



LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-TR-418367

SciDAC's Earth System Grid Center for Enabling Technologies Semi-Annual Progress Report for the Period April 1, 2009 through September 30, 2009

D. N. Williams, I. T. Foster, D. E. Middleton

October 20, 2009

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.

SciDAC's Earth System Grid Center for Enabling Technologies

Semi-Annual Progress Report for the Period April 1, 2009 through September 30, 2009

Principal Investigators

Dean N. Williams³, Ian T. Foster¹, and Don E. Middleton⁵

The Earth System Grid Center for Enabling Technologies Team:

Rachana Ananthakrishnan¹, Frank Siebenlist¹

Arie Shoshani², Alex Sim²,

Gavin Bell³, Robert Drach³

Jim Ahrens⁴, Phil Jones⁴

David Brown⁵, Julien Chastang⁵, Luca Cinquini⁵, Peter Fox⁵, Danielle Harper⁵, Nathan Hook⁵, Eric Nienhouse⁵, Gary Strand⁵, Patrick West⁵, Hannah Wilcox⁵, Nathaniel Wilhelmi⁵, Stephan Zednik⁵

Steve Hankin⁶, Roland Schweitzer⁶

David Bernholdt⁷, Meili Chen⁷, Ross Miller⁷, Galen Shipman⁷, Feiyi, Wang⁷

S. Bharathi⁸, Ann Chervenak⁸, Robert Schuler⁸, Mei-Hui Su⁸



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

¹ Argonne National Laboratory, Chicago

² Lawrence Berkeley National Laboratory

³ Lawrence Livermore National Laboratory

⁴ Los Alamos National Laboratory

⁵ National Center for Atmospheric Research

⁶ National Oceanic and Atmospheric Administration

⁷ Oak Ridge National Laboratory

⁸ University of Southern California, Information Sciences Institute

Table of Contents

<i>SciDAC's Earth System Grid Center for Enabling Technologies</i>	<i>1</i>
<i>1 Executive Summary</i>	<i>4</i>
1.1 Overall goal for this reporting period	4
1.2 Highlights	5
1.2.1 ESG-CET Mid-term Project Review	5
1.2.2 LLNL ESG CMIP3 Portal and R&D Highlights	6
1.2.3 NCAR ESG Gateway Portal R&D Highlights	7
1.2.4 ORNL ESG Gateway Portal R&D Highlights	9
1.2.5 LANL ESG Node Highlights	9
1.2.6 LBNL Storage Resource Manager Highlights	9
1.2.7 PMEL Product Delivery Services Highlights	10
1.2.8 ANL Security, Data, and Services Highlights	10
1.2.9 USC/ISI Mirroring Highlights	11
1.2.10 Rensselaer Polytechnic Institute	11
2 Overall Progress	11
2.1 ESG-CET Gateway	11
2.1.1 Domain object model	12
2.1.2 Publishing Services	12
2.1.3 Deep Storage Data Retrieval	12
2.1.4 Support for High Performance Data Download Client	12
2.1.5 Data Versioning	12
2.1.6 Semantic Search	12
2.1.7 Modeling Metadata	12
2.1.8 Federated Authentication	13
2.1.9 User Attribute Services	13
2.1.10 Metadata Exchange	13
2.1.11 Data Publishing Operations	13
2.1.12 Packaging	13
2.1.13 Documentation	14
2.2 ESG-CET Data Node	14
2.2.1 Data Node Installation	15
2.2.2 ESG-CET Software components to be installed:	16
2.2.3 Operational workflow:	16
2.2.4 Support:	17
2.3 ESG-CET Cyber Security	17
2.4 Storage Resource Management (SRM) Data Movement	18
2.4.1 DataMover-Lite (DML)	18
2.4.2 Bulk Data Mover (BDM)	19
2.5 Data Replication	20
2.6 Product Services	21
3 ESG-CET Group Meetings	21
3.1 ESG-CET Executive Meeting (May 10th)	22

3.2 ESG-CET Project Review (May 11 – 12)	22
4 Collaborations	22
5 Outreach, Papers, Presentations and Posters	22
5.1 Outreaches Activities	22
5.1.1 WCRP/WGCM Meeting	22
5.1.2 Unidata and the Climate and Forecast Conventions (CF)	23
5.1.3 SciDAC 2009 Conference	23
5.1.4 Supercomputing 2009 (SC09) Bandwidth Challenge	23
5.1.5 UK e-Science Envoy Visit to Chicago and NCAR	23
5.1.6 SciDAC Outreach Center	23
5.1.7 GFDL Meeting on Climate Services	23
5.1.8 USGS Visit to NCAR	23
5.1.9 Computing in the Atmospheric Sciences, Annecy, France, September 2009:	23
5.1.10 NOAA Geophysical Fluid Dynamics Laboratory	24
5.1.11 NOAA Office of Climate Observations (OCO)	24
5.1.12 Hybrid Coordinate Ocean Model (HyCOM) consortium (NOAA, Navy, et. al.)	24
5.1.13 US Integrated Ocean Observing System (IOOS)	24
5.1.14 Global Earth Observation Integrated Data Environment (GEO-IDE)	24
5.2 Papers:	25
5.2.1 Grid Computing Environments Workshop Paper	25
5.2.2 TeraGrid Conference Paper	25
5.2.3 IEEE Computer Graphics and Application Paper	25
5.2.4 Eos, Transactions, American Geophysical Union Paper	25
5.2.5 IEEE International Conference on Data Mining Paper	25
5.2.6 IEEE International Conference on Data Mining Paper	25
5.3 Talks:	25
5.3.1 Dr. Steve Koonin's Visit to LLNL	25
5.3.2 Dr. Steve Koonin's Visit to ANL	25
5.3.3 Space Mission Challenges for Information Technology	26
5.3.4 TeraGrid 2009	26
5.3.5 OceanObs 2009	26
5.3.6 Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology	26
5.4 Posters:	26
5.4.1 Supercomputing 2009 (SC09)	26

1 Executive Summary

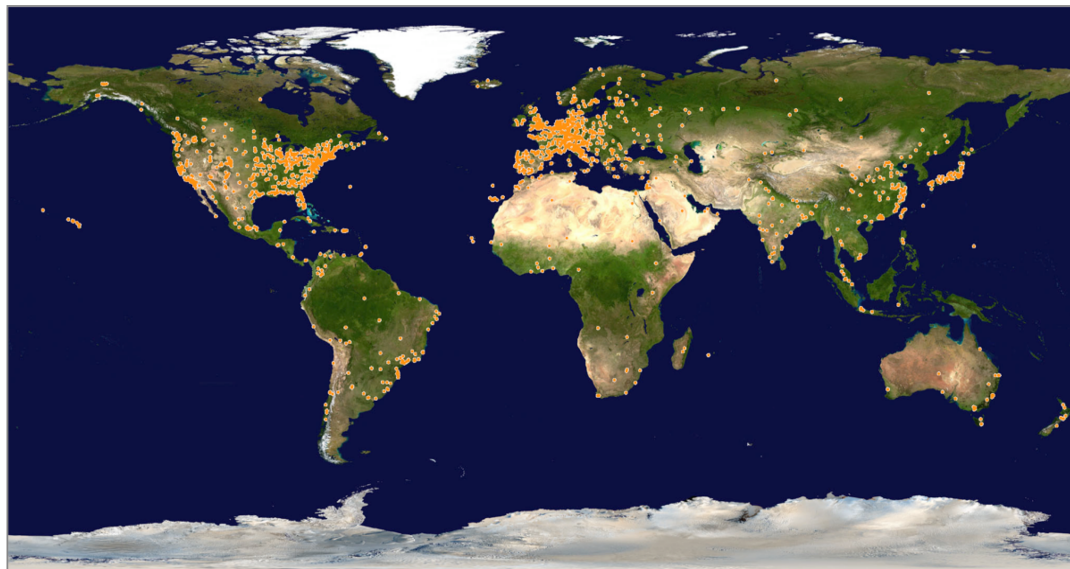
This report summarizes work carried out by the ESG-CET during the period April 1, 2009 through September 30, 2009. It includes discussion of highlights, overall progress, period goals, collaborations, papers, and presentations. To learn more about our project, and to find previous reports, please visit the [Earth System Grid Center for Enabling Technologies \(ESG-CET\)](#) website. This report will be forwarded to the DOE SciDAC program management, the Office of Biological and Environmental Research (OBER) program management, national and international collaborators and stakeholders (e.g., the Community Climate System Model (CCSM), the Intergovernmental Panel on Climate Change (IPCC) 5th Assessment Report (AR5), the Climate Science Computational End Station (CCES), the SciDAC II: A Scalable and Extensible Earth System Model for Climate Change Science, the North American Regional Climate Change Assessment Program (NARCCAP), and other wide-ranging climate model evaluation activities).

