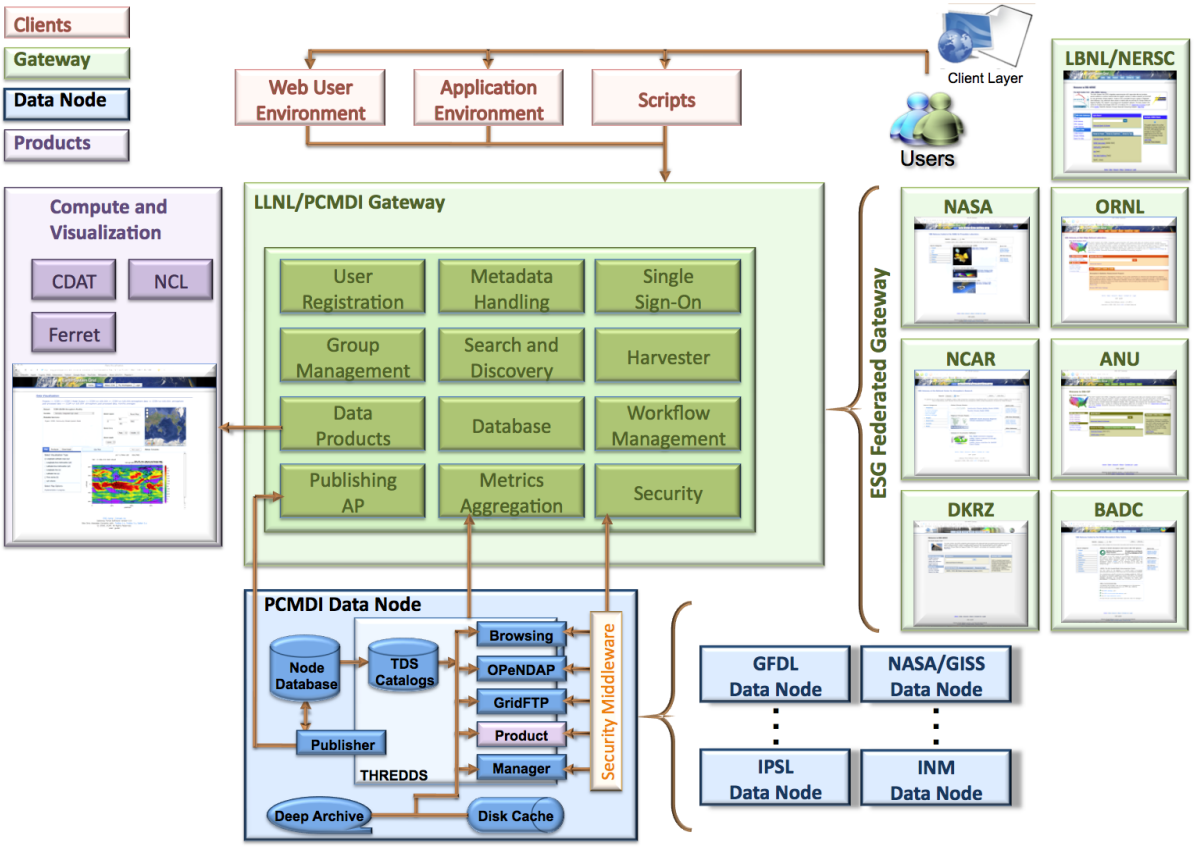


Figure 3*: Users can access ESGF data using Web browsers, scripts, and client applications. ESGF is separated into gateways (green) and data nodes (blue). Gateways handle user registration and management and allow users to search, discover, and request data. Data nodes are located where the data resides, 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.*

3.3 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. Currently serving over 25, 000 users, the total ESGF federated archive reaches upwards of 2 petabytes, most of which is visible on Table 7.

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 3.4). 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 and a more comprehensive collection of model output. For an up-to-the-minute overview of the data availability via ESGF, visit http://cmip-pcmdi.llnl.gov/cmip5/esg_tables/transpose_esg_static_table.html, which shows the IPCC model runs for designated experiments. For 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 340 TB. Visit the CMIP5 archive status website for an up-to-date summary of the total size, number of data sets, models, and participating modeling centers.

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/	1.3.2	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/	1.3.1	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	1.3.2	Production	CMIP5	drach1@llnl.gov
NCAR	http://www.earthsystemgrid.org	1.3.2	Production	CSSM, NARCCAP, PCM	esg-support@earthsystemgrid.org
NCI	http://esg.nci.org.au/esgcet	1.3.2	Production	CMIP5	muhammad.atif@anu.edu.au
ORNL	http://esg2-gw.ccs.ornl.gov/	1.3.2	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 report highlights the overall success of the project in meeting its deliverables.

