4TH ANNUAL

EARTH SYSTEM GRID FEDERATION AND ULTRASCALE VISUALIZATION CLIMATE DATA ANALYSIS TOOLS

FACE-TO-FACE CONFERENCE REPORT

DECEMBER 2014

A GLOBAL CONSORTIUM OF GOVERNMENT AGENCIES, EDUCATIONAL INSTITUTIONS, AND COMPANIES DEDICATED TO DELIVERING ROBUST DISTRIBUTED DATA, COMPUTING LIBRARIES, APPLICATIONS, AND COMPUTATIONAL PLATFORMS FOR THE NOVEL EXAMINATION OF EXTREAME-SCALE SCIENTIFIC DATA.

LLNL-TR-666753

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

The work was also undertaken with other collaborating international government funded organizations (namely from Australia, Germany, UK, France, China, Japan, Netherlands, Italy) and other US government agencies (NASA and NOAA).

Contents

1 Abstract

The climate and weather data science technology infrastructure community met December 9–11, 2014, in Livermore, California, for the fourth annual Earth System Grid Federation (ESGF) and Ultrascale Visualization Climate Data Analysis Tools (UV-CDAT) Face-to-Face (F2F) Conference. The conference was hosted by Lawrence Livermore National Laboratory with support from the Department of Energy, National Aeronautics and Space Administration, National Oceanic and Atmospheric Administration, the European Infrastructure for the European Network of Earth System Modelling, and the Australian National Computational Infrastructure. Both ESGF and UV-CDAT remain global collaborations committed to developing a new generation of open-source software infrastructure that provides distributed access and analysis to simulated and observed data from the climate and weather communities. The infrastructure, tools and methods being developed under these international multi-agency collaborations are critical to understanding extreme weather conditions and long-term climate change.

The F2F conference fosters a stronger climate and weather data science community and facilitates a stronger federated software infrastructure. Further, scientists supported by the agencies that now participate in the F2F represent most of the major centers that manage the major internationally significant environmental data collections and their associated infrastructure for analysis, and bring their considerable expertise to the conference. The 2014 F2F conference detailed the progress of ESGF, UV-CDAT, and other community efforts over the year and sets new priorities and requirements for existing and impending national and international community projects, such as the Coupled Model Intercomparison Project Phase 6, and the newly established DOE Accelerated Climate Modeling for Energy project. Specifically discussed at the conference were project capabilities and enhancements needs for data distribution, analysis, visualization, hardware and international network infrastructure, standards, and resources.

2 Executive Summary

From individual researchers to modeling centers, those working with climate data sets of any size are faced with many technological changes and integration issues. These challenges were a source of rich discussion during the international 2014 Earth System Grid Federation (ESGF) and Ultrascale Visualization Climate Data Analysis Tools (UV-CDAT) Conference, held December 9–11, 2014, in Livermore, California, USA. This report summarizes the conclusions reached during the conference, including the impact of current technology developments and improvements since the last conference, held in 2013, and the need for new and improved technologies to meet community requirements.

The report draws on the years of collective experience from the community of hundreds of climate, computational and data scientists, and high-performance computing (HPC) technology specialists working together to improve climate research by capturing leading technologies from the in-depth presentations and discussions. These forward thinkers, as well as senior executives, represent varying stages of the data ecosystem, as shown in **Figure 1**. The report is meant to function as a guide for future developments. In addition, it will help the community track and monitor the evolution of the international climate data ecosystem.

An international group of government agencies, including the U.S. Department of Energy (DOE), European Union Commissions, Australian National Computational Infrastructure (NCI), and Asian government universities and research centers, provide the funding for the applied research and development activities shown in **Figure 1**. The funding includes work on the development of the international open-source federated data archive software stack that joins the decentralized multi-petabyte archive, local and remote client applications for searching and accessing data, software installation at modeling and data centers, hardware resources for remote data reduction and manipulation, international network resources for large-scale data movement and replication between centers, analysis and visualization tools for diagnostics and model metrics, and community resource pooling for technical and science support. These activities represent only a fraction of the work underway.

In addition to individual software, hardware, and network development, a portion of funding feeds into building and strengthening collaborations to enable the integration and leveraging of all existing and new tools. By working together towards a common strategic plan and mission goals, the funding agencies support highly visible scientific community projects with the data dissemination tools and resources needed to advance climate and weather research. Without these funding agencies coming together to support the data ecosystem, none of the work in **Figure 1** would be possible.

The growing international interest in ESGF and UV-CDAT development efforts has attracted many others who want to make their data more widely available and easy to use. For example, the World Climate Research Program, which provides governance for CMIP, has now endorsed components of the software foundation to be used for ~70 other model intercomparison projects (MIPs), such as obs4MIPs, TAMIP, CFMIP, and GeoMIP. At present, more than 40 projects are represented.

To accommodate the needs of a growing community, major centers and projects in the U.S., U.K., Germany, Australia, Japan, and a number of other countries have implemented mandatory or voluntary community software-supporting standards defined by the community. Each project is responsible for supporting its own community of users. This way, the community benefits from the wide range of services and activities that the software foundation provides to improve science in communities and increase access to data. Through the software alliance, governed under the worldwide multi-agency consortium, the team has developed an operational system for serving climate data from multiple locations and sources.

The multi-agency sponsors base the international foundation of community software on hundreds of decentralized community developers. In all cases, the software that is produced is open source and freely available. The software, supported by one or more funding agencies, is developed collaboratively through a consensus-based development process using Agile software development methods. The community foundation software products include:

- ESGF, for worldwide data management and access;
- CoG, a content management system and wiki for scientific projects;
- UV-CDAT, visualization and analysis tools for ultrascale climate data;
- Climate and Forecast (CF) convention, a definitive description of the data each variable represents along with the spatial and temporal properties of the data;
- Climate Model Output Rewriter, which produces fully CF-compliant netCDF files;
- Earth System Documentation (ES-DOC), metadata standards used to describe Earth system models;
- Globus, for secure file transfers; and
- Live Access Server, for visualization, analysis and subsetting of climate and reference data sets.

With these and many other software libraries, packages, and sub-components, the integrated infrastructure enables real-time comparison of model output to observational measurements in a controlled environment, thus eliminating tedious activities associated with climate research.

In addition to software, the partner agencies are making significant investments in developing hardware, network services, and resources. As one of the goals of this working consortium, we are committed to ensuring that all services and resources, including software, hardware, and networks, are easy to use and have the necessary user documentation. The partner agencies also manage directly, or engage other critical infrastructures such as data centers supporting hardware, tertiary storage, cloud and compute clusters, high-performance computers, and multi-agency network providers, such as ESnet, Internet2, JANET, SurfNet, DFN, and AARNet. By coordinating our efforts with existing and upcoming hardware and network infrastructures, we will empower climate projects and the climate data community with the best possible platform for scientific advancements. This is evident in the existing data ecosystem, where more scientific papers have been published more recently than ever before.

The data ecosystem, depicted below in **Figure 1**, indicates how data are generated, stored, documented, managed, manipulated, and made available for intended use. That is, the data and information are processed through each component, persistent provenance and workflows are captured for documentation and reproducibility/repeatability, and information facilitating knowledge discovery is recorded. The ecosystem components described in **Figure 1** are:

- ***Critical Complex Data Generating Systems*** that generate petabytes of data from sophisticated technology sources, ranging from high-end supercomputers, clusters, and computer servers to sensitive environmental detectors, lab analyses, and orbiting satellites;
- ***Data Collection and Management***, which collects, stores, and organizes data for easy user discovery and accessibility;
- ***Data Analytics*** for pattern discovery, structure identification, dimension reduction, image processing, machine learning, and exploratory visualization;
- ***Data-Intensive Computing*** for describing applications that are input/output bound and enabling large and complex data manipulations; and
- ***Decision and Control*** for decision control and knowledge discovery.

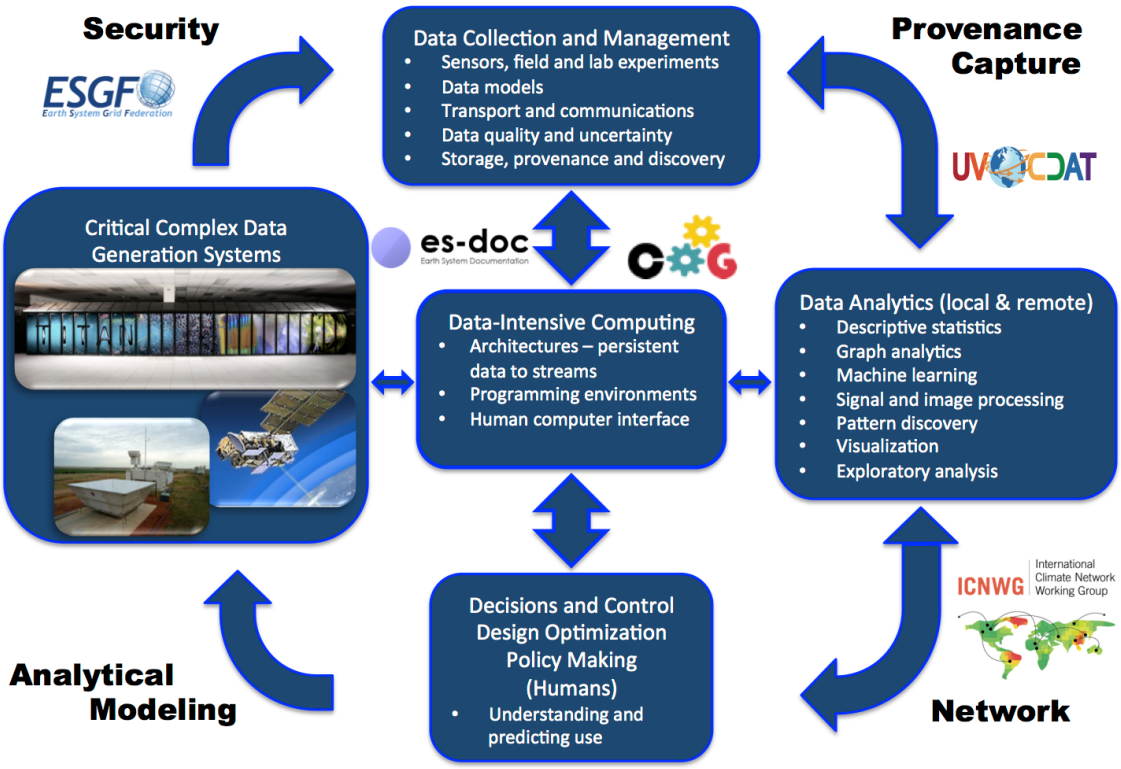

Figure 1. *The diagram depicts the many components for the data ecosystem, where provenance capture is pervasive throughout. Data from the "Critical Complex Data Generation Systems" are housed and securely managed at many worldwide sites with the ESGF software stack. Local and remote computation is necessary, as the increasing data size and algorithm complexity is leading to more data-intensive and compute-intensive user requests. For data backup with easier data access, the network must be able to move petabytes of data between data centers. Finally, analytical modeling assists users in making smart choices in using community resources for moving and computing large-scale data.*

3 ESGF Governance

Most of the software, hardware, and network components that make up the data ecosystem possess a governing board. The sheer number of governing boards across the international agencies involved makes process management and community coordination quite challenging. Therefore, before beginning the process of bringing all the disparate components together, it is important to understand how each of the governance processes works at a conceptual level. Fortunately, at the conceptual level, most governing bodies behave in similar ways.

A major topic of the conference was ESGF governance. As shown in **Figure 2**, the conceptual ESGF governance structure generally consists of a steering committee of sponsors, an executive committee of experts, and project teams. These committees are normally tied to one or more projects in the community with a specific mission or goal in mind. This governance model drives innovation and quality of services. Moreover, the governance model helps to balance the conflict that exists between new development and day-to-day operations. Additional challenges under governance involve developing project roles in the community that help to define relative tasks and responsibilities that are not expressed clearly enough for implementation purposes.