The ESG-CET executive committee consists of David Bernholdt, ORNL; Ian Foster, ANL; Don Middleton, NCAR; and Dean Williams, LLNL. The ESG-CET team is a collective of researchers and scientists with diverse domain knowledge, whose home institutions include seven laboratories and one university: Argonne National Laboratory (ANL), Los Alamos National Laboratory (LANL), Lawrence Berkeley National Laboratory (LBNL), Lawrence Livermore National Laboratory (LLNL), National Center for Atmospheric Research (NCAR), Oak Ridge National Laboratory (ORNL), Pacific Marine Environmental Laboratory (PMEL), and University of Southern California, Information Sciences Institute (USC/ISI). All work is accomplished in close collaboration with the project's stakeholders, domain researchers, and scientists.

1.1 Overall goal for this reporting period

During this semi-annual reporting period, the ESG-CET team continued its efforts to complete software components needed for the ESG Gateway and Data Node. These components include: Data Versioning, Data Replication, DataMover-Lite (DML) and Bulk Data Mover (BDM), Metrics, Product Services, and Security, all joining together to form ESG-CET's first beta release. The launch of the beta release is scheduled for late October with the installation of ESG Gateways at NCAR and LLNL/PCMDI. Using the developed ESG Data Publisher, the ESG II CMIP3 (IPCC AR4) data holdings – approximately 35 TB – will be among the first datasets to be published into the new ESG enterprise system. In addition, the NCAR's ESG II data holdings will also be published into the new system – approximately 200 TB. This period also saw the testing of the ESG Data Node at various collaboration sites, including: the British Atmospheric Data Center (BADC), the Max-Planck-Institute for Meteorology, the University of Tokyo Center for Climate System Research, and the Australian National University. This period, a total of 14 national and international sites installed an ESG Data Node for testing.

During this period, we also continued to provide production-level services to the community, providing researchers worldwide with access to CMIP3 (IPCC AR4), CCES, and CCSM, Parallel Climate Model (PCM), Parallel Ocean Program (POP), and Cloud Feedback Model Intercomparison Project (CFMIP), and NARCCAP data.



Over 2,700 sites
120 countries

Figure 1: Geographic distribution of the users that downloaded data from ESG web portals.

1.2 Highlights

1.2.1 ESG-CET Mid-term Project Review

May 11th and 12th, SciDAC and BER program management conducted a mid-term assessment review of the ESG-CET project. Present at the review were six distinguished panel reviewers knowledgeable in climate, computer, and computational science, and software design and development. The thoughtful and insightful feedback from the panel was extremely helpful to the ESG-CET executive team. As suggested from the review panel comments, the ESG-CET team is presently working to increase computer science journal publications. We are also working more closely with stakeholders to improve communications, system data integrity, and collaboration with SciDAC's Outreach Center, and to achieve a better software stack installation solution. The following summary indicates how well ESG-CET did in the overall ranking process. For a more detailed look at reviewer comments and ESG-CET responses, please contact SciDAC or BER program management.

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

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: The SciDAC and BER program will continue to fund ESG-CET through its recommended term.

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

1.2.2 LLNL ESG CMIP3 Portal and R&D Highlights

The [WCRP CMIP3 Multi-Model Database](#) ESG portal continues to provide the world’s climate scientists with the most complete collection of climate simulation data. This portal serves over 35 terabytes of data to more than 3,000 scientific groups worldwide with an average download rate of 700 GB/day. Thus far, more than 530 research papers have been published based on the CMIP3 Multi-Model archive and over 800 TB of data have been downloaded.

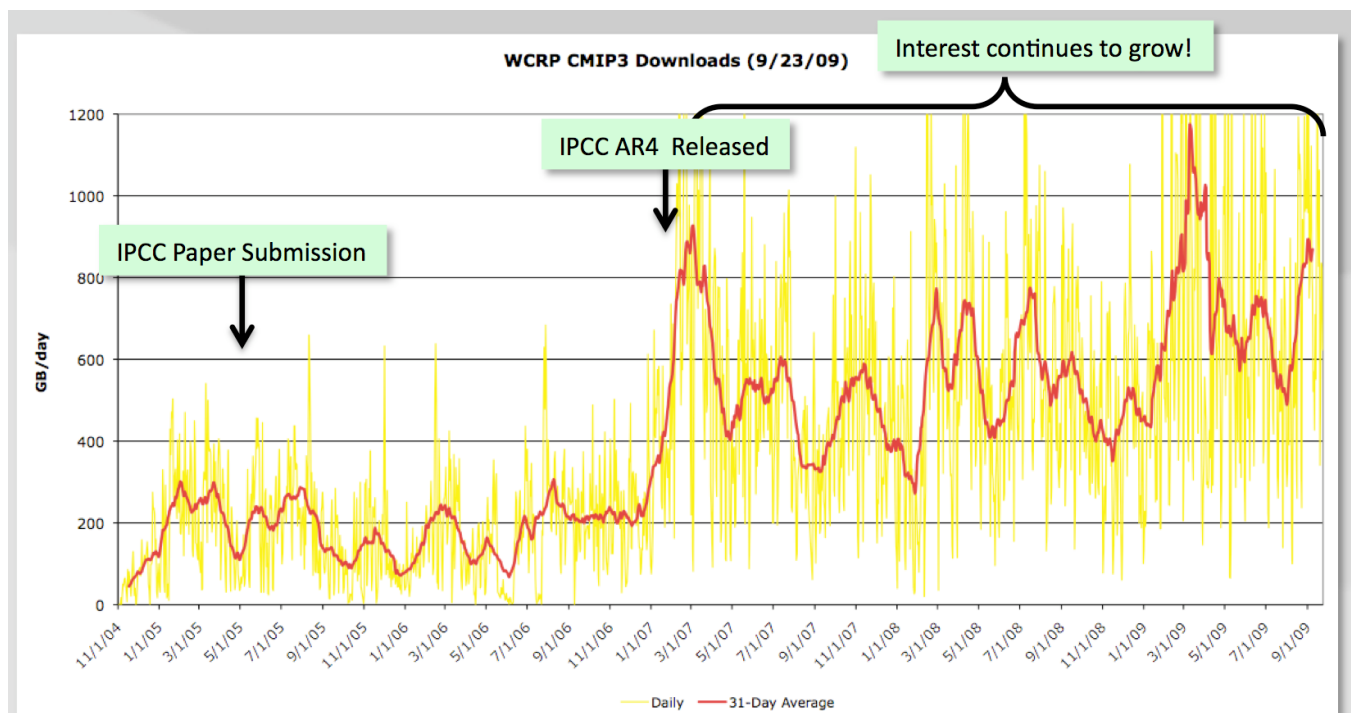


Figure 3: ESG Monthly LLNL/PCMDI CMIP3 Gateway and FTP Downloads.

Leading ESG Data Node Development, PCMDI has helped to deploy ESG Data Nodes at LLNL, NCAR, ORNL, ANL, GFDL, LANL, NASA JPL, NASA Goddard, UC Davis, ANU, University of

Tokyo, JAMSTEC, BADC, DKRZ, University of Michigan, USC/ISI, and NCDC. To assist in this rapid deployment, PCMDI spent considerable efforts on Data Node installation, cutting installation time down from two or more weeks to just under an hour. We achieved this speedy installation process via (1) supporting Red Hat Linux EL5 and CentOS, with a bash script being created to preload and install various software components needed to build the Data Node; and (2) creating a virtual machine (VM) image, complete with CentOS and required libraries. We also created pre-installed Data Node for immediate use. We are currently working to enable the use of Yum as a third method for installing a Data Node. Yum is an automatic updater and package installer and remover for RPM systems. It automatically computes dependencies and determines what steps must be taken to install packages.

1.2.3 *NCAR ESG Gateway Portal R&D Highlights*

While our software development work is focused towards the creation of the next generation system, the ESG web portal at NCAR continues to serve the user community. The total number of registered users is approaching 15,000, and the projected cumulative data downloaded by the end of 2009 is expected to be around 150 TB (see Figures 4 and 5).