3.4 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).

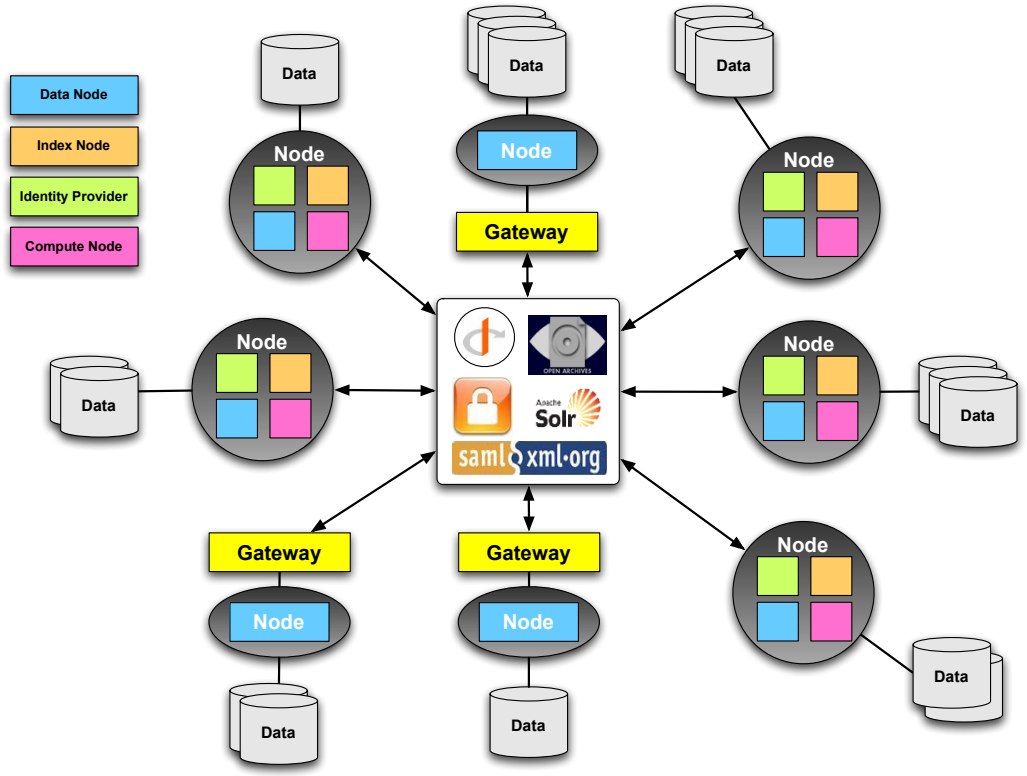


Figure 4: *The Earth System Grid Federation combines the new P2P architecture with the more traditional client-server model. The P2P system was expressly designed for extensibility and scalability, and it supports geospatial and temporal search, dashboard showing system metrics, user interface for notification, 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.*

3.4.1 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 DAP, 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's Data Access Protocol.

3.4.2 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.

3.4.3 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 Data Access 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.

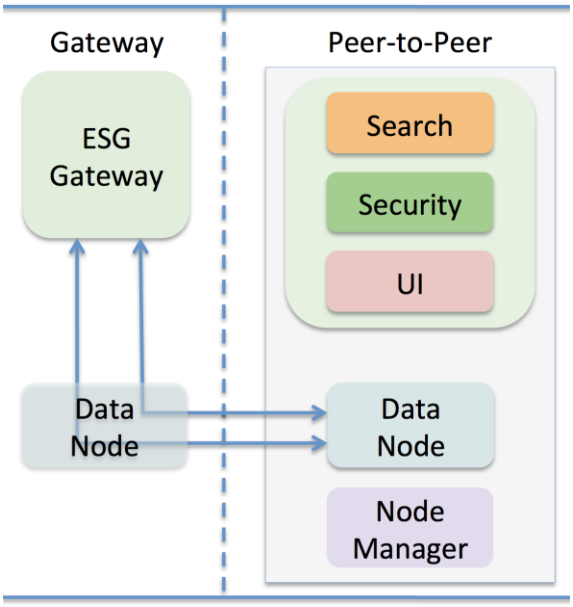


Figure 5: *Ecosystem: co-exiting stable systems*

4 Innovative Technology and Integration

4.1 Product Services

4.1.1 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.