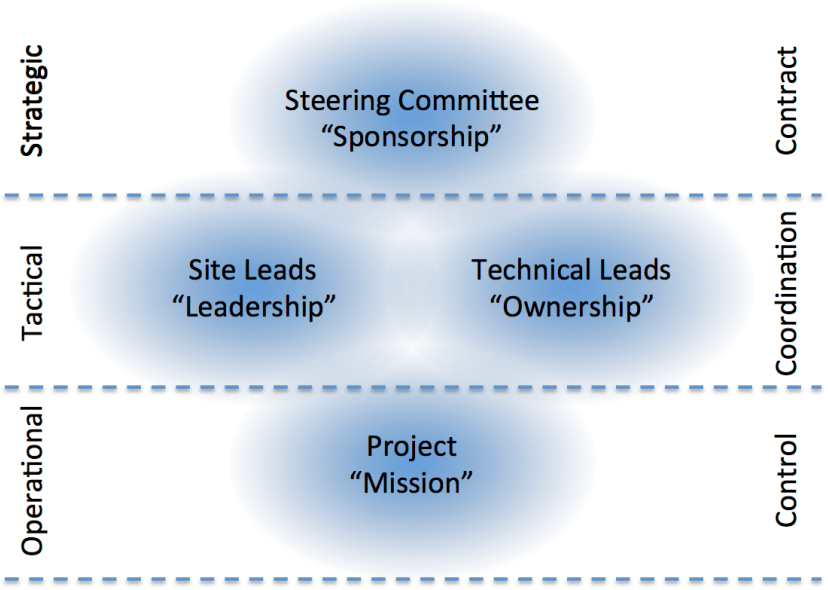

Figure 2. *Common process governance structure of most data ecosystem components.*

For the data ecosystem, there are many aspects of governance that must be considered for its successful execution, all of which are equally important:

- Sponsorships and funding opportunities—coordinated financial support from international agencies to complete a specific task or assignment needed by the community in support of research activities, the end result generally leads to an open-source product for community consumption.
- Organizational structure and goals—shapes and defines project leadership and is the foundation for guidance, showing competence in reaching project goals and milestones efficiently and effectively.
- Decision-making model or process—process resulting in the selection or direction of a component or process, generally done as a collective process involving key decision makers and/or key developers.
- Management model—how we conduct or govern ourselves in formulating project values, strategies, transitions, and shortcomings; coordinating activities; allocating resources, etc.
- Implementation phases—involves product installation, user training and documentation, and system documentation.
- Roles and tasks (associated with budget)—determines appropriate leads and schedules for tasks.
- Rewards—motivation for job completion and overall project success.
- Component lifecycle—the distinct stages between when the components are created and when they are destroyed (or refreshed with the next evolution of technology).
- Monitoring and control—usage tracking and monitoring of the entire ecosystem by users, administrators, and sponsors.
- Performance evaluation—analysis of software and hardware and network equipment performance.
- Reporting—reporting to sponsors, stakeholders, and team members an account of the state of the enterprise and upcoming events (e.g., this report).

In recent years, the need to direct and organize process management across the ESGF infrastructure has grown as the importance of the data ecosystem has grown. However, efforts to achieve change and improvement must be conducted in keeping with clear definitions of the various supported projects; otherwise, introduction of initiatives and actions to restructure processes may not produce the expected results, and overall success will not be guaranteed. It has therefore been vital to make progress toward a well-defined process governance model, aligned with the sponsor strategies and priorities, that supports the overall ecosystem development, and improves the quality of the operations. This will permit better coordination and communication among process initiatives by formalizing roles, responsibilities, structures, and also metrics by which we can be measured and improved.

As stated in the ESGF governance document, a number of committees have recently been established, represented in Figure 3. The make-up and responsibilities of these committee's are expected to be further developed and mature over time, and the following represents the current working model.

The ***Steering Committee*** is composed of the major ESGF sponsors that oversight development and operations that align with priorities of major scientific projects, the overall strategic development of the infrastructure for the future, and develop of the governance itself – including its relationship with other governance bodies that closely interact with the ESGF infrastructure or the range of funding dimensions to the ESGF. The Steering Committee oversights and reviews the operational and technical development activities of the ESGF that are carried out by an ***Executive Committee***.

The ***Executive Committee*** is a broader representative group that has been established to closely consider software development decisions that are consistent with the needs of the funding agencies' projects, gather community requirements, set software directions, coordinate operational activities, and keep stakeholders (i.e., **Steering Committee**, fellow PIs, and the climate research community) informed of progress. The ***Executive Committee*** will bring matters to the Steering Committee to table the considerations of individual or centralized development, and operational priorities and how they are being managed – which includes the deliberation of how best to gather and execute the community requirements. Importantly, the ***Executive Committee*** is tasked to make sure mission-critical tasks are completed.

Importantly, the ***Executive Committee*** interfaces with the end-user community. Committee members have a balanced insight into the scientific requirements, emerging technologies, and strategies that must work across ALL agencies. They do this by engaging with the sponsors and sponsor-funded projects; holding Webinar meetings, face-to-face (F2F) conferences, and teleconferences; exchanging information via mailing lists and project websites; and combining these technologies to ascertain user feedback, which will then be used to strengthen the ESGF software end product. The current ESGF team is experienced and has worked together successfully to meet the communities' data needs for a long time. The ESGF software effort also benefits from advisory panels. Requirements are highlighted by the advisory panels and articulated to funding agencies. For example, there was a panel composed of sponsors and community project leads at the 2014 ESGF F2F conference to provide input and harden community requirements.

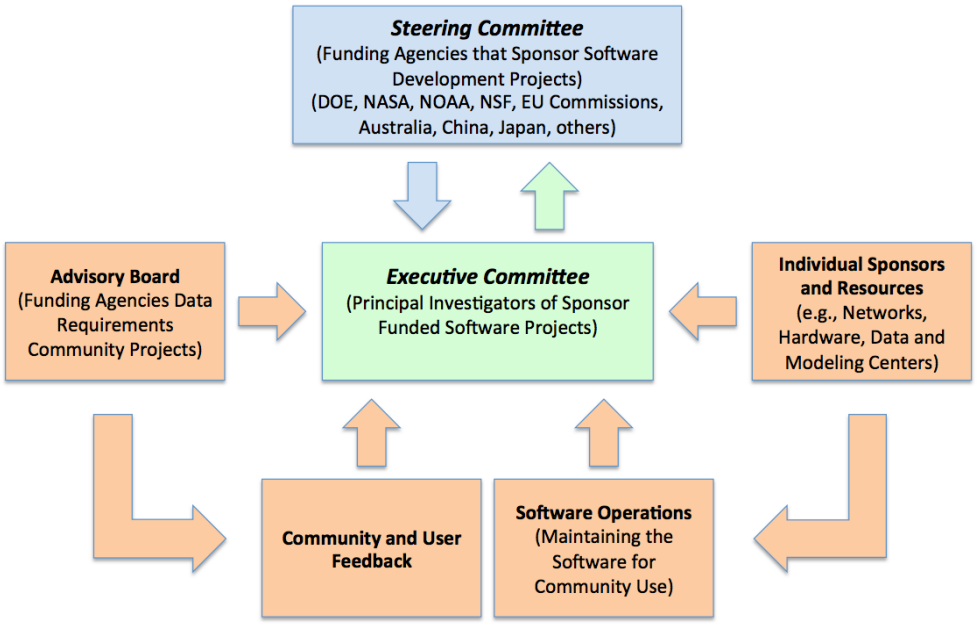

Figure 3. *High-level organizational chart showing the line of communication between the Steering Committee (i.e., international funding agencies and sponsors) and their sponsored PIs, representing the Executive Committee. In keeping this fluid, key advisory panels may be established. Therefore, this figure is only indicative of the process.*

For example, there may be a single panel just for CMIP6 (e.g., the Working Group on Coupled Modeling [WGCM] Infrastructure Panel [WIP]).

Both committees work closely together with supported projects and non-funded organizations to define and implement new technologies for the scientific community, especially those in the areas of data management, distributed computing, networking, analysis, and visualization. Through its inclusive structure, the ***Executive Committee*** maintains close connections to international efforts sponsored by the ***Steering Committee*** (i.e., DOE, National Aeronautics and Space Administration [NASA], National Oceanic and Atmospheric Administration [NOAA], National Science Foundation, European Union [EU], Australia, China, Japan, and others) in the areas of climate and data science. The ***Steering Committee*** relationship with the ***Executive Committee*** offers a unique opportunity to leverage major programs/projects across multiple international funding agencies in the service of community science research.

The full ESGF governance document can be found on the ESGF site: http://esgf.llnl.gov.

4 Science Drivers: Projects' Use Cases

Climate and Weather science is a prominent example of a discipline in which scientific progress is critically dependent on the availability of a reliable infrastructure for managing and accessing of large quantities of heterogeneous data on a global scale. It is an inherently collaborative and multi-disciplinary effort that requires sophisticated modeling of the physical processes and exchange mechanisms between multiple Earth realms (atmosphere, land, ocean, and sea ice) and comparison and validation of these simulations with observational data from various sources, possibly collected over long periods of time.

The 2014 F2F conference focused on several key data sets that are important to community researchers and on programmatic goals that span many different funding agencies. In this section, we describe a few funding agency use cases in more detail to illustrate the challenges associated with accessing and analyzing data from disparate projects. That is, our goal is to address the particular requirements and ultra-scale capabilities needed to access, analyze, and visualize large data sets that are responsible for helping scientists and policy makers understand climate change. Several sets of archives in particular will provide focus for our proposed work, namely:

- The Coupled Model Intercomparison Project Phase 6 (CMIP6) and other multi-model intercomparison data sets, including the NASA-sponsored satellite observational data intercomparisons;
- The Coordinated Regional Climate Downscaling Experiment (CORDEX) for projections of how the climate of the Earth may change regionally in the future;
- The Accelerated Climate Modeling for Energy (ACME) as it exploits advanced software and new HPC machines in the development and application of a fully coupled, state-of-the-science Earth system model for scientific and energy applications; and
- The interaction of data sets from various domains on which climate may have an effect.

Although the ESGF and UV-CDAT conference received more than 20 project abstracts, only 15 were selected for conference presentation due to time constraints. Of those 15, only a few are described below as sample use cases.

4.1.1 Use Case 1: CMIP and obs4MIPs

The climate community has worked for the past decade on concerted, worldwide modeling activities led by the Working Group on Coupled Modeling (WGCM), sponsored by the World Climate Research Program (WCRP), and leading to successive reports by Intergovernmental Panel on Climate Change (IPCC). The fifth assessment (IPCC-AR5), released in September 2013, was the latest report. These activities involve tens of modeling groups in as many countries, running the same prescribed set of climate change scenarios on the most advanced super-computers and producing several petabytes (PB = 10^{15} bytes) of output containing hundreds of physical variables spanning tens and hundreds of years. These data sets are held at distributed locations around the globe, and must

be discovered, downloaded, and analyzed as if they were stored in a single archive, with efficient and reliable access mechanisms that can span political and institutional boundaries.

The same infrastructure must also allow scientists to access and compare observational data sets from multiple sources, including, for example, Earth Observing System (EOS) satellites such as those found in obs4MIPs. These observations, often collected and made available in real-time or near real-time, are typically stored in different formats and must be post-processed to be converted to a format (e.g., netCDF-CF) that allows easy comparison with models. The need for providing data products on demand, as well as value-added products, adds another dimension to the capability demands. Finally, science results must be applied at multiple scales (global, regional, and local) and made available to different communities (scientists, policy makers, instructors, farmers, and industry).

Because of its high visibility and direct impact on political decisions that govern human activities, the end-to-end scientific investigation must be completely transparent, collaborative, and reproducible. Scientists must be given the environment and tools for exchanging ideas and verifying results with colleagues in opposite time zones, investigating metadata, tracking provenance, annotating results, and collaborating in developing analysis applications and algorithms. This virtual collaboration environment that facilitates and advances scientific discovery is precisely the data ecosystem environment that was presented at the ESGF and UV-CDAT F2F conference.

4.1.2 Use Case 2: CORDEX

Regional climate downscaling (RCD) techniques, which include both dynamic and statistical approaches, are being increasingly used to provide higher-resolution climate information than is available directly from contemporary global climate models (GCMs). The techniques available, their applications, and the community using them are broad and varied, and it is a growing area. It is important, however, that these techniques and the results they produce be applied appropriately and that their strengths and weaknesses are understood. This requires a better evaluation and quantification of the performance of the different techniques for application to specific problems. A coordinated, international effort to objectively assess and compare various RCD techniques, built on experience gained in the global modeling community, is providing a means to evaluate RCD performance, to illustrate benefits and shortcomings of different approaches, and to provide a more solid scientific basis for impact assessments and other uses of downscaled climate information.

WCRP views regional downscaling as both an important research topic and an opportunity to engage a broader community of climate scientists in its activities. CORDEX has served as a catalyst for achieving this goal.

One of the CMIP successes has been in getting data out to the community in a coordinated manner, using a single and documented format and file structure. It has been decided that CORDEX will use the same ESGF infrastructure as CMIP. The same facility now exists for CORDEX data. The IS-ENES2 community took responsibility for implementing several adjustments to the process:

- DRS has been adapted for dynamical downscaling;
- Attribute service (data access and term of use) is operated by LIU;
- Versioning of data sets has been done at the variable level; and
- Quality control (QC; by DKRZ) has been done prior to the publication.

Ensuring that CORDEX data is evaluated and used in regions throughout the world requires the maintenance of regional training and capacity-building programs. Several extremely important regional CORDEX workshops and training and outreach activities have been held in between 2011 and 2014. The Africa-CORDEX training program has (so far) spurred the generation of a large amount of CORDEX data available for Africa and the creation of at least three peer-reviewed articles, led by three regional groups of African scientists, using CORDEX data. The most recent workshop, an Asian ESGF training workshop, was organized by the Monsoon Asia Integrated Regional Study in China and held in December 2014. Nicolas Carenton (Institut Pierre Simon Laplace [IPSL]) and Michael Kolax (Swedish Meteorological and Hydrological Institute [SMHI]) were the primary speakers for the two-day training workshop.

4.1.3 Use Case 4: ACME

Based on their workflow requirements, general use cases within DOE-ACME can be separated into three distinct categories. The first is the process for developing a new capability within the model, which requires many small runs with rapid turnaround of the workflow steps, significant interaction with software tools, and automated testing and version control. The structure of the output varies and needs to be easily accessible through short-term, local archives. Plotting and analysis need to be more interactive, nimble, and extensible for the user as development proceeds.