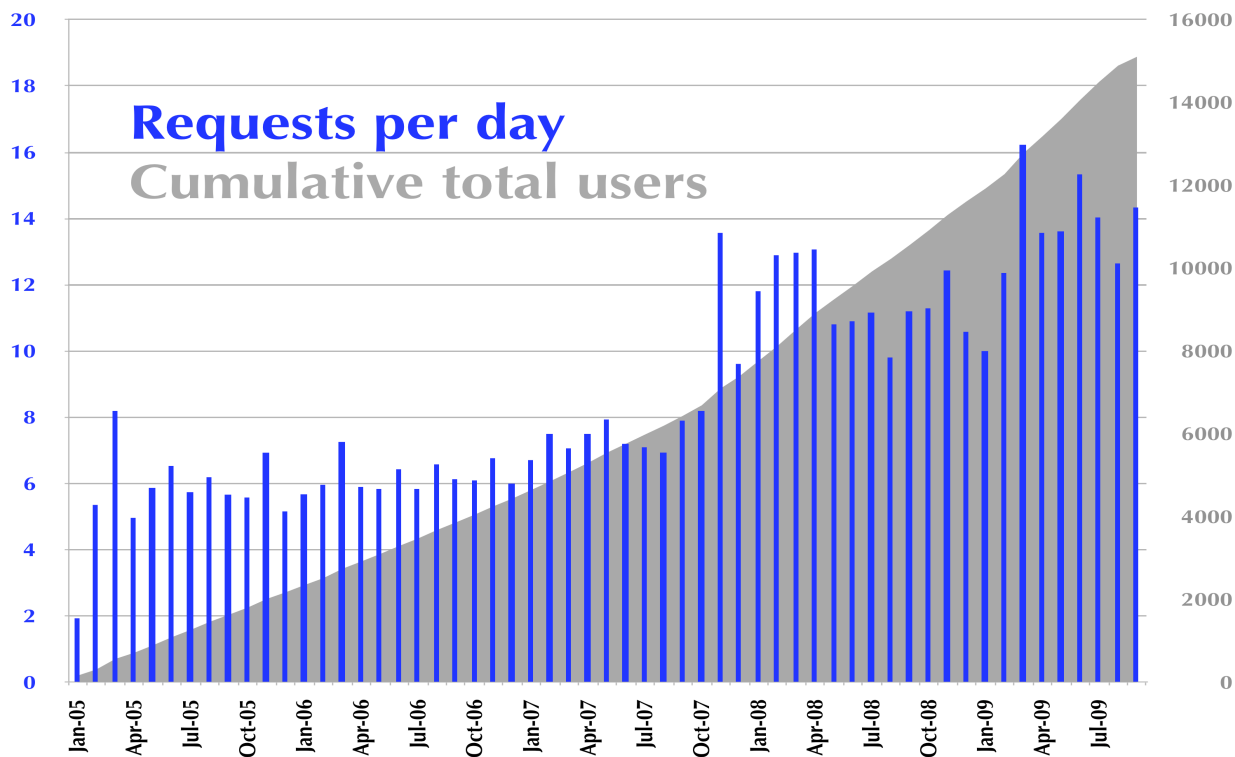


Figure 4: Number of registered users at the ESG NCAR operational portals.

NCAR ESG-CET portal download volume

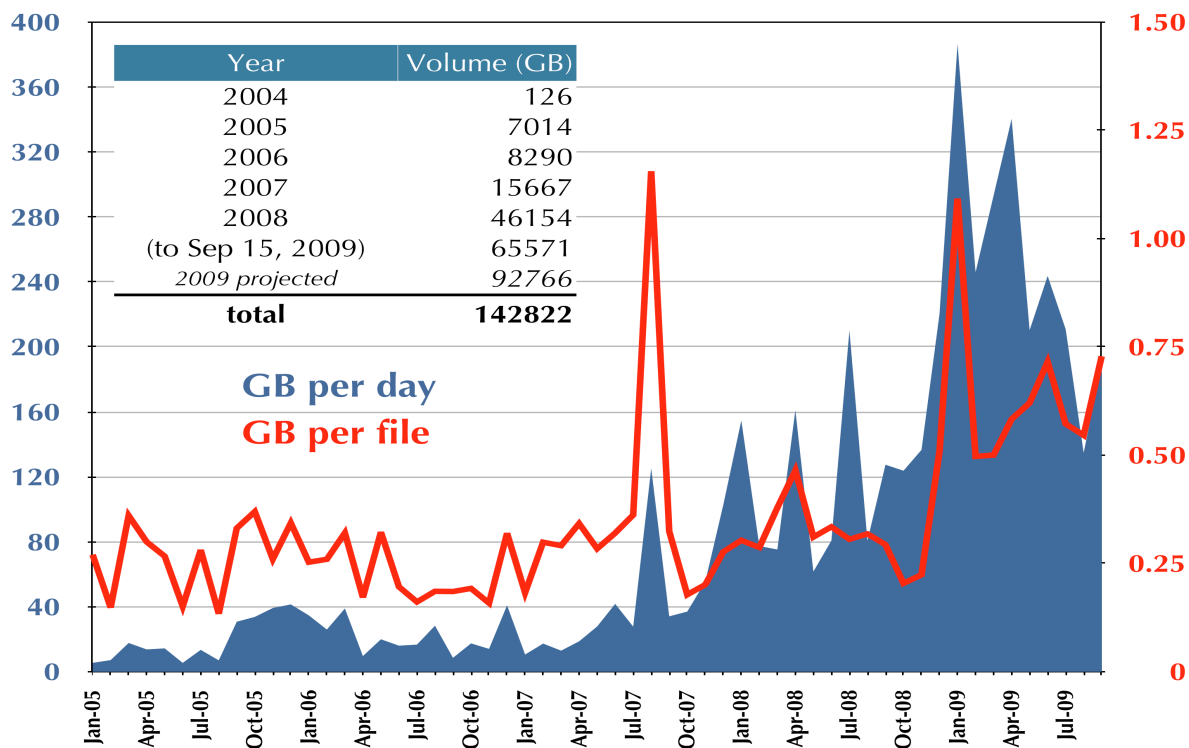


Figure 5: Data download metrics for the ESG web portal at NCAR.

During the past six months, we have expanded, re-factored, and exercised the Gateway software stack to support the first release of the ESG-CET system scheduled for the end of October 2009. The functionality targeted in this first release will include support for the current LLNL/PCMDI and NCAR data holdings (respectively, the CMIP3 data archive and distribution and the CCSM, PCM, and NARCCAP data distributions) and readiness for the upcoming CMIP5 data streams. In particular, we have worked to integrate the Gateway software with the LLNL/PCMDI publishing application running on the Data Node, the BestMan middleware for file retrieval from deep storage, the TDS and GridFTP servers, and the MyProxy server for X509 certificate retrieval.

During this reporting period, we have brought to an almost operational status the infrastructure for establishing a first level of federation services among ESG Gateways, and for achieving interoperability with other partner Data Centers. These federation services include user single-sign-on via OpenID, user attributes exchange (including access control information) via a SAML-based attributes service, and metadata exchange via the OAI-PMH protocol.

Successful collaboration continues between ESG, ESC, and Metaphor projects with the goal of providing metadata services for accurately describing models, model components, and simulations, and for linking models to model output data.

We have started work to provide a Gateway software distribution that can be easily installed, operated, and upgraded by any center that wishes to join the ESG collaboration. While much work still needs to be done to achieve this goal, the first ESG Gateways are being brought to operational status at NCAR, LLNL/PCMDI, and ORNL.

Lastly, in tandem with ESG-CET's heightened focus on use-case driven requirements, development, and design, NCAR has enhanced its software engineering process and infrastructure. This work has emphasized better collaborative development, comprehensive issue tracking and prioritization, and continuous integration and release management. We intend to open up the issue database to public access during the next reporting period, and expect this to further improve our communication and collaboration in the ESG-CET context.

1.2.4 ORNL ESG Gateway Portal R&D Highlights

In order to better support ORNL's plans for a large-scale ESG installation and archive associated with CMIP5 and other activities, we have added several researchers associated with our National Center for Computational Sciences (NCCS) to the team. Galen Shipman, who leads the NCCS Technology Integration Group, and Feiyi Wang and Russ Miller from the TechInt group, have already started assisting with test installations of the new software stacks and working with other NCCS staff to plan an ESG deployment within NCCS.

Using existing facilities on ORNL's TeraGrid network, we now have an installation of the new Data Node software stack operational, including testing of the ability to publish data to the PCMDI Gateway. We are also in the process of standing up a Gateway at ORNL. (We have an initial Gateway operational and we are in the process of customizing it.)

These new installations run alongside our production systems, which include a Data Node for the CCSM portal at NCAR and the standalone C-LAMP portal. During the last six months, ORNL has delivered close to 8 TB of climate model data to ESG users from its production systems.

Finally, we have begun to make connections to researchers in the ORNL Environmental Sciences Division associated with environmental data archives including ARM, CDIAC, and the biogeochemical dynamics DAAC, all of which are interested in the possibility of publishing some of their datasets via the ESG. We are pursuing some small initial experiments with them, as time permits, to better assess what work would be required on both sides to allow effective use of the ESG for their data.