4.1.2 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 (as shown in Figures 29 and 30).

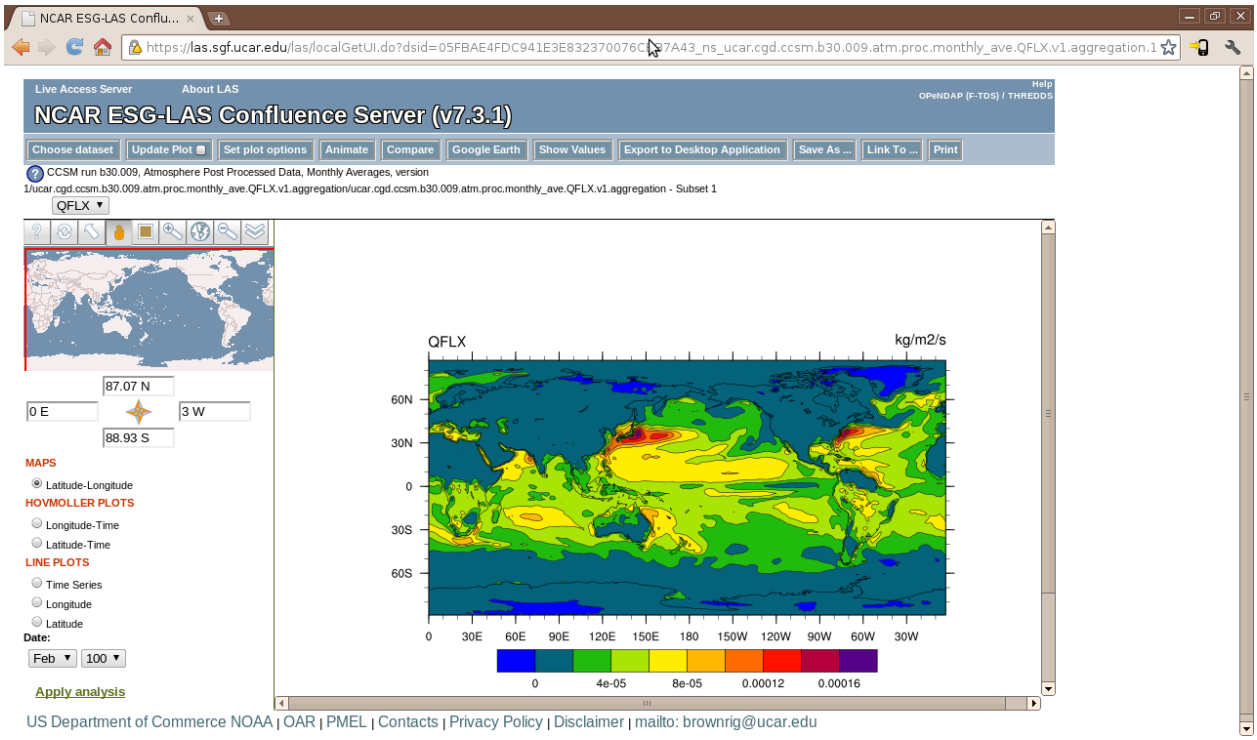


Figure 29: *Data preview using NCAR's NCL as the visualization engine for the product services*