Secondly, exploratory use-cases and their workflows involve numerous and varied length and spatial-scale simulations with single or multiple components activated, potentially using ensembles for uncertainty quantification and optimization to explore parameter space and model fidelity. Output is shared within small groups of project scientists both locally and externally using short- to medium-term archiving. Interactive, web-based, visualization tools are required to incorporate HPC information that is especially useful at this stage for diagnosing issues before full production runs begin. Provenance is also necessary to record testing and evaluation steps required for paper and data publishing in development-focused journals.

Thirdly, production runs of the model comprise the most substantial and diverse set of use cases. Collections of ensembles are performed over months and may be transferred to multiple staff as they proceed. Large jobs are queued on U.S. Leadership Computing Facility systems where large data sets are created, and complete provenance and archiving infrastructure is required for data publishing to other collaborators and eventual public release.

4.1.4 Use Case 4: Diversity of Use Cases

The diversity among and within use cases necessitates a flexible infrastructure as well as a strong collaboration with other domains and computational scientists to embed the needed tools within the data ecosystem (i.e. getting scientists to use the software and provide feedback for ongoing development). The list of project feedback is daunting, but community developments (described in Section 6) will enable more interaction with projects and users and be adapted for many different use cases. The following use case illustrates the interface of the researcher/user with the data in order either to perform scientific research or to understand environmental concerns relevant for setting policy. The processing effort involves moving vast amounts of data (spanning several government agency analysis centers) to and from various ESGF sites around the world.

Many diseases have strong correlations with short and long-term weather patterns, with causal relationships that can be as simple as cold noses being more susceptible to the common cold, or as complex as the intricate interactions between weather, such as migratory animal populations, pathogens carried via airborne vectors, acorn crops, acorn-eating mice populations and their ticks that spread Lyme disease to neighboring human populations. Regardless of the complexity of the underlying mechanism, the first step towards gaining a predictive understanding of these mechanisms is to provide a common platform where disease outbreak data can be integrated with observed and interpolated climate data sets. Existing epidemiological databases record historical case numbers for dozens of important diseases over time across geographic regions. For example, the Tycho database (www.tycho.pitt.edu) tracks weekly numbers of cases and deaths for more than 50 diseases across U.S. states and cities, in some cases going back to 1888. This data, along with key demographic data on population density and composition, vaccination rates, and health expenditures, can be readily imported and made available through ESGF. Domain experts in climate science and epidemiology alike can then search this data for previously unknown climate drivers, visualize these relationships using UV-CDAT, and develop predictive models for individual disease, in an international collaborative effort that may rival the current climate modeling community.

Public health experts to predict real-time risks of specific disease outbreaks based on recent weather patterns can use such models—essentially a weather-driven disease forecast system. Climate scientists will also be able to use such models to predict the spread of tropical diseases due to global climate change, as is currently the case with Dengue and Chikungunya virus spreading into the Southern U.S. Working with the scientists, policymakers can

use these short and long-term disease predictions to drive allocation of funds into prevention and vaccination campaigns, treatment centers, emergency response preparedness, and research on diseases that pose novel risks.

5 Feedback from Projects and Their Requirements

5.1 Summary of Requirements

As discussed in the previous sections, governance and use cases are real issues that determine how requirements affect different aspects of operations and software development as they relate to projects and data. Therefore, encouraged by many supporting funding agencies, a significant fraction of projects utilizing ESGF and UV-CDAT to disseminate and analyze data attended the conference to voice their concerns and requirements for current and future community data use cases. At the forefront were discussions centered on maintaining essential operations and the development of new and improved software to handle ever-increasing data variety and complexity, velocity, and volume. This section is the summary of requirements for computing and data sciences' activities critical for the community to meet its scientific mission, both as individual projects and as a federation of collective projects.

Results of the analyzed requirements from each project and as a collective are already under way by many of the ESGF, UV-CDAT, and collaborating sub-working teams that are already engaged in high-priority innovations to sustain the climate communities computational and data activities. Although, there are many projects utilizing the data ecosystem (e.g., ESGF disseminates data for more than 40 well-known climate projects), only a few projects were able to provide feedback and requirements, due to time constraints: CMIP6, WIP, NASA's satellite and weather program, CREATE-IP, obs4MIPs, CORDEX, climate4impact, NCPP, NMME, HIWPP, ACME, and GeoMIP. With this limitation and not in any particular order of priority, the summary requirement findings revealed the need for:

- ***Governance*** as a high-priority issue. The community would like to see a more formal agreement of software expectations and requirements and clearly defined membership roles. In addition, other governance bodies are required for user support, Internet network coordination, the projects themselves, critical operations, and external software coordination. For better communication, future requirements and governance issues must be available to everyone. This is to include road maps, timelines, meeting minutes, milestones, deliverables, roles and responsibilities, and anything else that facilitates better communication of governance.

  The ESGF node software stack must also be developed further to better co-exist with other systems to execute on its charter to support a wider variety of priority communities and needs. The governing bodies will help to recognize the people sponsoring development activities. Although we are recognizing the federation, it is important to allow autonomous processes independent of the international federated agencies. The current model is very good for nurturing the current mission and community. By thinking longer term, we need to figure out how to help broader communities engage and possibly look at hybrid governance models. For example, the Earth Science Information Partnership (ESIP) allows agencies to buy-in but have less direct influence on projects or their direction.

- ***Provenance*** capturing is needed throughout the end-to-end process of the data ecosystem described in **Figure 1**. Data provenance challenges with large-scale heterogeneous data in climate research include efficiently analyzing captured provenance, capturing cause-and-effect provenance, capturing simulation runtime provenance, capturing experimental and design provenance, capturing hardware and compute provenance, and capturing data movement provenance, to name a few. The integration of all these provenance captures will be challenging in itself and will need to be extended to all development areas of the data ecosystem, such as publishing, version control, data exchanges, computing, etc.—all over distributed nodes and heterogeneous research environments.

Fortunately, data provenance has been studied over the years and there are many conventional provenance techniques focused on determining how a given data set came into being (e.g., what machine processed the data using which model and owned by whom). For climate research, particular interest has been placed on identifying which data causes changes in the simulation, the status of an executing simulation or analysis for derived variables, and provenance of simulation models. For observational data sets, provenance will be applied when feasible. Therefore, the data provenance for our data ecosystem will focus on provenance of simulation results, simulation execution, the simulation model, and model intercomparison data used for post analysis. In addition, in the federated environment, we must support distributed provenance capture and provenance queries and searches.

- ***Controlled vocabulary*** is a set of words, phrases, or terms that are acceptable values for completing certain metadata fields. For consistency across the many projects, a panel to define the standard definitions for each term should be in place to govern the defining process. Controlled vocabulary services must be established to provided site-wide, nation-wide, and federation-wide consistency. Vocabularies are needed for many other components of data ecosystem such as Data Reference Syntax (DRS). That is, components such as DRS make use of controlled vocabularies to facilitate documentation and discovery. For the CMIP project, the controlled vocabulary is used to help develop category-based data discovery services. Many of the other projects have also stated a need to use controlled vocabularies for consistent data discovery and documentation across the federation. Finally, controlled vocabulary will complete a search relatively effectively and help to eliminate returning extraneous information.

- ***Operations*** are recognized by the community as key for sustaining successful data dissemination to a community of disparate community projects. It is also recognized in the decision-making process to refresh component software and hardware technologies. Often, operations are lost in the development process and not funded by agencies that are interested in organizational science. However, the fact is, without operations folded into the development process and into the long-term science project, sustainable long-term science would be problematic. All aspects of operations are important to the science community, such as maintaining provenance, user support, documentation, software releases, and overall performance. Nearly every component of the data ecosystem must have sustained operations folded into its development.

  A part of sustaining the operations of the data ecosystem is sustaining the consistency of the operational (or production) data sets. Scaling ESGF technology to handle large-scale data sets consistently and throughout the federation is an operational challenge that will require all supporting funding agencies to consider putting into software and hardware development. As a suggestion, the governance board could empower an operations team to bring development and operations closer together and to communicate more effectively to manage the data ecosystem production applications.

- ***Server-side (remote) computing*** is increasingly becoming a critical operation as rapid advances in technology, storage capacity, computing capability, and the size of data sets becomes too large for practical data movement. Emerging collaborative data systems in the community, such as ESGF, provides petabytes of data and is on a trajectory to provide tens of petabytes of data in the next few years. For such large data repositories, it is necessary to provide product and data manipulation services where the data is co-located to reduce data movement.

  For ease of use, flexible, intuitive and highly interactive web-based interfaces are needed to not only interface with the data but to also interface with remote high-performance computing resource (e.g., clouds, Hadoop, and Linux clusters) and underlying data manipulation software. As discussed at the F2F conference, this software-as-a-service model will better facilitate direct access to the underlying data, whereby specific software tool such as UV-CDAT would manipulate the data and return an OPeNDAP product that popular analysis tools (e.g., Ferret, NCL, GrADS, UV-CDAT) could further the data manipulation process for the end user. Security, handled through ESGF, would authorize hardware, software, and data access through the analysis tools, which are OPeNDAP-enabled. This strategy,

discussed by the ESGF compute working team, obviates the need to download massive multi-file data collections for an analysis that may require only a small regional subset of the full data set.

- ***High-speed data network*** connections are needed to perform large-scale data transfers within a reasonable time between primary climate data and modeling centers. Specifically for CMIP6, tens of petabytes of data will need to be replicated among various data centers internationally. To help meet this requirement, we would propose a few different tools and goals:

    1. The adoption of a "tier" system for describing the ESGF data centers involved with particular projects. Tier 1 data centers would be expected to support high-speed data replication to other Tier 1 data centers. Tier 2 centers would have lower performance requirements but would be serving more local science research, and so on. Currently, the Tier 1 data centers are Australian National University (ANU), the British Atmospheric Data Centre (BADC), the German Climate Computing Center (DKRZ), and Lawrence Livermore National Laboratory (LLNL). In the future, perhaps the Beijing Normal University (BNU) in China and the University of Tokyo in Japan will be added to the list or Tier 1 climate data centers.
    2. GridFTP (or its cloud-based derivative Globus Transfer) would be the suggested transfer protocol to implement at the Tier 1 sites, since the existing HTTP-based tools used for ESGF will not scale to the data rates required for petabyte-scale data replication.
    3. A firm goal of 16Gbps (2 GB/sec) by 2016 for data transfers between Tier 1 sites, just before the full production of CMIP6. This will allow the replication or transfer of 2.5 petabytes of data in one week between two designated sites.

    The performance engineering activities necessary to achieve the high levels of performance are being coordinated under the auspices of the International Climate Network Working Group (ICNWG). Formed under ESGF, the ICNWG is a coordinated, multi-agency, international collaboration of institutional and backbone network providers dedicated to the adoption of networking best practices in support of the study of climate change. In addition to the institutional networking organizations that support the modeling and data centers, the network organizations involved in ICNWG include networks in Australia (AARnet), Germany (DFN), the Netherlands (SURFnet), the U.K. (Janet), Europe as a whole (GÉANT), and the U.S. (ESnet and Internet2).

- ***Publication workflow*** represents the data and metadata that are published, stored, and served from ESGF nodes located at the data modeling centers and yet are searchable and accessible as if they were stored in a single archive. In order to achieve satisfactory performance in publication as a service, metadata holdings at each site need to be partitioned and published into high-level, frequently queried discovery information. Remote published queries issued to any node are to be automatically distributed to all other nodes in the federation, so that complete, up-to-date searchable and discoverable results are always returned to the user. Additionally, automatic and incremental metadata publication technologies are needed to create identical, synchronized local copies of remote metadata catalogs, to speed up query performance over large geographic distances.

    Quality control must be done prior to publication and file attributes should be checked before publication. Getting the data ready and prepared is very challenging, a process that must improve for CMIP6 and other model intercomparison projects. Additional requirements are remote publication, the ability to publish various file formats with unknown file attributes, and publication for tape archives.

- ***Data QC*** is a framework within the data ecosystem and is used every time new data is introduced or generated for use. Data quality standards are in flux as sets of different standards are put upon the data as they move throughout the system. For example, the initial publication of simulation data may check for the netCDF data format and the CF convention. However, this data check many not be relevant for other forms of data, such as observational or image data. For an open ecosystem representing heterogeneous data collections, mechanisms for defining quality requirements are a must. For this reason, data quality

requirements must be independent, allowing projects to describe and include their own set of standards relating to their project's data relevance, accuracy, accessibility, timeliness, coherence, and project use.

For data quality assurance, inconsistencies must be documented and included in errata pages as well as operations to correct the data for use and the notification out to the community that the data was updated. As the community uses the data, it should be associated with a rating process to help determine QC process (accuracy, completeness, etc.). To further the QC process, digital object identifiers (DOIs) will be issues as a statement for community vetting and for journal publication.