1.2.5 LANL ESG Node Highlights

LANL continued to serve data both through the NCAR ESG portal and through direct web access. To better facilitate access, LANL obtained new hardware to replace its previous oceans11 server, and worked to replicate existing data into the new ESG enterprise system.

Work was also completed on the re-design of the SCRIP interpolation package. New features include subroutine interfaces (rather than off-line computations), reformulated algorithms to improve robustness, new SCRIP subroutines into ESG (through CDAT and Ferret applications), general ocean-specific analysis routines (meridional heat transport, barotropic stream function, etc.), and the addition of Gridspec support to SCRIP interpolation routines.

Working with Jim Ahrens' visualization team, LANL continued prototyping ParaView for analyzing large ocean data sets accessing data generated as a part of the Climate End Station runs at ORNL. This work includes incorporating parallel visualization and analysis capabilities in ESG.

1.2.6 LBNL Storage Resource Manager Highlights

The Berkeley Storage Manager (BeStMan) is a new implementation of the Storage Resource Management (SRM) standard, based on a modular design that makes it easier to interface to various storage systems, including Mass Storage Systems (MSSs). Versions of BeStMan have been developed

for HPSS at LBNL/NERSC and ORNL, and for MSS at NCAR, and deal with different security mechanisms. The following tasks have been accomplished:

- BeStMan servers have been installed, passed initial tests, and are now running at LBNL/NERSC, NCAR, and ORNL.
- New ESG Gateway portals are continually testing underlying HPSS and MSS access.
- The ESG Metadata Publishing tool for browsing all file information stored on deep storage use BeStMan servers in all three sites (LLNL/PCMDI, NCAR, ORNL). Support for this activity is ongoing.

1.2.7 *PMEL Product Delivery Services Highlights*

At the heart of PMEL's contribution to ESG-CET is the Live Access Server (LAS), a development activity with origins that date back to 1993. LAS is an XML-configurable workflow engine that has been utilized in support of widely varied climate applications. The introduction of LAS into the ESG in 2006 brought the accomplishments of those efforts into the ESG project, and expanded the positive impacts of ESG-specific enhancements to LAS back to those other communities. In particular, ESG enhancements to the model-intercomparison capabilities of LAS have benefited the dissemination of operational global ocean circulation model outputs from the Hybrid Coordinate Ocean Model (HYCOM), a project that DOE has supported in the past under ocean modeling funding. HYCOM LAS data services are available at <http://las.hycom.org/las/getUI.do>.

ESG enhancements have also benefited the seasonal to inter-annual climate forecast community, through the Global Ocean Data Assimilation Experiment (GODAE), with a number of LAS servers installed internationally (<http://www.godae.org/Ocean-products.html>), including the installation at the US GODAE Server (<http://www.usgodae.org/las/servlets/dataset>). ESG developments have also had positive impacts on the ocean-climate observations community through LAS. The enhancements made to LAS in support of the ESG data publication process have found their way into the Observing System Monitoring Center (OSMC -- <http://osmc.noaa.gov/Monitor/OSMC/OSMC.html>) and related ecosystem-monitoring activities at http://www.pfeg.noaa.gov/products/las_5_1.html.

As ESG-funded enhancements are added to LAS, there will be continued positive impacts across a wide range of significant projects that utilize the software. In addition, LAS is currently being used as the visualization tool for the Surface Ocean Carbon Atlas (SOCAT). It provides access to and maps of the collected underway carbon observations, as well as secondary tools to monitor quality control of the data by carbon community members, which also benefits from expanded capabilities in LAS for data intercomparison being developed for ESG-CET.

1.2.8 *ANL Security, Data, and Services Highlights*

ANL has worked with the ESG team to provide Single Sign-on (SSO) solutions for the ESG-CET project. A Web SSO and PKI SSO solution has been integrated with the ESG Gateway to provide single sign-on across web and non-web applications of the project.

In the Web SSO realm, we are using the OpenID standard and leveraging the OpenId4J open source development effort. We have extended the OpenID to allow for a white list of trusted Identity Providers (IdPs) that is enforced by the OpenID Service Providers (SPs). We have also added pluggable attribute providers, so as to include arbitrary attributes as a part of the OpenID authentication. The relying party can use these attributes for authorization purposes.

To facilitate attribute processing in the PKI realm, the Argonne team worked to provide a SAML-based attribute-processing solution. This solution provides strict data types attribute handling and allows for better interoperability with external partners. We leveraged the open-source SAML processing library (OpenSAML) to provide a framework and some samples for attribute providers, which publish attributes and have them embedded as a part of either OpenID protocol or PKI credential. Currently, the framework is being integrated for use with the MyProxy server, which provides the basis for PKI SSO framework, such that a user's attributes are embedded in the credential issued to the user. These attributes are strongly typed and can be parsed and consumed by the resource for enforcement. This work has been integrated with the ESG Gateway for the October testbed.

1.2.9 *USC/ISI Mirroring Highlights*

The ISI team has successfully integrated several ESG clients and services necessary to provide data mirroring capabilities, including MyProxy service, ESG metadata catalog, NCAR authorization token service and data transfer operations using DML and wget.

1.2.10 *Rensselaer Polytechnic Institute*

RPI has implemented a prototype library that reads an NcML file utilizing the joinExisting aggregation type, opens local file access to the nested netCDF locations defined by the NcML aggregation and builds a representation of the aggregated dataset's structure in memory. NetCDF structure queries (nc_inq_x(...) methods) can be answered directly from the in-memory cache.

The nc_get_vara() data retrieval method has been implemented such that it accesses the nested netCDF locations and builds an aggregated data response. Other data retrieval methods have not been implemented, but follow a similar structure and should not take a great deal of time to implement.

Short-term tasks include a code review using static analysis tools (fortify), memory analysis tools (valgrind, shark), and peer review. Following the review and subsequent updates, the code implementation of the netCDF v3 API will be completed and the NcML classes will be integrated into a branched copy of the RNI client library such that the client library will support local netCDF, OPeNDAP, remote netCDF over ppt & gsiftp, and limited support of local NcML files.

2 Overall Progress

During this reporting period, progress was made in key areas that are necessary to meet ESG-CET objectives, goals, and milestones. This section provides greater technical depth and presentation of the components needed for the beta release.

2.1 ESG-CET Gateway

The first release of the Gateway software stack is scheduled to take place at the end of October 2009, with the initial deployment of the NCAR and LLNL/PCMDI Gateways in support of the existing data holdings and of the new incoming CMIP5 data streams. In the past six months, all Gateway development has been targeted to prepare for this upcoming release, in terms of inserting or completing needed functionality, testing, bug fixing, packaging, and documentation. The following is a brief progress report for the main Gateway-related activities that took place during this time.

2.1.1 *Domain object model*

The domain object model, and its corresponding relational model persisted in the database, were expanded and polished to accommodate the storage and access of all metadata necessary to support the full functionality of the first release of the Gateway services. The relational model is now composed of approximately 150 tables organized in several schemas.

2.1.2 *Publishing Services*

The Gateway web application was instrumented with the machinery necessary to parse THREDDS data catalogs, create hierarchies of domain model objects, and persist them to the relational database. Additionally, this functionality was exposed as a Hessian web service that can be invoked by the publishing application running on the Data Node, requiring secure mutual authentication via digital certificates.

2.1.3 *Deep Storage Data Retrieval*

Work was completed to integrate the deep storage retrieval functionality provided by the BeSTMan middleware into the Gateway, and to develop user interfaces and controls to invoke, monitor and access file retrieval requests submitted to the back-end. Thorough testing versus the NCAR and NERSC archives was completed. Testing with the ORNL HPSS is under way.

2.1.4 *Support for High Performance Data Download Client*

The Gateway application was augmented with the capability of supporting high performance, multi-files download via the Data Mover Light (DML) client. DML can now be downloaded and configured directly through the Gateway, and started via a file list in XML format that the user can request for either files stored on rotating disk, or transferred to the Gateway cache from deep storage archive.

2.1.5 *Data Versioning*

Much effort has been spent in analyzing and capturing use case and requirements for data versioning, mostly driven by the need to support the replication of the LLNL/PCMDI CMIP5 archive at a certain point in time to European Data Centers (specifically, BADC and DKRZ). The required changes to the domain object model and the Gateway interfaces have been prototyped, but because of time constraints they will not be part of the first software release.

2.1.6 *Semantic Search*

An alternative user interface has been developed on top of the ESG semantic services for data search and discovery, which more closely resemble the general characteristics of standard business web sites, and which might be more intuitive to the user. A final decision about which interface will be included in the Gateway first release has not yet been made.