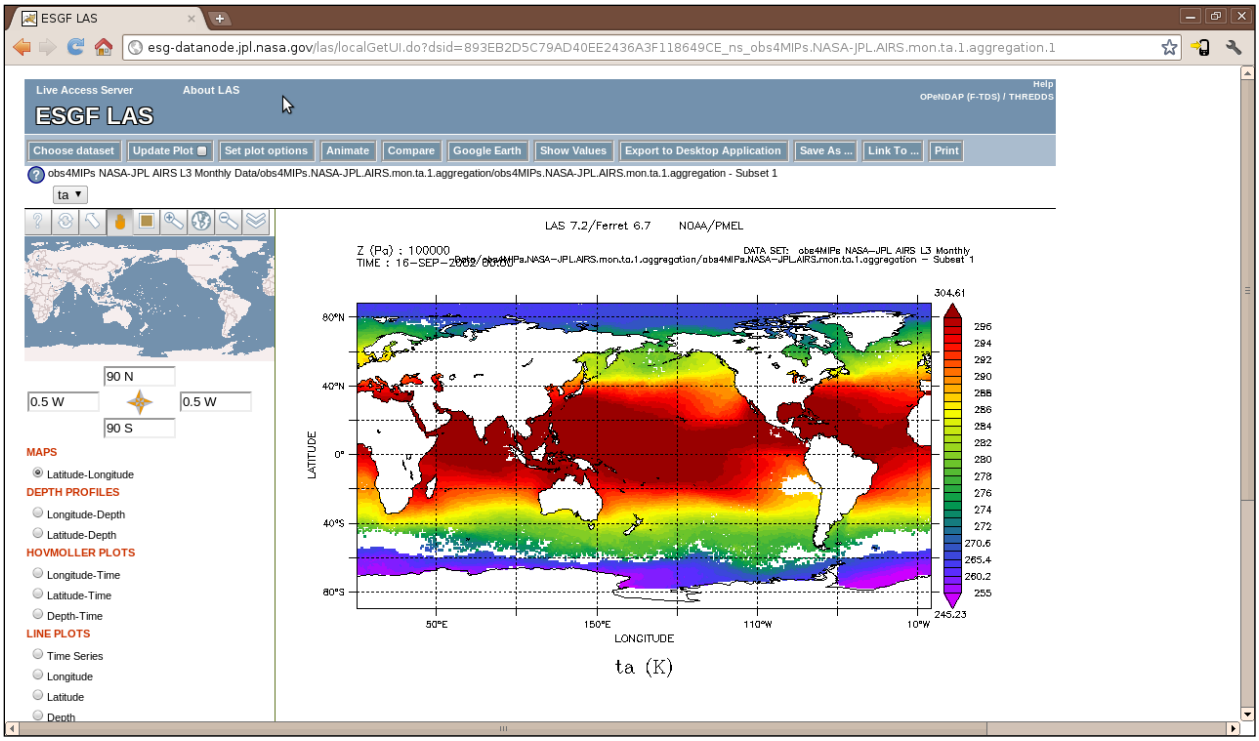


Figure 30: *Data preview using PMEL's Ferret as the visualization engine for product services from the NASA/ JPL data node.*

4.1.3 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) shown in Figure 31. 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 him or her to re-enter the ESGF product services at that precise point and pursue further investigation of the data sets.

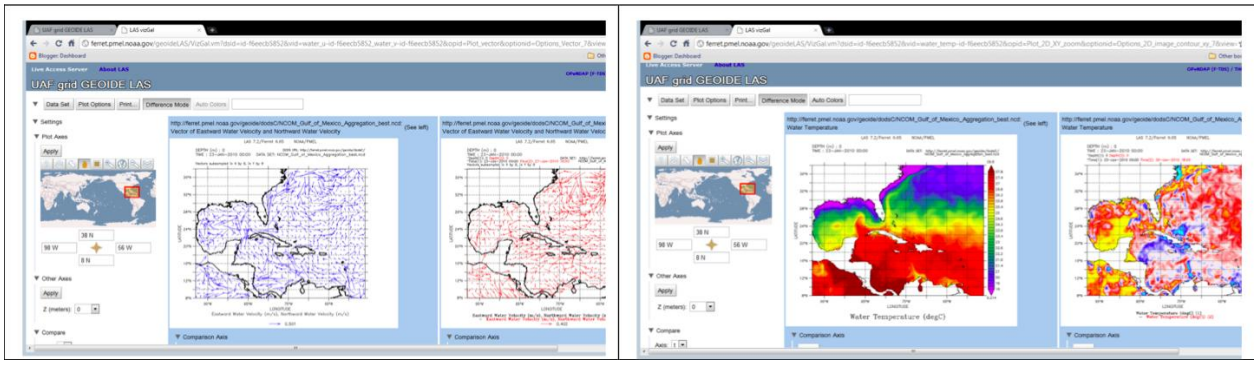


Figure 31: *The vizGal user interface client being used to compare vector and scalar data via differencing.*

4.1.4 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 DAP 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.

4.1.5 Federation

The ESGF is a federation of data held by many different ESGF 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 DAP client library calls to the remote OPeNDAP server.

4.1.6 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 OPeNDAP Hyrax. The product services layer, itself, has no awareness of individual files. See section 4.4.7 for more details.

4.1.7 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 a DAP service 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 DAP client 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.

5 Data and Overall Community Impact

There is a growing national and international interest in the benefits ESGF provides to 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 the community defined. Community benefits include a wide range of services and activities that ESGF provides 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 who want 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 has an impact on the following:

- Current CMIP5 activities and preparation for future assessments (for example, the U.S. National Climate Assessment [NCA]);
- Current CSSEF activities that push the design of ESGF to fit into a U.S. DOE-led testbed infrastructure for evaluating the uncertainty of climate models;
- Development of data and metadata facilities to include observations and reanalysis products in CMIP5 (called obs4MIPs);
- Enhancements and improvements to the current climate research infrastructure capabilities through involvement of the software-development community and 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. 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.