- ***Data versioning*** is not simply for keeping track of modified data, although this is a primary use case. Along with data provenance and QC, data version control can also be about slight variations in the data within a data set or project. In either instance, versioning must be made clear and easily accessible to users for delta differences and intercomparisons. If multiple versions of data sets exist, end users will be directed to the latest version by default. Older replicated data sets at federated sites will also redirect the user to the latest version of the data by default. Data replicated at hosted sites will be automatically notified of the data updates, and administrators will be given the opportunity to manually or automatically update their sites. New versions of data sets will also automatically notify users who have downloaded the data.

  Version control for data is about tracking activity so that we can keep a clear record of data changes and identify the issues that warranted the changes.

- ***User support*** is to provide technical assistance for the many components of the ecosystem—that is, answering questions or resolving platform issues for users in person, via phone, or electronically. Currently, developers and some climate scientists are providing computer hardware, network, and software support for installation, usage, information searches, and science knowledge or interpolation. User support can be project specific; however, questions pertaining to users can be universal. The question becomes, who pays for user support and for how long? For some issues, a frequently asked questions list has been established through Askbot, but only 4% of the community uses it. Up to 80% of the community communicates via email. Because of the uncertainty of who supports the user community and to what level, 15% of all inquiries go unanswered. This is based on an average of ~2 ESGF questions a day over the December 2013 through September 2014 time period. Moreover, accountability and responsibility of the people answering questions is unclear and the process is uncoordinated.

  The overall vision for our proposed user support framework is to achieve user satisfaction and reliability of user support. Studies have given holistic perspectives on the support process and multiple representatives with different specialties and job descriptions have their own conception of the user support process. To prevent top scientists from getting bogged down with minor tasks, students should be hired to do same job, or self-help mechanisms should be set up. In addition, training should be offered at key conferences such as the Geophysical Union (AGU), European Geosciences Union (EGU), and the American Meteorological Society (AMS). No matter the solution, funding and communication will be key in addressing the user support needs.

- ***Security*** infrastructure is critical to ESGF. However, it is one of the biggest pain-points for the community and must be made simpler for use. Because of the restrictive nature of projects and data centers, data restraints requirements must remain in effect, which means that there can be no relaxing of data security requirements at this time. In fact, addition to safely securing the data, security must be expanded to other ESGF services such as publication, computing, and networks. This will ensure that only those who are authorized will have access to data and resources provided by ESGF node/group participants.

  Working with Amazon, and perhaps modeling security efforts after systems such as Facebook will help with the ESGF security process. As a step in the right direction of security simplification, MyProxy will

be removed from the EGSF software stack—it is not needed in the proposed open standard to authorization (OAuth 2.0) security framework for sharing credentials.

- ***Metrics*** collections are essential to successful organizational management of large projects such as ESGF and UV-CDAT and must be done continuously and automatically. For our purposes, metrics collected will be on the data, services, and system performance as they relate to individual community projects and as a collective of the federation as a whole. The projects require that automated processes capture the dynamic metrics (e.g., distinct user counts, resources used count, products delivered counts, and volume delivered) from the system and put them into a metrics database for viewing. There must be two types of metrics for viewing: (1) administration metrics for viewing system performance; and (2) usage metrics for reporting data and resources used by individual, groups, and projects. Basic metric development requires knowing criteria of a good metric, how to collect data that evolve into the metrics we track and how to use metrics to the betterment of our supported projects. Our metrics must be descriptive, controllable, efficient, understandable, precise, and credible.

- ***Data compression*** is absolutely critical for multi-petabyte archives such as CMIP6. For this reason, netCDF-4 data compression is required for large and small subsets of data to minimize the overall data archive size. Using netCDF-4 data compression is expected to reduce the overall archive four-fold. Accessing or updating a subset of data (or files) would require first the decompression (or uncompressing) of the file before delivering to the end user. From reported experiences at the F2F conference, the decompression and delivery time is negligible at best. To be determined is the level of compression. For example, Level 1 will compress files faster than Level 9. Conversely, Level 9 compression is intended to compress smaller than Level 1. For our use, only the variable data will be compressed and not the metadata information or its attributes.

  NetCDF-4 also provides shuffling, which may improve compression and chunking of compressed variables. Chunk shapes may have large effects on compression and performance. Large chunks improve compression, but may slow subset access. Therefore, we may want to study the chunking a bit more closely for optimal chunk sizes for individual projects. Still, we do expect to use netCDF-4 data compression, and we are experimenting with compression options to reduce the overall archive size and the data volumes.

- ***Search capabilities*** must be extended to allow users to find the information and resources they need to access and manipulate before intended use. Today, the typical data analysis and processing workload for which the system is designed is for users to go to a node, search for, and download data. In the future, the requirements are for users to search the federation for data, algorithms, and other resources and manipulate the data before downloading the end product (i.e. data, visualization, workflow steps). In addition, documentation has to be connected with files in a way that is immediately discoverable and accessible along with the end products (data files, visualizations, etc.).

  Seamless handling of non-netCDF files, including searchability and accessibility, is a requirement for many projects. It should also include PDF documentation, visualizations, movie files, and the like. Currently, attributes exposed in search depend on some aspects of the publishing system. Not all attributes are exposed. Search is built around CMIP5 needs, so as the community expands, the search attributes should expand as well. The time period in particular is not currently in the database, and that is important because some individuals only want a certain time span in the models or observations. Individual files can currently only be searched by hand.

  Search is not as easy as picking a few facets around which the CMIP5 requirements are built. Therefore, as the community expands, so too will the search attributes. For example, obs4MIPs has other attributes that do not line up with CMIP. For this reason, there are requirements for a stand-alone search. In any case, there is a need for much more powerful search tools than the simple facet-based search.

- ***Installation*** of the ESGF and UV-CDAT software stacks provides the community with proven installation tools such as build, test, and validation procedures to ensure that the software is set up properly. However, challenges still persist in installing both complex software stacks. For example, it has been reported that some have spent weeks installing ESGF at their site. Therefore, requirements are for easier installation and software installation to be completed in under an hour for anyone in the community. Although in most cases there is some testing of the software, there must be more integrated test suites and regression testing for post-installation. To compound these requirements, the software stacks port to various flavors and versions of Unix, Linux, and Mac (for UV-CDAT only) operating systems. Better documentation on how to get the software installed is required by the administrators and end users. For the betterment of installation, the installation team is moving to RPMs to help with the speedup and scripting capability of future releases. RPMs will also help in automating the entire installation process.

  In addition to RPMs, future releases of virtual machines will greatly assist installations and help administrators avoid worrying about super-user privileges and various security concerns.

  Data service tests must be integrated into the installation procedure. This means that sample data for testing must be incorporated with the release of the software stacks. Having the data incorporated with the test procedures provides installers with confidence they have installed the software stack properly.

- ***Data transfers*** must be reliable and secure for high-performance file transfer and synchronization. Many users explained how they developed separated sub-workflows that renewed the authentication certificates every eight hours, otherwise downloads would stop. Moreover, users indicated that there were good nodes and bad nodes to transfer data. For these reasons, we must have a fire-and-forget transfer mechanism that includes auto fault recovery that works within ESGF's seamless security protocol.

  Currently, ESGF utilizes Wget (HTTP) scripts to transfer the majority of data, which moves files onto the endpoint (or destination) local directory structure. This process is a bit complicated and very slow for massive data movement. Therefore, the requirement is to move over to utilize Globus transfer for both replication and user data movement. With Globus, users will be able to easily share data without having to move data anywhere else as long as other users have Globus access. This means that there will be options to either download directly to a user's machine or transfer data to another server. To accomplish faster transfers for end users and replication, the data transfer will utilize GridFTP over TCP and be integrated with the efforts of the high-speed network. Ultimately, the purpose of the data transfer team is to improve the speed and performance of data transfers between data centers and end users.

- ***Data subsetting*** using OPeNDAP is in place, however, generating OPeNDAP URLs was not a requirement for most of the projects, including CMIP5. Because subsetting is a critical function for data reduction and manipulation, it has been required from most projects to produce OPeNDAP URLs for project subsetting search and download capabilities.

  Since most analysis tool applications can read OPeNDAP files, this was the natural choice. This effort must be coordinated with the server-side computing effort. As such, OPeNDAP will allow end users to fetch subset or computing data to their institution and allow users to use their own analysis tool application for further manipulation. Moreover, subsetting must occur across multiple files and handle a number of types of geospatial subsettings, as well as temporal subsetting and pressure level subsetting. This includes wrapping and unwrapping for 360-degree geo-graphics systems.

- ***Data formats*** involve making a value judgment on particular data sets as appropriate for certain types of use by individual projects. For this reason, format enforcement for data consistency is extremely valuable, and ESGF has benefitted from uniform file formatting (i.e., netCDF format and CF convention). No matter which project is accessed, the end user can rely on the fact that the data will be the same and does not have to be converted for use. Most, if not all, projects have tried to adhere to the netCDF-CF format and convention. However, there are times when netCDF and CF are not appropriate, such as saving and

searching for image or movie files, or handling variables and attributes not defined in CF. In these non-netCDF and/or CF use cases, ESGF needs to better accommodate data models outside the normal realm of CMIP.

For the use case where netCDF and CF are required, the publisher should enforce the format restrictions. For the use case where it is inappropriate to use netCDF-CF requirements, the publisher should allow for any type of data set to be published into ESGF in such a way that will enforce format restrictions, but allow enough flexibility to accommodate project requirements other than CMIP. That said, the community is struggling with unfamiliar data formats and tools, and users new to the community are not familiar with netCDF-CF. We need an approach that will accommodate current and new users and many other formatted data and conventions.

- ***Documentation, user guides, video tutorials, and example problems*** are highly desired by the projects and user community. For scientists in other domains or neophytes in this domain, professional guidance is also desired, but this is a subject for a future report. A general concern with searching is that it is possible complete a search relatively effectively if the searcher knows what he or she wants, but if the searcher does not, it can be unclear where to go to find what is needed. ESGF documentation seems to be spread over two different websites: http://esgf.llnl.gov and the ESGF GitHub wiki. Some of the information is up to date, while other information is out of date—for example, the installation procedures or the user's manual.

  As a suggestion, there should be an ESGF and UV-CDAT installation and user's session at the AGU, EGU, and possibly AMS conferences. This might encourage the developers to update the installation and user manuals for teaching at the conferences and inform the users at the same time on how to install with the latest software script and use the software stacks.

- ***Cloud computing*** and awareness along with **cluster computing** are required by the community to remove data access and manipulation. Traditional HPC environments involve a computing system coupled to storage. For such HPC cases, the storage must not go down. At the opposite end of the spectrum is an object storage environment, such as a cloud, which is very large, scalable, and fault tolerant. This, we believe, will be the storage, distributing, and computing environment of the future. With this environment, the cost for storage and computing will be greatly reduced. The hope is to get computing for free with very little data storage costs. Some in the community have aggressively expanded their hardware infrastructure to include cloud technology. This means that ESGF and UV-CDAT must be flexible in deploying, expanding, and adapting to cloud solutions as part of its vast global ecosystem of hardware, networks, software, and services.

6 Feedback from Modeling and Data Centers and Their Requirements

Producing and/or maintaining high volumes of scientific data within a high-performance computing and storage environment present unique production and operating challenges. Often, the only realistic choice for long-term storage and backup are robotic tape drives within a hierarchical management system. As described at the F2F conference, how the data is organized can be problematic and have considerable impact on performance and operating costs. In the worst case, poor data organization means that data are never accessed simply because it takes too long or cannot be located. Fortunately, this was not indicated as a high priority within the currently used ecosystem.

However, the centers did describe multiple users finding, accessing, and downloading multiple data copies to their local area work space. Access constraints (security policies and firewall restrictions) set by HPC centers or data centers can lead to such inefficient behaviors. Redundant efforts waste considerable resources and significantly slow scientific productivity. Additionally, some applications require that large volumes of data be staged across low-bandwidth networks simply to access relatively small amounts of data. Finally, when data usage changes, or

storage devices are upgraded, large data sets may need to be reorganized to take advantage of the new configuration, making the entire process expensive and extremely time consuming for the centers. For these and other reasons (described below), there is a need to develop a more intelligent and integrated data ecosystem (i.e., **Figure 1**) across the federation that provides services that anticipates usage, storage, and better communication and Internet connects between the centers.

Feedback and requirements stem from several of the world's premier climate modeling and data centers (i.e., IPSL, NCI, DKRZ, BADC, NCAR, GFDL, and LLNL). Since many of their requirements are represented in Section 5, only the requirements representing significant differences are listed below. Not in any particular order of importance, the summary requirement findings revealed the need for:

- **Improvements and standardization in data publishing.** Data requirements at modeling and data centers vary greatly depending on the needs and services of their communities. For example, to be considered a Tier 1 data center in the ecosystem, data centers must provide 10 PB of rotating disks and compute resources for data reduction and manipulation, as well as be connected to other Tier 1 data centers via the ICNWG 8 Gbps connection protocols. Eight Gbps is the minimum Internet connection speed to replicate 1 PB of data in 14 days at another node site—a requirement needed for the automated process of verification and updating missing data. These requirements are established for the CMIP6 project, which in most cases (but not in all cases) sets the standard for many other projects. At all sites, preparing, finding, replicating, and archiving data for long-term storage is expensive. To help keep the cost down, the centers would like to see data publishing more streamlined and automated.