2.1.7 *Modeling Metadata*

The ESG collaboration has worked to support the efforts of the Earth System Curator (ESC) project to develop the full infrastructure for capturing and displaying accurate model metadata within the Gateway web application. The information displayed within the model “trackback” pages has been greatly expanded to include tens of scientific properties, a detailed data grid specification, the full hierarchy of model components composing the simulation, and all simulation inputs (forcing, boundary conditions, etc.) and output data streams. Additionally, services have been developed to further the user’s understanding by retrieving the exact definition of all metadata terms from a community-maintained vocabulary web service, which is expected to be made available in the near future at the British

Oceanographic Data Center (BODC). In this area, ESG and ESC are working in direct collaboration with the European Metaphor project to develop common vocabularies, map ontological models, and provide general interoperability.

2.1.8 *Federated Authentication*

We have made much progress during the past six months towards both federation among ESG Gateways and interoperability with other non-ESG partner data centers. We have finalized the infrastructure required to support user single-sign-on via the OpenID protocol, and demonstrated successful cross-site authentication between the NCAR and LLNL/PCMDI Gateways and with the BADC web portal. This use of OpenID technology will allow users to register at only one site, where their credentials will be stored, and then carry their authenticated identity as they navigate and access data throughout the ESG federation.

2.1.9 *User Attribute Services*

ESG, in collaboration with its European partners, has settled on the use of SAML (the Security Assertion Markup Language) as the enabling technology to exchange user attributes and access control information among sites. Each ESG Gateway or partner Data Center will deploy a SAML-based Attribute Service, which other Gateways can securely query to retrieve attribute information about a specific user. This information is needed both to authorize the user to access specific datasets that are controlled by an access group managed by another site (for example, IPCC AR5 data stored at BADC, for which access permission needs to be granted by LLNL/PCMDI administrators), and to store complete data access metrics. Prototype SAML Attribute Services have been developed independently on top of the ESG and NERC Data Grid (NDG) user information management stacks, and first interoperability between the two services has been successfully demonstrated.

2.1.10 *Metadata Exchange*

Work is almost completed to bring the software infrastructure for metadata exchange among Gateways and Data Centers to an operational status. This infrastructure, based on the well-known and widespread OAI-PMH protocol, will allow each Gateway to act as both a metadata repository (for exporting records to other centers) and a metadata harvester (for importing records from other centers), in a wide variety of metadata formats including RDF, DIF, DC and in the future ISO. The Gateway web application includes an administrative interface that can be used to trigger on demand OAI harvesting from a remote repository of choice, and to monitor and report the status of the harvesting job.

2.1.11 *Data Publishing Operations*

Following the installation of the Data Node software stack at NCAR, work has begun to script the publication of all current ESG NCAR data holdings (CCSM, PCM and NARCCAP simulations) to the prototype Gateway. This work is approximately 30% completed, although full publication will require some upgrades to the Data Node publishing application to support non-standard model runs, multiple deep storage archives, and multiple data access services.

2.1.12 *Packaging*

In order to facilitate the deployment and configuration by site administrators, the packaging and build process of the Gateway software stack has been converted to Maven. Among other things, Maven allows consistent tracking of library dependencies; centralization of deployment parameters in a single site-specific project file; and easy management of software builds for different profiles.

2.1.13 Documentation

Work is underway to provide a comprehensive Gateway Administrator Guide containing all relevant information about Gateway installation, maintenance, operations, and functionality, including the underlying architecture, data model, and exposed services. At this time, this guide is still largely incomplete, but it will represent one of the main team priorities once the first Gateways have been deployed operationally.

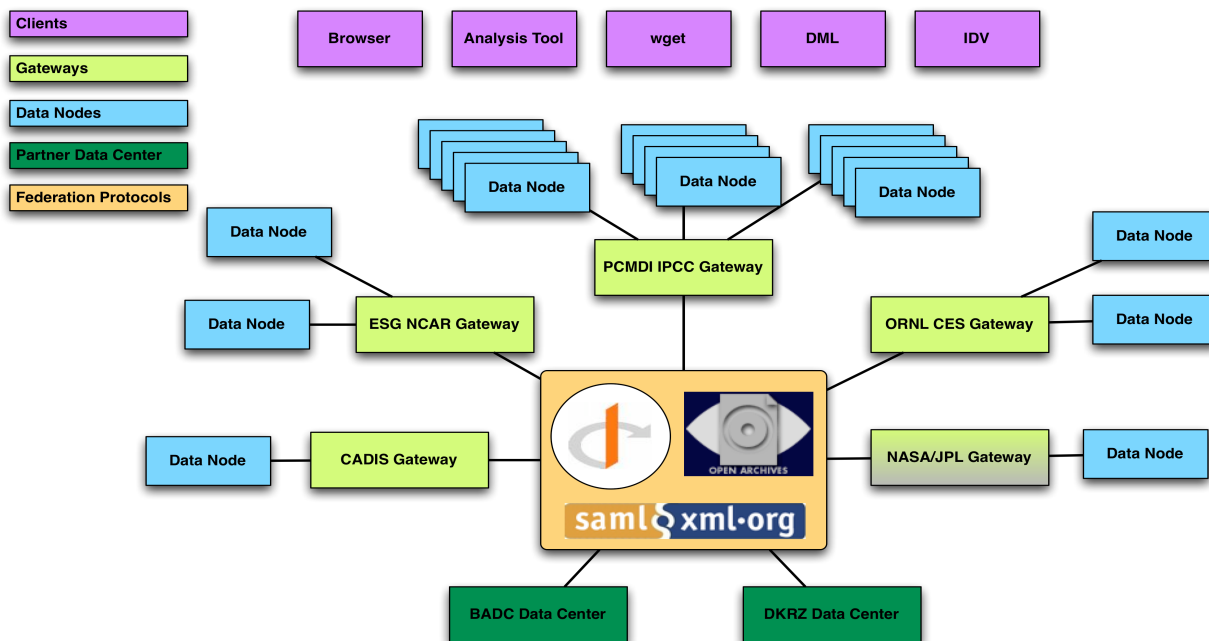


Figure 6: The ESG Federation as envisioned in Fall 2009, with the technologies used to achieve interoperability highlighted at the center of the picture (OpenID, OAI-PMH, and SAML).

2.2 ESG-CET Data Node

The ESG-CET system consists of two major components: 1) Gateways that support portal services, which serve as interfaces to end-users who can search, discover, and request data and data products, and 2) Data Nodes where the data actually resides. The Data is “published” to the Data Node, which makes the data visible to a Gateway and enables its delivery to end-users. It is expected that Gateways will only be installed by a small number of centers devoted to serving data (e.g., LLNL/PCMDI, NCAR, GFDL, BADC, DKRZ, ANU, JAMSTEC), whereas it is hoped that most climate modeling centers will install the Data Node software through which they can serve their model output.

Deploying a Data Node software stack involves scanning netCDF Climate and Forecast (CF) convention files and publishing data to a secure ESG-CET Gateway. For groups that are unable to serve their own data through an ESG Data Node, model output can be sent to LLNL/PCMDI, which will then be made available to end-users through the LLNL/PCMDI Gateway.

For CMIP5 it should be noted that however data are published, it will not be accessible to users until a questionnaire has been filled out, providing documentation of the model and the simulations performed. This questionnaire is being developed by the METAFOR project at the request of the WCRP’s Working Group on Coupled Modelling (WGCM).

2.2.1 **Data Node Installation**

We provide a summary of the requirements associated with a data node deployment.

2.2.1.1 *CMIP5 Hardware, storage, and network requirements:*

The hardware, storage, and network requirements for an ESG-CET Data Node are:

- 1) Hardware of your choice running the latest version of CentOS or Red Hat Enterprise Linux (RHEL). For example:
 - a. Dell PowerEdge 2850: 2 core x 3.2 GHz Intel, 4 GB RAM, RHEL 5.3, 26 TB disk
 - b. Dell R610 2-CPU E5540, 32 GB RAM, CentOS 5.3, 24 TB disk, direct attachment to 250 TB RAID archive
 - c. Other versions of Linux/Unix will probably work, but will not be supported at this time. As time progresses, additional operating systems will be added to the support list.
- 2) Storage hardware systems are up to the host site, but for CMIP5 fast access we are requiring spinning disk and **not** tape storage (even though tape storage is supported by the ESG-CET framework).
- 3) The faster the network, the better. Some groups will have 2 x 10 Gbps network access while others will have less than 1 Gbps. Initially, LLNL/PCMDI will have 2 x 10 Gbps network access and is working to connect to a 100 Gbps network.