Table 7 below represents some of the national and international ESGF data archives the community requires.

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, MLS, TES	H	Observational	Satellite	Atmosphere	Collecting	30 GB
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	20 TB
NCAR	PCM	M	Model	Gridded	Atmosphere Land Ocean	Published	21 TB
NCAR	CCSM	H	Model	Gridded	Atmosphere Land Ocean	Published	1000 TB (1 PB)
NCAR	CADIS	H	Model	Gridded	Atmosphere Land Ocean	Published	140 GB
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	Land	Published	339 MB

				forcing data (U.S. sites)			
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

6 Collaborations

Partnerships and our intent to collaborate are reflected by 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 CMIP5, CSSEF, CCSM, and other data-intensive climate-relative community projects as mentioned in section 5, *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; in some cases, interest in generalizing and enhancing ESG-developed technology and disseminating to a larger audience in other scientific domains was expressed. For example, we 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). As 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' success. For this reason, we positioned ESG-CET team members to 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 project's 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.

7 References

[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.

[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).

Appendix A PMEL Outreach, Papers, Presentations, Poster, Books

Outreach

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

  - *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.

- ***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.

- *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.

- ***April 1, 2009 through September 30, 2009***
  - *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-CET teleconferences and meetings leading to a vigorous bi-directional exchange of ideas and technology. NOAA/PMEL (Steve Hankin, ESG-CET 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-CET.
  - *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***
  - *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 teleconferences 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.

- ***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.

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

  - *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.

  - *Atmospheric Radiation Measurement (ARM) Collaboration*

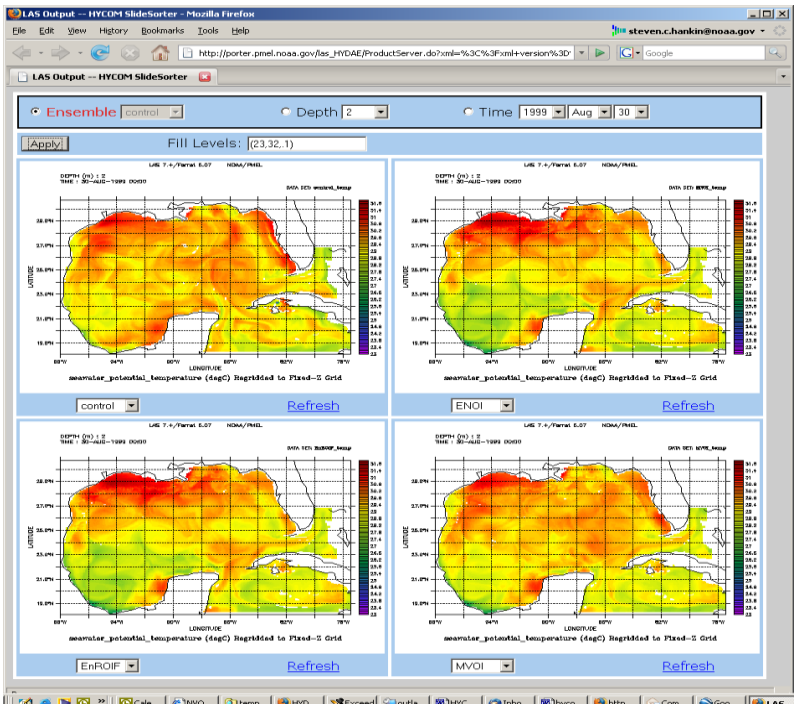
    The team at Argonne has started collaborating with Environment Science Division at ANL, specifically to work with scientists at Climate Research Station on the Data Domain to Model Domain Conversion Package (DMCP) (see URL: http://www.atmos.anl.gov/DMCP/). This recently initiated effort has been exploring 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 shown below, showing the HyCOM model intercomparison.)



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-CET 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, also effecting an active two-way technology transfer between NOAA and ESG-CET.

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

  - *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.

Appendix A Outreach, Papers, Presentations, Poster, Books

A.2 Papers with PMEL authorship

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

  - *SciDAC '11 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: A Data Infrastructure for Data-Intensive Climate Research," *Journal of Physics: Conference Series, SciDAC '11 Conference Proceedings.*

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

  - *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***

  - *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*.

- ***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"*.

- ***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.

A.3 Books with PMEL contributions

- ***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.