  For the federation, different algorithms for compressing output are needed for centers to minimize storage costs.

- **Federation-wide analytics on how data are used, for long-term project planning and interagency coordination.** Analysis, particularly large-scale analysis, stress central processing unit (CPU) center resources, whether they are HPC in-situ, cluster or cloud analysis. To reduce CPU usage, centers post-process derived variables for commonly computed variables such as yearly ensembles and zonal means. In recent years, work has gone into the areas of providing information-rich data access and analysis workflows that utilize both data compression and reduce CPU usage. These workflows and others are needed to assist centers in the automation of data processing and accessing.

  Modeling and data centers point out that version history for replicas and data sets—a crucial notion during analytics phases—needs attention. The replica notion is important for data centers but adds confusion during analytics phases, when what is important is to reach the data and not whether this data is a replica. With version history, what is needed is a way to figure out rapidly whether an analysis used up-to-date data versions, and if not, whether an out-of-date version is scientifically acceptable.

- **Detailed governance**. Governance is a real issue for the centers. In addition to what was mentioned in Section 5, governance requirements must include operations and installation of the hardware, software, and networks on behalf of the projects. This includes available roadmaps, timelines, milestones, deliverables, and responsibilities of the centers. Overall governance not only affects the projects, but also affects the centers' operational systems and overall costs.

7 Technology Developments

The primary ecosystem product will be a scalable architecture capable of supporting extreme-scale and collaborative science, across a range of communities and projects as listed in Section 4. A set of management tools and services will make it easy for researchers to customize, manage, and use the infrastructure to navigate, manage, manipulate, and share the data and associated metadata related to a project contained within ESGF. These tools will allow researchers to:

- **Manage:** Create and destroy data, specify various attributes (e.g., schemas), and control who is allowed to update contents and access metadata;
- **Populate:** Incorporate files, plus associated metadata either via explicit commands or implicitly by specifying directories or job queues to be associated with the project, or by executing discovery procedures that mine information from file metadata or other sources;
- **Explore:** Search, browse, and access contents, via graphical and command line interfaces will include active views that update in real time as data elements are added and modified;
- **Manipulate:** Invoke operations on specified data contents, for example, to generate derived data products, move older data to archival storage, and transfer data to remote locations; and
- **Disseminate/Collaborate:** Provide interfaces for data and metadata from in a form suitable for access by many remote users, and enable controlled remote access to data and resources.

With these tools in place, it will become straightforward for a researcher to:

- Compare the results of a set of computations with other previous computations or experimental results and find files with similar properties;
- Determine the location, identity, format, authorship, age, access patterns, and space consumption of all files (i.e., provenance) associated with a project;
- Discover and operate on data that match specified criteria across project domains of interest with large variety of collections of data; and
- Publish scientific outputs, and a rich collection of associated metadata, in a manner that permits access by either a small, controlled set of research collaborators or an international community.

We intend for our ecosystem solution to meet the needs of a range of projects from different climate domains (i.e., simulation, observation, and reanalysis) and with degrees of technological and infrastructure sophistication. Our approach is to offer applicable components as open-source modules that can be customized and extended by the user communities. (In practice, we expect that many communities will use a mix of hosted and downloaded components.) We will also document the architecture, component structure, and application-programming interfaces (APIs) so that other developers can contribute to the development in the future and participate in the ecosystem. Below are the list of technical components under development and timelines of their completion for the coming year.

7.1 ESGF Technology Development

7.1.1 International Climate Network Working Group

7.1.1.1 Project Needs

Over the last year, Dean Williams and Eli Dart formed the ICNWG under the ESGF (see **Figure 4** showing the overlapping institutions and support for the working group). Additionally, supported by Enlighten Your Research Global (EYR-Global), this working group creates a forum to support the proposed EYR-Global project ("International Networking for Climate") and participating ESGF sites as they troubleshoot performance between internationally hosted ESGF nodes. In addition, the working group created a website to track the working group's progress (http://icnwg.llnl.gov).

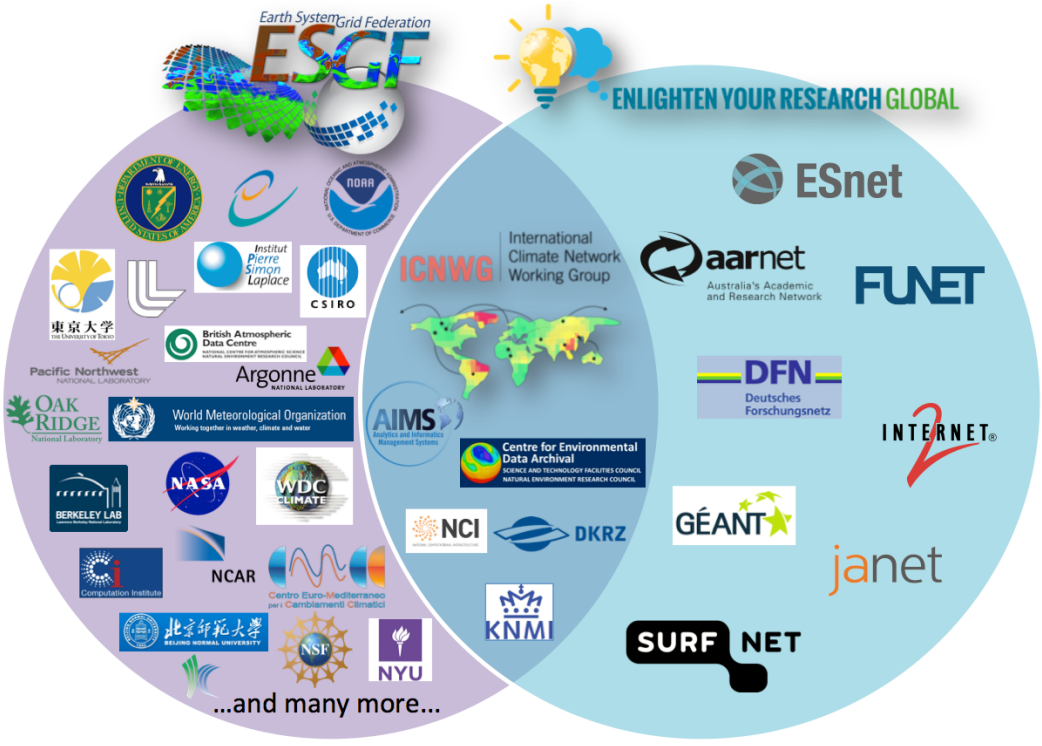

Figure 4. *Diagram showing overlap between ESGF organization and the EYR-Global Program. Data centers noted in the crossover region are supported by EYR-Global network organizations (on the far right). ESGF has many more institutions that participate in their program; however, five data centers are pioneering upgrades or updates to their infrastructures to improve large-scale data replications.*

ESGF sites that are participating in this working group include: AIMS in the U.S., CEDA in the U.K., DKRZ in Germany, NCI in Australia, and KNMI in the Netherlands.

When starting this project, data transfer performance was variable between all sites, between 10 KB/sec and 40 MB/sec. In many cases, the software used to transfer the data is Wget or similar HTTP-based tools. ICNWG aims to transition the collaboration to GridFTP for data replication, driven by Globus. A data set will be acquired in February 2015 to test disk-to-disk and memory-to-memory performance and provide a realistic approximation of production-level data replication performance.

In general, local network, server, and storage infrastructures must be upgraded to accommodate the data transfer performance required for this project. Many sites have made significant progress in these areas, and we have enumerated the details of their work in this report.

This project is a huge undertaking that requires coordination between different countries with different policies and traversing multiple network domains. To improve network connections from end to end, summing to thousands of miles of fiber and tens of network domains, this group will continue to work to support reliable, long-scale data transfers across the international domains. The ICNWG will continue to test and troubleshoot the network paths to each site, with ESnet's help, for the duration of the project (until 2016).

7.1.1.2 Summary of Progress

Four out of the five participating sites have deployed perfSONAR nodes. perfSONAR measures the network performance capabilities at the end sites by using the tools bwctl (run every few hours) for throughput testing and owamp (running continuously) for low-bandwidth one way delay measurement and packet loss testing. The results are then stored on a server, which can be viewed using a web interface, MaDDash (Monitoring and Debugging Dashboard), for easier performance troubleshooting and understanding (see **Figures 5** and **6**).

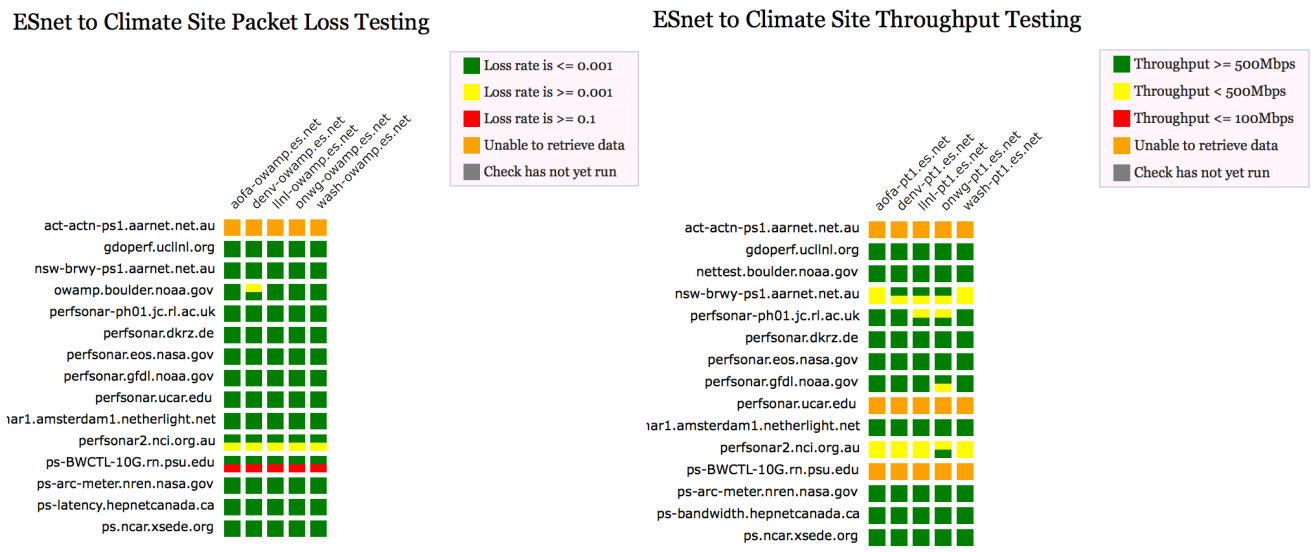

Figure 5. *(Top) OWAMP dashboard showing packet loss between ESnet perfSONAR boxes and ICNWG participating organizations. (Bottom) BWCTL dashboard showing bandwidth tests between perfSONAR nodes (pt=performance testers). These are single-stream test results. When running tests four-way parallel, some sites are meeting network speeds already. (Please note network organizations and data centers have or already had perfSONAR deployed that makes testing network paths easier.)*

In the next six months, the ICNWG will have its own server, icnwg.es.net, to accumulate data between all the perfSONAR nodes. The server is hosted by ESnet, which will manage and maintain icnwg.es.net for perfSONAR data. This will allow all nodes to test to each other in a mesh-like environment, which will show more realistic results for data replication that may not traverse ESnet—for example, between Australia (NCI) and Germany (DKRZ).