2.2.1.2 *Data Node Software prerequisites:*

Once your hardware (with operating system), storage, and network are in place, make sure the following software prerequisites are installed before initiating the Data Node installation script:

- 1) gmake (GNU make) – usually preinstalled on Redhat Linux
- 2) svn (subversion client) – usually preinstalled on Redhat Linux
- 3) JDK 6+ (Java 1.6+) – usually installed on Redhat Linux
- 4) An account on an ESG-CET MyProxy server, with Publisher role. This can be created from the ESG-CET Gateway Portal (must contact the host Gateway before publishing)
- 5) X11 libraries and headers (xserver-xorg-dev, libx11-dev, libxt-dev, libxrender-dev). These are only used to build the publisher graphical user interface (GUI). If absent, the publisher scripts will still build, but the GUI will not be available.
- 6) Development library headers are also needed, namely zlib, termcap, and readline. They are used when building the PostgreSQL database. It is recommended that the operating system be installed with all the development headers for the distribution.
- 7) A host with port 80 accessible for incoming traffic from the Gateway. *[Note: In addition to the installed ESG-CET Gateway software, port 443 must also be open for incoming SSL requests.]* The Gateway reads published catalogs from the THREDDS server listening on those ports.

2.2.1.3 *Installation script:*

The installation script contains SVN software download commands and bash scripts to obtain, build, and configure the ESG-CET Data Node for you. The installer is queried to answer a few questions about the Data Node configuration for their particular environment (e.g., “Do you want to install an analysis and visualization product server?” or “Do you want a GridFTP server?”). Default choices are provided to the installer and in most cases recommended. The time it takes to install the ESG-CET Data Node is approximately 1 hour. Experienced installers have installed the Data Node software stack in 30 minutes or less. The ESG-CET Data Node script installation can be found at the following URL:

- http://rainbow.llnl.gov/dist/datanode/ESG_DataNode_install.tgz

2.2.1.4 *Virtual Machine (VM) installation:*

The ESG-CET software stack may be installed on a traditional 'bare metal' Linux distribution, as described above, or distributed via a virtual machine (VM). A VM is a software platform that provides the ability to run a complete and sovereign operating system (a guest OS) as an application inside another operating system (the host OS). The guest OS executes software applications identically to a physical machine. The ESG-CET VMs contain the fully installed CentOS operating system (with requisite libraries et. al. installed) and the ESG-CET Data Node software stack. (This eliminates the need to check for software prerequisites.) VMs take advantage of the benefits of hardware virtualization – specifically, better security, hardware insulation, portability, ease of backup, and protection against potential software conflicts. For ready operation with minimal configuration, follow the links to the CentOS ESG-CET Data Node VM installation files:

- http://rainbow.llnl.gov/dist/datanode/ESG_CentOS_Linux_2.6.x_kernel.v4.dn.tgz
- http://rainbow.llnl.gov/dist/datanode/ESG_CentOS.v4.dn.md5

[Please note that the first link connects to a VMWare CentOS virtual machine and is 1.6GB!]

2.2.2 **ESG-CET Software components to be installed:**

The components of the ESG-CET Data Node that will be installed are:

- 1) PostgreSQL (relational database) – contains the results of scanning and harvesting metadata from the input (netCDF CF) files. The database also contains other information required for system operation.
- 2) CDAT/CDMS – CDAT is based on the Python object-oriented language. CDMS is a sub-module of CDAT used by ESG-CET to scan data.
- 3) Python ESG-CET package – Publication modules, scripts, and GUI.
- 4) Apache Tomcat (Java servlet container) – web application server that runs the THREDDS application.
- 5) THREDDS Data Server (TDS) – a servlet for cataloging ESG-CET Data Node files and data resources used for reading catalogs and publishing to the Gateway.
- 6) Live Access Server (LAS) – servlet for generating visualization products *[Note: LAS product service is optional and not yet ready for distribution. Coming soon.]*
- 7) Globus Toolkit / MyProxy – In our distributed system it is important to provide data security and integrity as well as user authentication. The MyProxy client is used for Single Sign-On to the Gateway.
- 8) GridFTP server – GRID-enabled FTP server.
- 9) Test publication – Publish a test data file.

2.2.3 **Operational workflow:**

As illustrated in the figure below, data is published using the Data Node software through a series of operations. The publisher:

1. Obtains a proxy certificate.
2. Scans a set of files: a) associates files to datasets, and b) caches file metadata in the node database.
3. Generates THREDDS catalogs and reinitializes the THREDDS Data Server (TDS).
4. Reinitializes the Live Access Server (LAS) based on the TDS catalog.
5. Publishes to a Gateway: a) contact Gateway via web service, and b) allow the Gateway to harvest the newly created TDS catalogs.

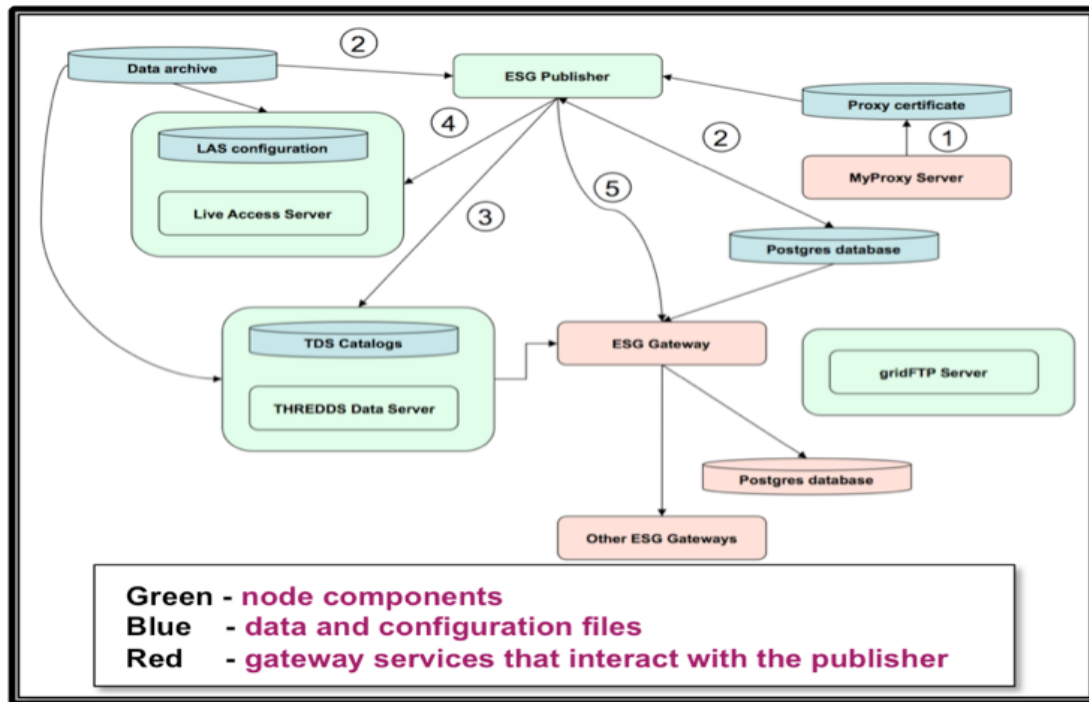


Figure 7: ESG Data Node where data is stored and published.

2.2.4 Support:

For ESG-CET software and installation help and support, please send your e-mail questions to: esg-node-user@lists.llnl.gov

2.3 ESG-CET Cyber Security

Argonne has facilitated an option for high performance download of data that users browse and choose through the Gateway. In the current ESG infrastructure, end user download of data was available only using HTTP. While simple to use, performance via HTTP is limited.

In the last six months, the Argonne team has worked with the Gateway team and the Data team to provide the option of downloading using GridFTP. This system works with the PKI SSO infrastructure to provide users with credentials for authentication with the GridFTP server and with the Gateway Token Authorization Service, to authorize download requests.

In this scheme, users log into the Gateway using their OpenID credential and browse the metadata to choose the data for download. If the data chosen is on a Data Node that provides gsiftp download, an option for using that protocol is presented, in addition to the default http protocol. If the user chooses gsiftp download, the Gateway generates a configuration file with relevant endpoints for use with download clients. The configuration file also contains an authorization token that indicates that the user has permissions to download this data.

We evaluated multiple download clients for performance, features, and usability, and chose Data Mover Lite (DML). The tool has been enhanced both in features and in usability based on critical feedback provided from the evaluation. DML now interacts with the MyProxy server, which provides the PKI SSO solution for the ESG federation. The tool allows the user to login to the MyProxy server to download a PKI credential for use for gsiftp download and leverages MyProxy to provision the user's

machine with required trust roots. DML then uses the configuration file downloaded from Gateway to pull down the data using gsiftp protocol.