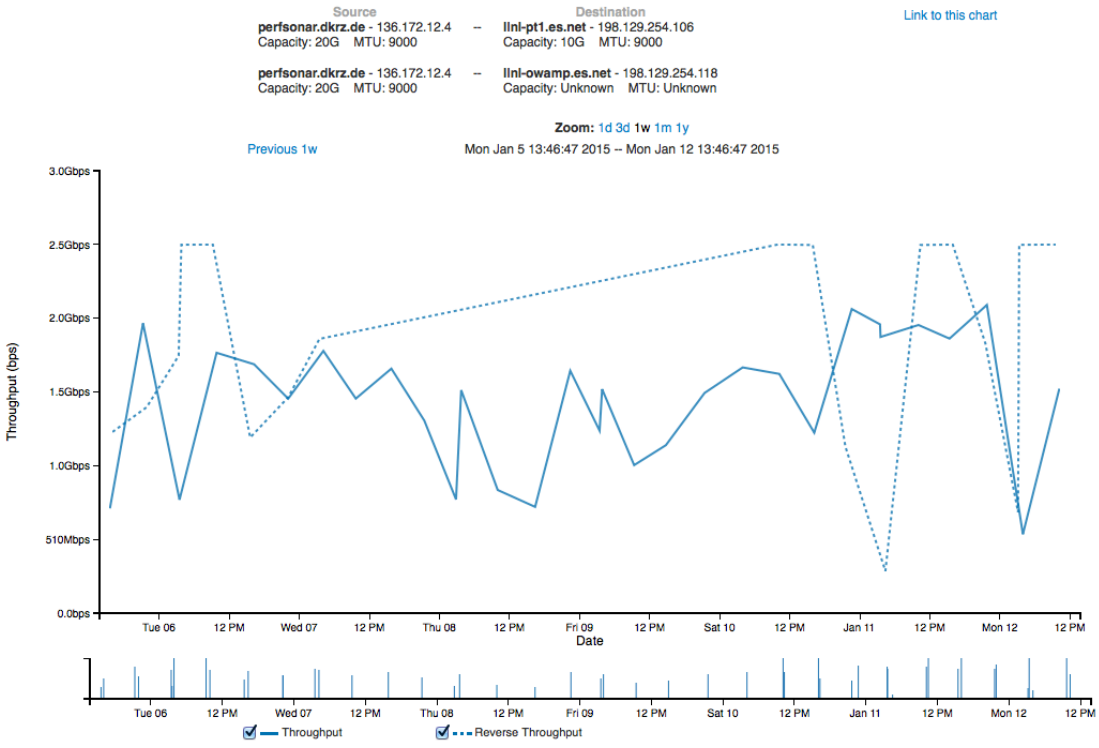

Figure 6. *Clicking on a cross section of the dashboard grid displays further information. This example shows the BWCTL tests between DKRZ and the Lawrence Livermore National Laboratory (where AIMS is located) for the past week. Notice bandwidth reaches an average of 2+Gbps for single-stream tests. By looking at the trends in the BWCTL and OWAMP graphs, engineers are able to narrow down issues in network performance due to failing hardware, asymmetric routing behavior, etc.*

The first goal for ICNWG was to achieve 500 MB/sec (4 Gbps, approximately 1 PB per month) of disk-to-disk throughput during the replication of a multi-terabyte data set, using production infrastructure. This has yet to be completed between all the sites, but there is assurance that the working group sites will reach this goal within the next six months. In 2015, the ICNWG will increase that performance by a factor of two (to 1 GB/sec, 8 Gbps, 1 PB per 14 days) between nodes. Achieving this capability on production systems will help prepare the global climate science infrastructure for the demands of CMIP6, and set the stage for continued scientific productivity in the critically important area of climate science. As a stretch goal, if possible, the group wishes to try for a second doubling of performance to 2 GB/sec (16 Gbps, more than 1 PB per week) between centers in 2016.

In the more immediate future, sites will be deploying or adding more data transfer nodes to their site. With proper tuning, these sites should be able to reach the ICNWG milestones.

Table 1. *Expected technical milestones for the year. Below is the original timeline for achieving the ICNWG's goals for each site at AIMS, CEDA, DKRZ, NCI, and KNMI.*

Timeline of Milestones for Network Service
March 2014 – September 2014
• Deploy 10G perfSONAR test server • Deploy 10G data server • Set up perfSONAR tests
July – December 2014
• File system tests for 10 G data servers—target 500 MB/sec • Achieve 500MB/sec (4 Gbps) network test throughput between perfSONAR test servers
January – February 2015
• 500MB/sec (4 Gbps) disk to disk transfers between data servers
February – March 2015
• Deploy second 10G data server • Configure second 10G data server for striped Globus/GridFTP transfers with first 10G data server
March 2015
• Test striped Globus/GridFTP transfers with one other center
June 2015
• Test striped Globus/GridFTP transfers with all centers
August 2015
• Demonstrate 1GB/sec (8 Gbps) transfers between all centers
Remainder of 2015
• Extra time for schedule slip • Prep for 2016 stretch goals
June 2016
• Demonstrate 2 GB/sec between all centers that are capable (stretch goal)

7.1.2 ESGF Installer

The ESGF Installation system will follow several axes of development in 2015. The general idea is to standardize the subcomponents' installation script arguments to make it easier to rewrite the installation controller. As long as the standard arguments are respected, the language that will be used for these subscripts does not matter. RPMization is also an important task, as it will make node installation significantly easier for system administrator–oriented skilled teams; this will be done gradually, and Globus is the first candidate. Once the subscript arguments have been standardized and rewritten, it will be time to move to a Tomcat system. This step

will make it possible to go on with RPMization of web applications. Future developments will focus on stripping down the ESG-node script to simply start and stop the system, as well as writing a new installation controller.

Table 2. *Expected ESGF installation milestones for 2015.*

Timeline of Milestones for ESGF Installer
April 2015
• ESGF 1.9 Release • Globus RPM • CoG Integration
August 2015
• ESGF 2.0 Release • Subscripts arguments standardization (node manager, security, ORP, IdP, search, web-fe, CoG, dashboard, desktop)
November 2015
• ESGF 2.1 Release • Use of Tomcat server

7.1.3 ESGF Build v.2

The ESGF Build platform is hosted on an IPSL machine and available at esgf-build.ipsl.upmc.fr. It runs a Jenkins server, which is listening to the different ESGF projects GitHub repositories. Whenever a code modification is pushed to any repository in the development branch, a build pipeline is executed and freshly compiled binaries are then available online.

The developments in 2015 will focus on binaries continuous deployment. Now that continuous build has been set up, the next step is to push the development binaries to the ESGF distribution mirrors automatically.

Table 3. Expected ESGF build milestones for 2015.

Timeline of Milestones for ESGF Build
February 2015
• Establish continuous deployment system prototype
June 2015
• Test continuous deployment system
December 2015
• Continuous deployment operational in development environment

7.1.4 ESGF Test Suite v.2

The ESGF Test Suite is a full python application designed to perform integration or post installation tests on ESGF nodes. At this point, the scope is to test a single data node and its three peer services (IdP services, index services, and compute services). The ESGF Test Suite can run high-level tests from a desktop so the tested node can be validated from the end-user perspective.

During 2015, the technical developments will focus on adding some new testing features to complete the suite—for example, federated search testing and Wget download testing. The next step will be then to make the suite executable from both any user's desktop and from an ESGF node itself. This will open new perspectives on testing internal services such as data publication.

Table 4. *Expected ESGF test suite milestones for 2015.*

Timeline of Milestones for ESGF Installer Test Suite
February 2015
• Federate search test scenario
March 2015
• Wget download test scenario
April 2015

• ESGF test suite will be executable both from a user's desktop and from an ESGF node
July 2015
• Publication test scenario

7.1.5 CoG

CoG is a web-based environment where users can host, manage, and share scientific projects. It provides an enhanced user interface (UI) to the data and metadata services of the ESGF and adds new capabilities for describing projects through standard metadata, collaborative authoring of wiki pages, configuration of project search functionality, and management of the federation among ESGF nodes. CoG is a web application based on the python–Django software stack and can therefore be integrated with other Django-based applications.

During 2015, CoG development for ESGF will focus on finalizing the process of replacing the current ESGF web front end, which began more than a year ago. This will involve integration of CoG with the ESGF installer, deployment at all ESGF nodes, and setup of bidirectional connections among all nodes in the federation. Additionally, the CoG functionality will be augmented as needed based on feedback from ESGF users and administrators, consistently with the over-arching goal of providing a superior UI to serve climate model and observational data (in particular, the current CMIP5 global archive and in preparation for the upcoming CMIP6 experiments). Finally, the CoG governance bodies (Steering Committee, Executive Committee, and User Review Group) will start to operate to plan, direct, and manage future CoG development in relation with ESGF.

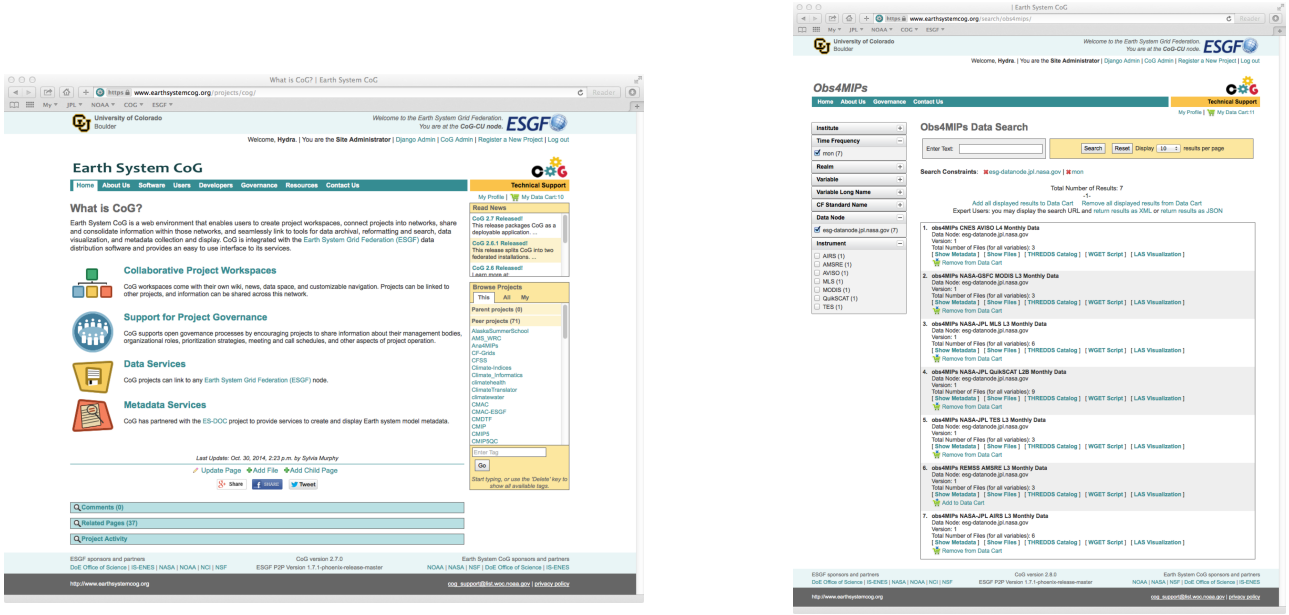

Figure 7. *A few examples of current CoG web pages: The home page of the CoG installation at the University of Colorado (left) and the Obs4MIPs search page (right)*

Table 5. *Expected CoG milestones for 2015.*

Timeline of Milestones for ESGF CoG User Interface
March 2015
• CoG deployed to ESGF nodes as part of the ESGF installer
• Run as standalone web application
May 2015
• CoG deployed to ESGF nodes as part of ESGF installer
• Run behind Apache HTTPD server

June 2015
• CoG replaces current ESGF web front end at all ESGF nodes
July 2015
• CoG monthly releases with enhanced functionality as directed and prioritized by the CoG governance bodies

7.1.6 Publisher

Publication allows a user with necessary privileges to push a data set into an archive and associate metadata with the data set for later discovery. This activity involves many steps, including moving the data set to the location where it will be stored and hosted, providing necessary metadata, extracting other metadata from the data set, verifying data format and metadata, and managing permissions on the metadata and data. The steps are both repetitive and onerous, and can be simplified and made more reliable by providing them as a service that the user can leverage.

Publication as a service allows the user to submit a publication task and provide the necessary information; the rest of the process is then completed for the user asynchronously. The "fire and forget" model reduces the burden on the user, simplifies the interactions needed for a user to publish, and can increase data quality by automating various steps and checks.

We currently have a proof of concept demonstration of the service, and this year will continue to build out the first version of the service. In addition to work on the new service, several improvements to the ESGF publisher tool were made, which included a usability fix, a "dry-run" operation mode, and a new tool that checks the synchronization of various metadata sources.

We expect the command-line ESGF publisher to remain an option for the foreseeable future. We plan to release a minor update to the publisher in early 2015. To further improve the command-line publisher, we plan to improve the configuration interface and method by which supplemental searchable metadata can be supplied for existing data sets. These changes are driven in part by requirements from the ACME project. Efforts will require coordination to understand the requirements, design the interfaces and software changes, carry out the implementation, and deploy the software to ESGF data nodes for testing.

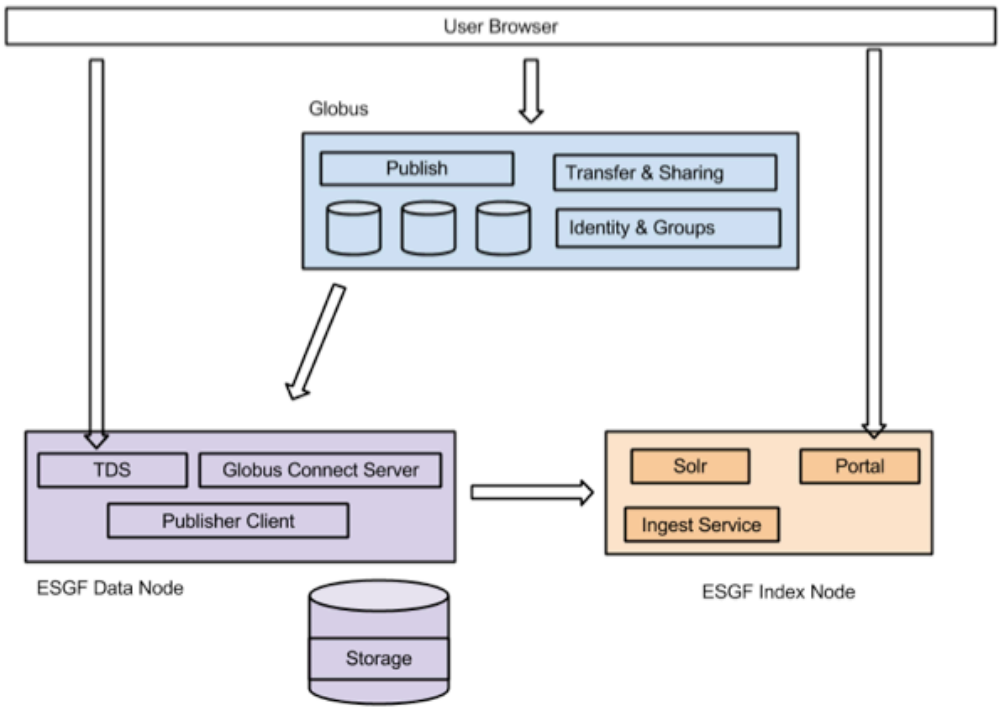

Figure 8. *The new workflow for Publication as a Service.*

Table 6. *Expected Publication milestones for 2015.*

Timeline of Milestones for ESGF Publication
February 2015
• Beta publisher release (release with ESGF node installer TBD) • Interface design for publisher changes
March 2015
• Software design for publisher changes
May 2015
• First alpha release of Publication as a Service to select projects
June 2015
• Requirements gathering completed for CMIP6
September 2015
• Software changes available for testing on ESGF nodes • Release of Publication as a Service with CMIP6 requirements

7.1.7 Integration of Documentation from External Repositories

Within the quality team, we consider how to increase the quality of ESGF data services. The main working task is the integration of external information into ESGF (e.g., version information, quality, or data citation). This implies the storage of unpublished events for provenance and the support of data collections (granularity of data citations, etc.). In 2015, the working team aims to provide stable sample implementations for the version information repository (local repository located at a data node) and the data citation repository (central repository).

Table 7. *Expected quality control milestones for 2015.*

Timeline of Milestones for ESGF Data Quality
June 2015
• Test installation for examples "version information" and "citation information" to show modeling centers and project coordinators
December 2015
• Stable implementation of version and citation information systems • Integration of versioning tool into ESGF nodes • Test integration of other external repositories (e.g., quality information)
June 2016
• Operability within ESGF

7.1.8 ESGF Identity, Entitlement, and Access Control

ESGF Identity, Entitlement, and Access Control (IdEA) is the system for managing access rights to data sets and other resources within the ESGF federation. It was initially developed to manage access to the CMIP5 archive hosted through ESGF and has been extended to manage access to other projects with ESGF such as CORDEX. Its scope includes systems for single sign-on—both web and command-line based systems with OpenID and MyProxyCA respectively—and registration and authorization to access federated resources with interfaces based on the SAML security specifications. Currently, this covers access to data sets and access restrictions on publishing data into the federation.

The access control system has faced a number of challenges in its use operationally since its first inception for CMIP5. These stem from the usability of services for end users, the complexity of using and maintaining public key infrastructure and other associated security configuration for node deployers and operators, and maintaining resilience of services in a large distributed federation made of many independent organizations.

At the 2013 ESGF meeting, a roadmap was defined to direct a series of technical developments to address a number of these challenges and update ESGF in line with the latest technologies and thinking in the area of federated identity management and access control. A working team has been established to coordinate and oversee the work.

Based on the roadmap over the course of the last year, a number of enhancements have been implemented by CEDA (Centre for Environmental Data Archival, UK). These include a simplification of the system of authentication with Wget scripts. These scripts are generated for users of the ESGF web front end, enabling them to perform a bulk download of data offline from their browser session. The new system removes the need for user X.509 certificates greatly simplifying the process.

In addition to enhancements to the existing system, a significant area of new development is the provision of support for user delegation. User delegation enables a user to grant a third-party service the right to act on his or her behalf when performing some action requiring the user's security credentials. One example would be a web portal displaying secured data from other nodes in the federation. In order to access this data, the portal needs to obtain some limited access to the user's credentials so that it can retrieve the data. This capability has not been supported by ESGF. As part of the roadmap, a system based on the popular OAuth 2.0 framework is being integrated into ESGF, enabling portals and other applications to obtain a certificate representing the user's identity and use the certificate to access secured data on their behalf. This work has also enabled the replacement of the MyProxy software with a simple web interface.

Using OAuth 2.0 also positions ESGF to take advantage of developments in single sign-on technology. The developers of OpenID 2.0 have created an OpenID Connect specification, which uses a completely new approach based on OAuth 2.0. A key focus for the work in the coming year will be the integration of OAuth 2.0 into ESGF in a phased approach towards full adoption of OpenID Connect for single sign on. This will need to closely co-ordinate with the installation team and the operational requirements of the federation. Pilot projects will be used to test new capabilities. The first of these will investigate integration of an ESGF OAuth 2.0 service with Globus Online to facilitate node data download.

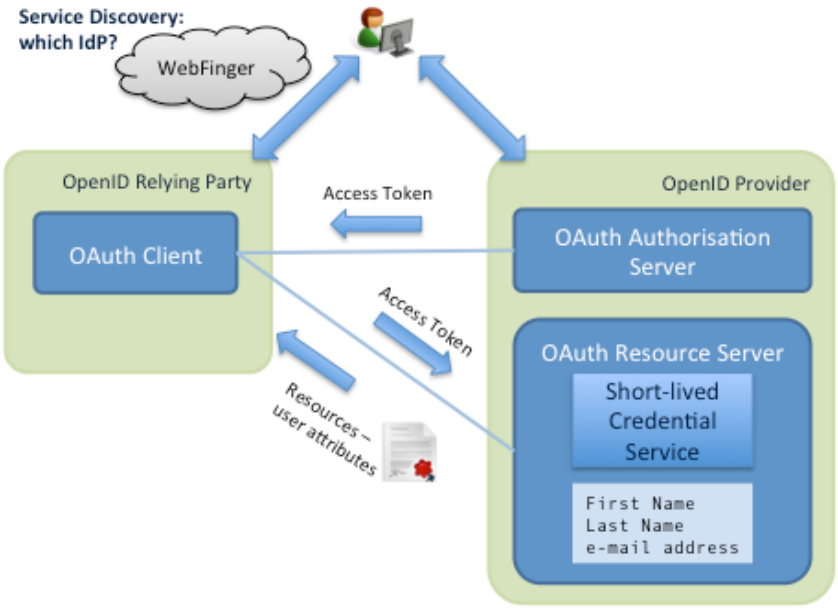

Figure 9. *Proposals to use OAuth 2.0 and OpenID Connect technologies for future provision of user delegation and single sign-on respectively. Development of the two can be linked since OpenID Connect builds on top of the OAuth 2.0 framework.*

Table 8. *Expected ESGF security milestones for 2015.*

Timeline of Milestones for ESGF Identity, Entitlement, and Access Control
February 2015
• Coordinate with installation working team to integrate improved Wget and OpenID sign-in enhancements into formal ESGF release • Initiate pilot for integration of ESGF OAuth 2.0 with Globus for ESGF node data download
March 2015
• Completion of study • Simplification of PKI trust roots for the federation
May 2015
• Complete pilot for integration of ESGF OAuth 2.0 with Globus for ESGF node data download
June 2015
• Identify candidates for further pilots for new OAuth 2.0 user delegation capability • Commence further pilots as required • Review operations policy and provision of IdPs for ESGF with the goal to reduce and simplify IdP provision
September 2015
• Provision of initial production of OAuth 2.0 user delegation services for ESGF IdPs
October 2015
• Commence plan and implement OpenID Connect development
December 2015
• Report back on progress with OAuth 2.0 services, pilots and OpenID Connect development

7.1.9 Data Transfer (Globus for ESGF)

Data transfer is a key capability in the ESGF stack, both for replicating data across sites and for users to move data to their machines or to other analysis clusters. The Wget solution still in use by many has limitations in terms of usability, scalability, and the need to always download to the machine where the Wget script is executed.

Globus Transfer provides reliable, secure, managed data transfer. It leverages GridFTP, a proven protocol for high performance data transfer, and provides platform capabilities allowing clean integration with the ESGF stack. Globus Transfer has been integrated as an option in the ESGF portal data cart, allowing the user to download to his or her own machine using Globus Connect, or transfer to another server.

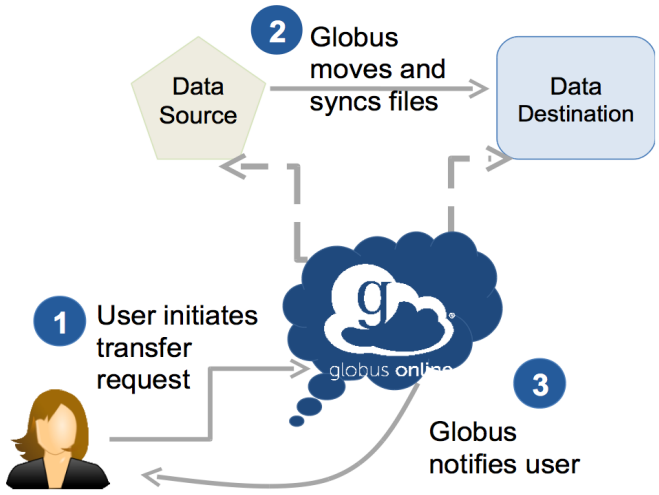

Figure 10. *Globus software data transfer and integration.*

Table 9. *Expected data transfer milestones for 2015.*

Timeline of Milestones for data transfer Publication

April 2015
• Integrate Globus transfer front end with CoG user interface
June 2015
• Upgrade ESGF data nodes to newer version of GridFTP server
July 2015
• Upgrade installation to include binary RPMs to eliminate need to build Globus components from source

7.1.10 ESGF Replication and Versioning Working Team

Versioning practices in ESGF are currently less than optimal, mostly because there is no unified and controlled process that is supported directly by ESGF services. Replication suffers from problems such as inconsistent versioning, a non-standardized replication workflow, and unobserved updates of data sets.

To address the versioning issues, the Replication and Versioning Working Team (RVWT) will employ persistent identifiers (PIDs) and associated state information as an anchoring mechanism. The goal is to assign PIDs to the files entering the ESGF data space at publication and track any changes such as file removal, updates, and extension of preliminary data sets. End user services that can exploit this information include, for example, obsolescence notifications and automatic redirection to latest versions. In parallel, the RVWT will simplify the publisher, define essential processes to be followed, and introduce a versioning system with automated numbering.

All activities are geared towards CMIP6 to ensure the development timeline focuses on 2015. The processes and technical solutions will be tested at single nodes first, particularly at DKRZ, with a focus on a stable service for PID assignment to individual files. The technical development builds on several existing solutions, including the Handle System for PID resolution and the PID Types API developed during activities of the Research Data Alliance. If possible, developments will be coordinated with similar PID activities in the frame of the EUDAT 2020 project. While other community projects are encouraged to participate in PID-based processes, resources are limited and immediate adoption cannot be expected. The working team will therefore develop hybrid solutions that can also continue to work with traditional tracking universally unique identifiers (UUIDs).

For replication, the work in 2015 will concentrate on the improvement and widespread use of the synchro-data tool (http://forge.ipsl.jussieu.fr/prodiguer/wiki/docs/synchro-data). In parallel, the performance of concrete ESGF-to-ESGF site bulk data transfers will be tested and monitored in coordination with the ICNWG networking group, starting with replication between PCMDI (U.S.), DKRZ (Germany), and NCI (Australia).

Table 10. *Expected ESGF replication and versioning milestones for 2015.*

Timeline of Milestones for ESGF Replication and Versioning
February 2015
• Implementation plan finished
• Dependencies set and relationship with existing and other ESGF developments
March 2015
• Develop tasks coordinated with working team
May 2015
• Synchronize data deployments with DKRZ, LLNL, and ANU for replication
• Synchronize data tool prototype supporting GridFTP transfer protocol
June 2015
• Unified versioning process implemented and locally tested
• Initial PID assignment to files via CMOR implemented and locally tested
September 2015
• Publisher enhanced with PID registration and PID-based versioning
December 2015
• Federation tests of PID and versioning processes performed at all participating nodes, processes reviewed

• Operational version of synchronized-data tool supporting multiple transfer options, including GridFTP

7.1.11 ESGF Dashboard and Monitoring

The Dashboard (FASM component) is a key milestone for the IS-ENES2 project. The system is a distributed and scalable monitoring framework to capture usage metrics, system status, and aggregated information at the single-site level, at the ENES archive level and at the global ESGF level. The Dashboard faces this important challenge through two main modules: the back end, responsible for collecting and storing a high volume of heterogeneous metrics, and the front end, which will provide the user with an intuitive web interface including local and global views, aggregated statistics, and monitoring information. The design phase of the system was completed in March 2014, with the production of the IS-ENES2 deliverable to monitor the system and Dashboard design. Since then, the existing views of the Desktop have been updated in order to be compliant with the design report.

During 2015, the activities will focus on improving and extending the existing framework to support new community requirements. A new set of sensors will be defined, focused on user metrics; in addition, new views and federation-level mechanisms for global statistics will be implemented. These activities will require the definition of a new version of the logging database in order to overcome the existing issues related to the current implementation of the node manager.