On the Data Node, GridFTP server installs have been configured to use a custom authorization module developed by the Argonne team that calls back to the Gateway Token Authorization service to validate the authorization tokens provided by the configuration file. In addition to the PKI authentication that establishes the user identity, the authorization step ensures that the data files are protected and only users with granted rights access the data.

2.4 Storage Resource Management (SRM) Data Movement

2.4.1 DataMover-Lite (DML)

A new requirement for DataMover-Lite (DML) that came up during this period was its operation with the authorization model that uses MyProxy. DML version 3 has been developed, integrated with the new ESG authorization model, and tested extensively. Both stand-alone version and web start version have been developed and tested. In addition, a new web-based interface has been developed to simplify the user's access. The screen dump below is an example of that interface.

The steps taken by DML and its interactions with MyProxy and GridFTP are described. An illustration of these interactions is presented at the figure below.

- 1) User browses ESG Gateway portal on https connection, and selects files to download. This generates DML input files with authorization tokens.
- 2) User initiates DML web start and opens the downloaded input file.
- 3) User logs into the ESG Gateway MyProxy server as designated in the DML input file to download a credential and provision the user's client for file transfer.
- 4) Upon successful authentication, DML communicates with GridFTP server for a file download on GSI connection and submits the authorization token for the file download.
- 5) Upon successful authentication at the GridFTP server, its authorization module verifies the authorization token for the file with the ESG Gateway Authorization web service.
- 6) Upon successful authorization, file downloading starts and the downloaded file gets stored at the user's target path.
- 7) DML continues for all files in the request.

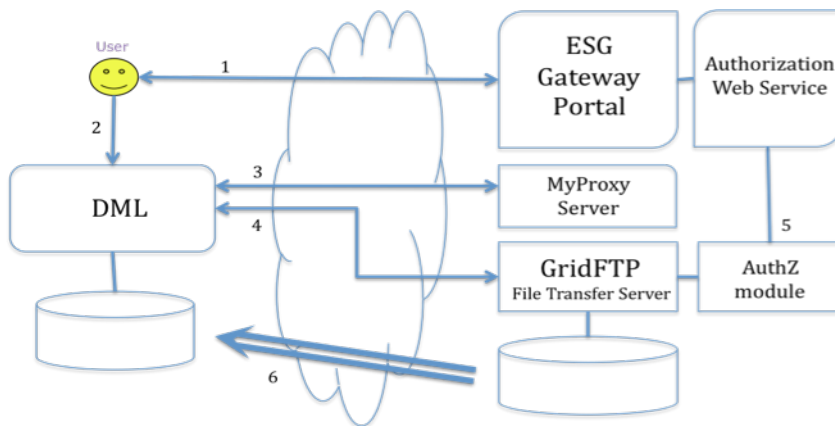


Figure 8: DataMover-Lite file transfer flow.

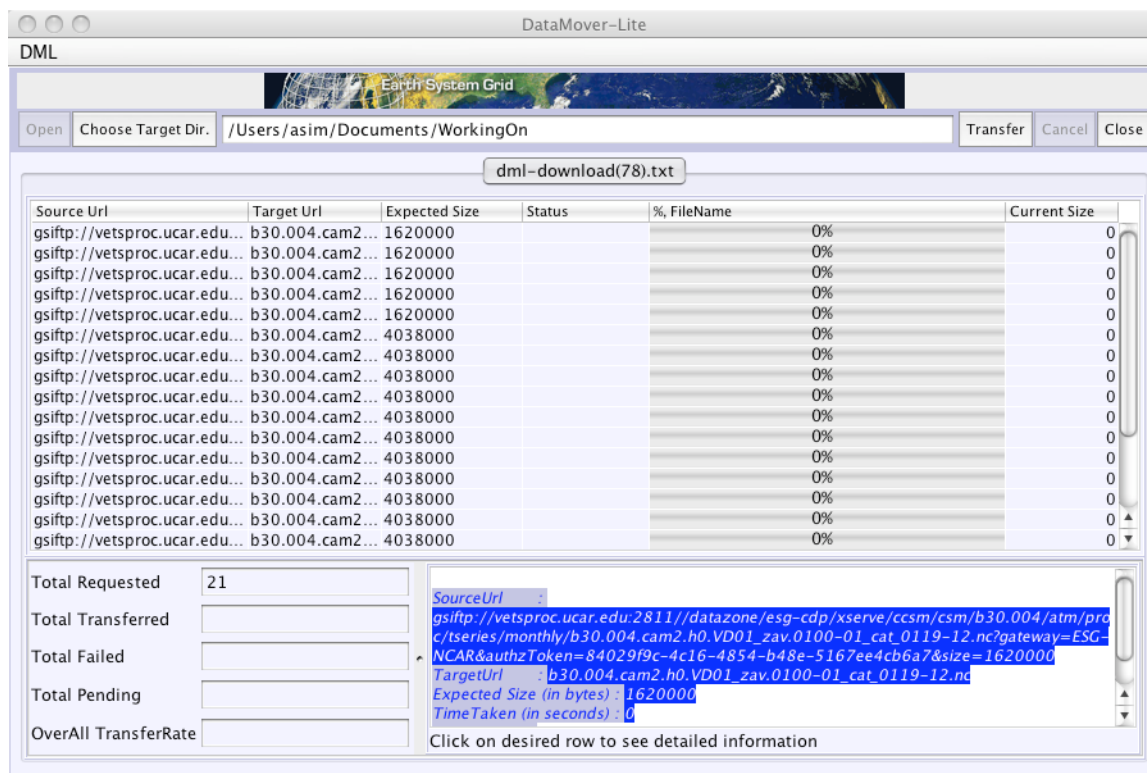


Figure 9: DataMover-Lite GUI interface.

2.4.2 Bulk Data Mover (BDM)

Bulk Data Mover (BDM) has been developed. The first working version will be released in early October 2009. Currently, it is undergoing extensive testing.

BDM takes a remote source directory and local target directory as command line inputs. It proceeds to build a target directory hierarchy based on the source directory structure. Since the requests can be for a large number of files, the information is stored in a relational database internal to BDM. This approach ensures that memory limitations do not interfere with BDM, and also facilitate recovery in case of interruption of long transfer periods.

Once the local (target) directory is built, BDM opens multiple transfer connections, authenticates the client, and starts copying multiple files concurrently through the opened connections. File copying continues until there are no more files to transfer, and then BDM closes the connections. The sequence of steps is described below:

- 1) The user provides the remote source directory and local target directory information to BDM as command line inputs.
- 2) BDM opens the remote GridFTP server, and recursively browses the source directory to find out directory structure and files information.
- 3) BDM save the user request in its local relational database in order to process the execution of large-scale file movement.
- 4) BDM creates the local target directory structure as in the remote source directory.
- 5) BDM opens multiple transfer connections and starts file transfers.

2.5 Data Replication

Substantial progress was made with respect to data mirroring and replication in the current reporting period. This capability was identified during the last year as an important capability for ESG and for collaborators in the UK and Germany who wanted to replicate key IPCC datasets from LLNL/PCMDI to their sites. The goal of this replication is to provide faster access for local scientists and to avoid single points of failure for the datasets.

In the past 6 months, several teleconferences between the ESG, BADC, and Max Plank Institute teams have helped to define the data mirroring use case and to identify issues that still need to be resolved, including how to designate replicas in metadata catalogs; how to discover replicated data sets; and how to update replicated data sets when updates occur or when new simulation files are added to an existing dataset.

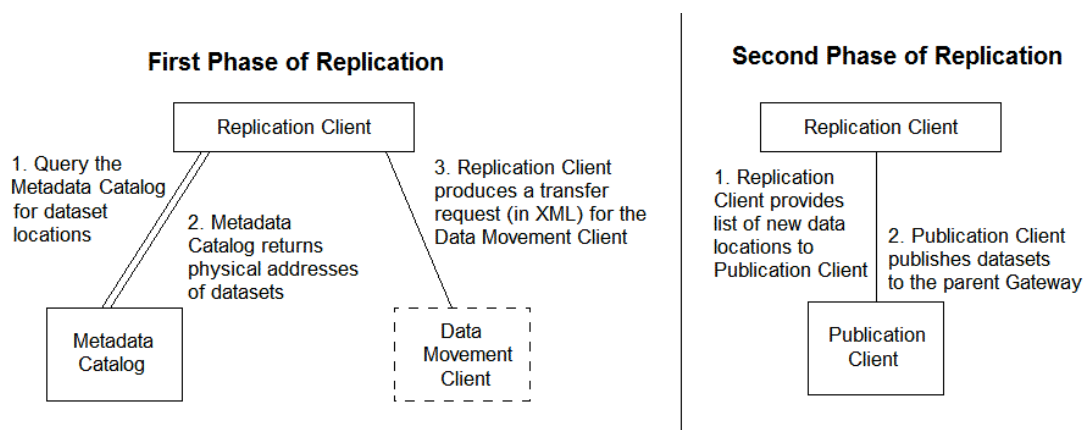


Figure 10: First and second phase of replication in ESG.

The team at ISI has been developing the basic data mirroring functionality, which is illustrated in the figure above. A mirror site submits a request to replicate a copy of a particular data set (or a list of individual files) to the replication client. The replication client then queries the metadata catalog to obtain a complete list of the files that make up this dataset. The replication client must also obtain authorization tokens that allow access to each file. Next, the replication client produces a transfer request for a data movement client that will be responsible for moving the large quantities of data to the mirror site. Our replication client will use the Bulk Data Movement (BDM) Client being developed at LBNL; an initial version of this service is expected in October 2009. Because the mirroring operations may transfer many terabytes of data, the data transfer operations managed by the BDM client may be slow, taking hours or days to complete. Once these transfers are complete, the replication client will provide a list of the newly copied datasets to the publication client, which is responsible for scanning these files, extracting their metadata, publishing their metadata into a local database, and exposing the newly mirrored data sets on the gateway associated with the mirror's data node. Note that the discovery metadata associated with mirrored datasets will indicate that the dataset is a replica rather than the originally published datasets.

Development of this initial capability has required modifications of the existing metadata service to provide individual file names in a data set and the development of an authorization token service.

Initially, we plan to use manual notifications to inform mirror sites when new files in a dataset are available. In the future, we will explore more automated notification systems.

An initial replication capability is planned for release and testing in late 2009. The exact timing will be heavily influenced by the release schedule for the BDM client.

2.6 Product Services

The ESG-CET team continued its Live Access Server (LAS) efforts with the development of tools like “vizGal”, which stands for “visualization gallery”. This tool provides the user with the ability to compare model fields and ensemble members, including visualizations of difference fields. This is a second-generation product for this capability and is built using the Google Web Toolkit. It is very robust, sophisticated, and useful for inter-comparing CMIP5 model output.

Also under way is the next generation geographic region selection tool that is based on the OpenLayers framework. This capability can be plugged into any LAS page or component that requires users to select a region on the globe. It will eventually be used to make selections directly on the plot output of map products.

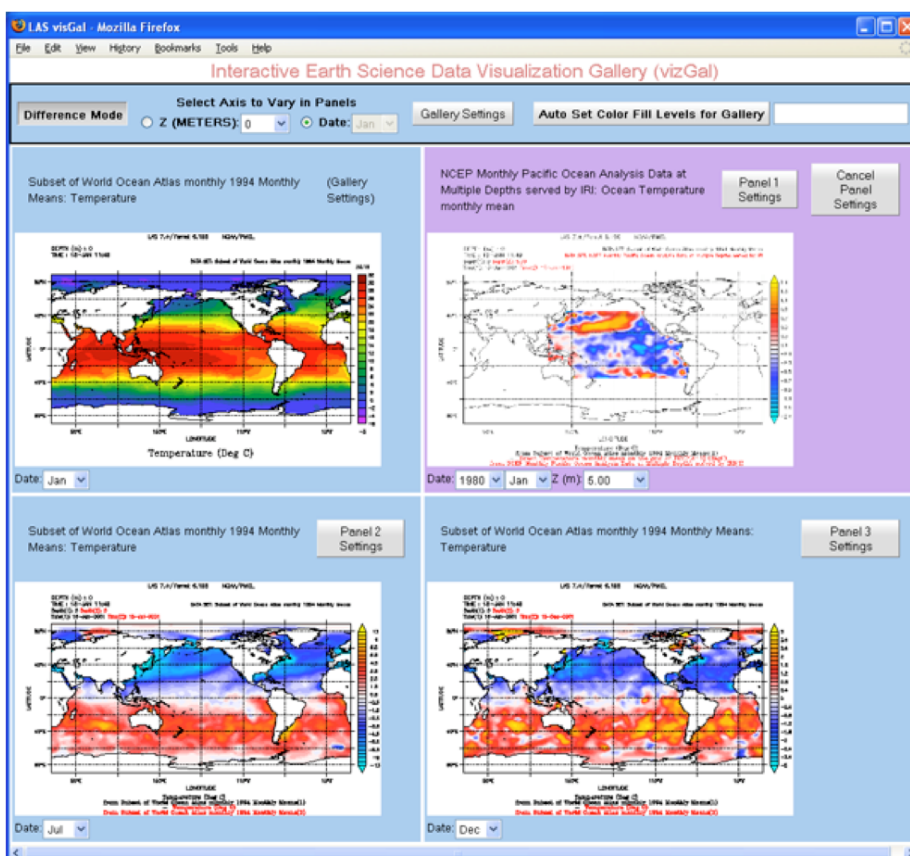


Figure 11: Live Access Server (LAS) visualization gallery (vizGal) intercomparison output.

3 ESG-CET Group Meetings

The ESG-CET executive committee holds weekly conference calls on Mondays at 9:00 a.m. pacific time. At these meetings, priorities and issues are discussed that make up the agenda for the weekly project meetings held on Thursdays at 12:00 p.m. pacific time via the Access Grid (AG). At the AG meetings, the entire team discusses project goals, design and development issues, technologies,

timelines, and milestones. Given the need for more in-depth conversation and examination of work requirements, the following face-to-face meetings were held during this next reporting period.

3.1 ESG-CET Executive Meeting (May 10th)

The ESG-CET executive committee convened in Rockville, MD (the day before the ESG-CET mid-term project review) to discuss the project, the mid-term review, and to set project goals.

3.2 ESG-CET Project Review (May 11 – 12)

The ESG-CET executive committee convened in Rockville, MD for the two-day ESG-CET project review. The executive committee presented an overview of the project, architecture designs, Gateway, Data Nodes, accomplishments, collaborations and partnerships, research and development, future work, and demonstrations. Overall, the review was successful and the team received high marks. (See section 1.2.1 ESG-CET Mid-term Project Review.)

4 Collaborations

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.

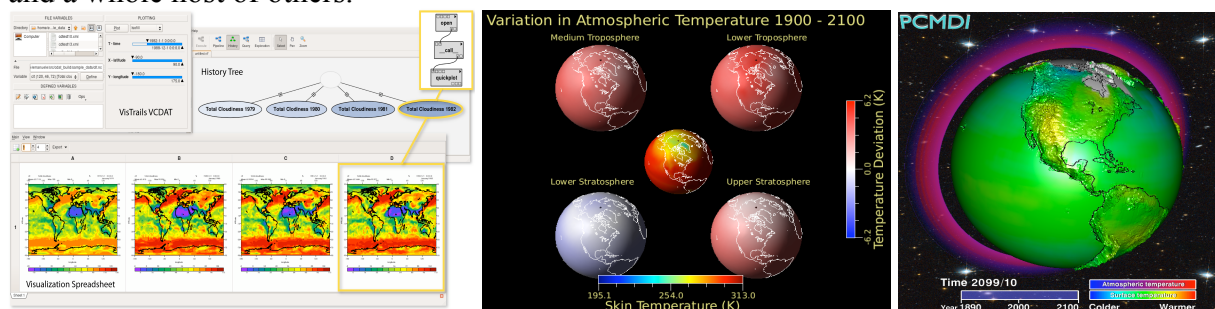


Figure 12: Example of CDAT workflow built inside of VisTrails, and visualization experiments on how best to show the variations of Microwave Sounding Units (MSU) in a three-dimensional display.

5 Outreach, Papers, Presentations and Posters

Outreach activities, papers, talks, and posters presented during this time period:

5.1 Outreaches Activities

5.1.1 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.

5.1.2 *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.

5.1.3 *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.

5.1.4 *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.

5.1.5 *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.

5.1.6 *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.”

5.1.7 *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.

5.1.8 *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.

5.1.9 *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.

5.1.10 *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.

5.1.11 *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.

5.1.12 *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.

5.1.13 *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.

5.1.14 *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 datasets. We have already populated this publicly available catalog with datasets from across several different NOAA line offices, including OAR, NMFS and NESDIS. This collection provides a rich set of data for inter-comparison and verification with the main ESG-CET collections.

5.2 Papers:

5.2.1 *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).

5.2.2 *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.

5.2.3 *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.

5.2.4 *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.

5.2.5 *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.

5.2.6 *IEEE International Conference on Data Mining Paper*

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

5.3 Talks:

5.3.1 *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.

5.3.2 *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.

5.3.3 *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.

5.3.4 *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.

5.3.5 *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.

5.3.6 *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.

5.4 Posters:

5.4.1 *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.