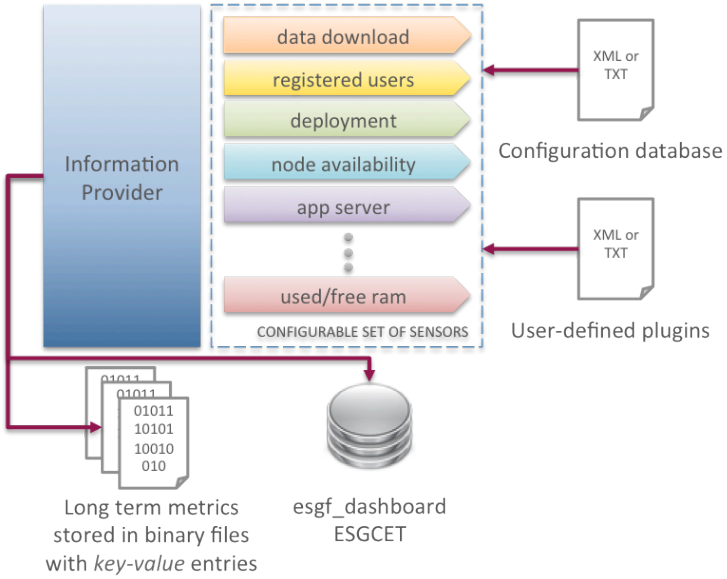

Figure 11. *Architecture of the ESGF Dashboard: the Information provider receives updated snapshots of the status of the federation through a configurable set of sensors.*

Table 11. *Expected data transfer milestones for 2015.*

Timeline of Milestones for ESGF Dashboard Publication
March 2015
• Definition and implementation of federation-level mechanisms for global statistics
July 2015
• New release with a refined and extended set of metrics and user interfaces
September 2015
• REST-ful interface to get access to the dashboard metrics
December 2015
• New release with views for fine grain or aggregated statistics per node, peer-group or federation level

7.1.12 ESGF Improved Usability and Support

The ESGF Support Working Team (SWT) is a collection of people from around that globe that are attempting to give ESGF users the best experience as possible. We have learned a substantial amount over the last two years. ESGF had transitions and changes in both their Wiki and website. At this year's F2F, Matthew Harris spoke of our experiences and the direction our group should go to continue to enhance user experience.

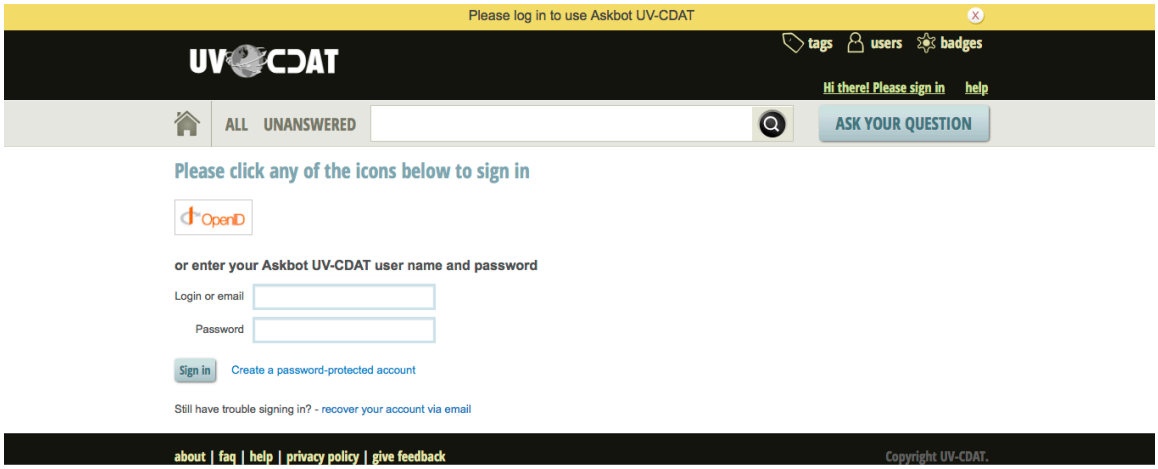

Figure 12. *Example of future Askbot Question and Answer site needed for ESGF users.*

Table 12. *Expected ESGF support milestones for 2015.*

Timeline of Milestones for ESGF Support
January 2015
• Find another co-lead for the ESGF Support Working Team (Torsten Rathmann has agreed to co-lead ESGF-SWT with Matthew Harris) • Set a bi-weekly time slot for ESGF-SWT teleconferences
March 2015
• The ESGF user community has two different beta versions of Askbot. Settle on one of the versions and reinstall and configure at LLNL • Currently none of the ESGF mailing lists are archived. Must make sure all mailing list are archived so that valuable information is retained
May 2015
• Once the ESGF mailing lists are archived, set up a method for users to search, access, and view mailing lists for knowledge discovery
July 2015
• The original wiki site is cumbersome and difficult to find needed material and thus a re-organization and re-structuring of the wiki will help users find wanted information

7.1.13 Scalable Node Manager

The node manager is a key component in the ESGF node software stack that gathers metrics, shares node information across federated nodes, and facilitates user-group management. The present implementation of the node manager has scalability limitations arising from the peer-to-peer protocol. We are redesigning the node manager to correct existing limitations in consistent reporting of information and address these scalability limitations. Our design centers on use of communication patterns that incorporate both hierarchical and peer-to-peer organization. We introduce the concept of "Super Nodes," which do additional heavy lifting for the node manager functionality. Additionally, other nodes will be able to assume a Super Node role upon failure of the original Super Nodes. The next-generation node manager will provide a consistent store, which should enable coordination of activities or failover capability for additional ESGF software components, beyond the several

components that presently rely on the node manager. Moreover, we will support current components such as the dashboard by providing a registration.xml.

The design phase for the node manager is still in progress, and we expect to elicit feedback from the community once we complete it. In the meantime, we can prototype several of the key node manager functions, such as communication protocols for health check and Super Node failure. We plan to rely on Django services, ensuring their integration with the current software stack. The integration of Node Manager Django services with the other ESGF services running under Django will be key to a successful deployment of the Node Manager-enabled ESGF node. Given that, we can plan on a deployment roadmap for existing ESGF nodes. We anticipate making considerable implementation progress in 2015 and should have testable code deployed to nodes by the year's close.

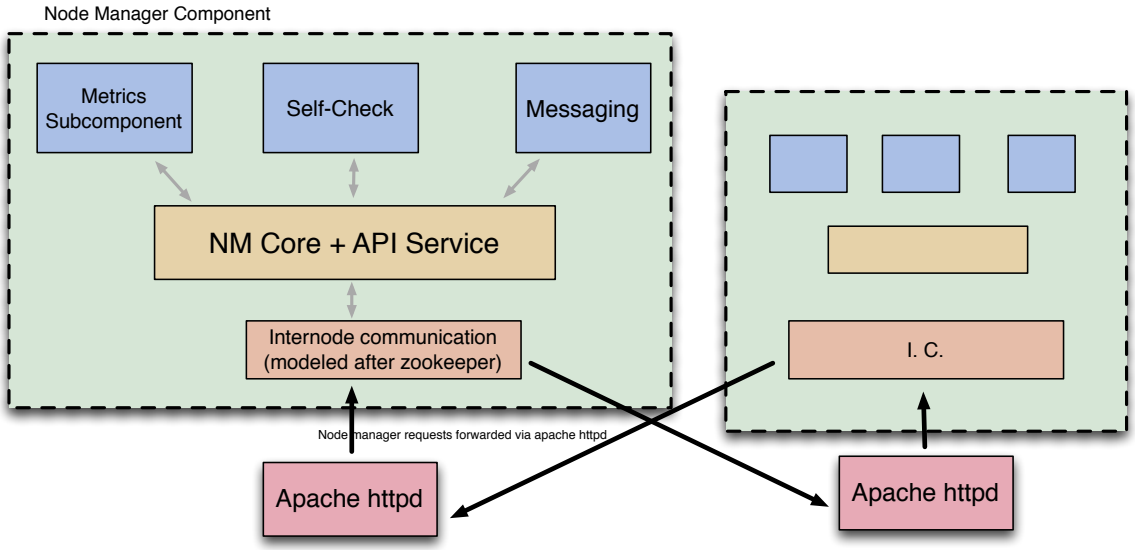

Figure 13. *New ESGF Node manager design for optimal sharing of federation information.*

Table 13. *Expected ESGF node milestones for 2015.*

Timeline of Milestones for ESGF Node Manager
February 2015
• Submit software design to community for feedback
March 2015
• Receive and incorporate design feedback
July 2015
• Full alpha version completed in test environment
September 2015
• **Alpha version tested at select sites**
December 2015
• Beta version tested at additional sites

7.1.14 ESGF Search Services

The ESGF Search Services allow clients to search for climate data across a global federation of distributed archives that includes tens of institutions around the world. The collective archive spans tens of projects, including the most prominent CMIP5 data used for the IPCC-AR5 report, and the supporting obs4MIPs, and ana4MIPs data sets. Internally, the ESGF Search Services are based on the popular Apache Solr engine and leverage its distributed search and replication capabilities (see **Figure 14**).

Despite its powerful functionality and increasing widespread adoption throughout the climate community, the ESGF Search Services infrastructure could benefit from adopting a more rigorous approach to defining and validating the metadata ingested into the distributed indexes. During the next few years (2015 and beyond), the technical development will focus on improving the functionality, reliability, and efficiency of search, as well as upgrading the infrastructure to support scaling to the next generation of climate models (CMIP6) and instrument observations (NASA decadal surveys and others), which together are expected to increase the size of the current archives from 10 to 100 times.

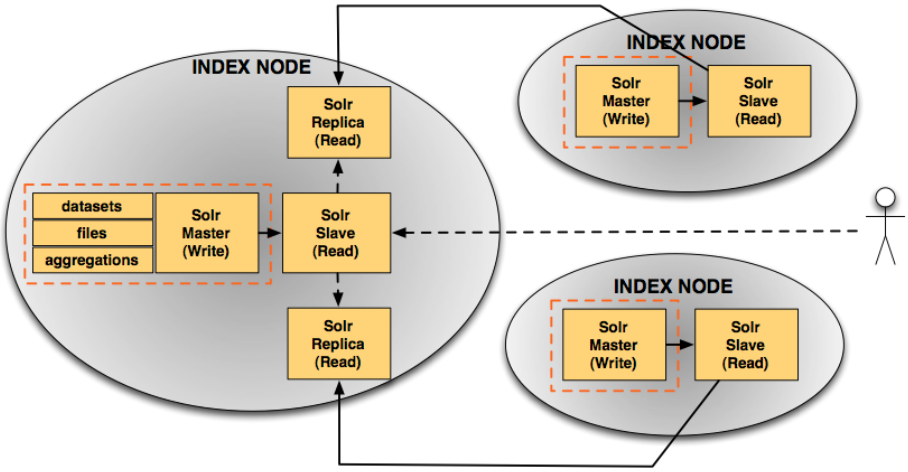

Figure 14. *Current architecture of ESGF Search Services, showing index replication and distributed search.*

Table 14. *Expected ESGF search milestones for 2015.*

Timeline of Milestones for ESGF Search Services
February 2015
• Establish prototype ESGF-Cloud node based on Solr-Cloud
April 2015
• Upgrade all ESGF Index nodes on Solr4, running on port 80
June 2015
• Upgrade ESGF metadata schema to support geo-spatial and temporal searches
August 2015
• Define, maintain, and enforce metadata schemas and Controlled Vocabularies
October 2015
• Revise and improve documentation for users and publishers
December 2015
• Define ESGF Virtual Organizations • Establish first government bodies

7.1.15 ESGF Provenance Capture

ESGF currently does not have the capability of true provenance capture. The review of our driving use cases demands however that we provide a number of services that will heavily rely on provenance information, these are: result explanation (how was this data set created), result comparison (where and how do these two results differ in the way they were created) and reproducibility (can we recreate these results). The latter requirement of reproducibility also offers an opportunity for training, if we can explain how results were created and let young researchers follow in the footsteps of their more experienced colleagues in the use of complex simulation and analysis tools.

Based on these observations it has been determined that to provide collaborators with the most complete picture of data set origin, provenance capture must encompass source code links, workflow history from any workflow engine, simulation history collected from log files, transaction records from data analysis tools and lineage

records that may be associated with observational data. Therefore, the future ESGF provenance component must provide services to capture provenance from difference native provenance sources, store provenance in a scalable and uniform way and link to native sources such as source code repositories so that collaborators can later leverage that information in cross-reference searching and data set verification.

In support of ESGF's provenance needs, we are looking to leverage the tool suite ProvEn, a comprehensive set of provenance services to support the extraction of provenance from heterogeneous resources, translation into interchange formats, and for providing the option of loading the transformed provenance into a registered data store. ProvEn is also extensible, allowing translated provenance to be correlated with other knowledge stores or databases. As mentioned earlier, translating provenance into a common form helps scientists browse, query, and infer from their scientific perspective. For ESGF, ProvEn will be extended with the results of the research on scalable provenance. ProvEn provides four primary provenance ingestion services, as follows:

1. ***Batch Service***. This service allows clients to send multiple heterogeneous provenance streams (e.g. workflow provenance, historical log, lineage) to ProvEn as one transaction. Batching provenance provides clients a translated uniform provenance result and allows the client to declare the group of provenance sources under one namespace that can be correlated directly back to corresponding data sets
2. The ***Extraction Service*** has a dual nature. From the native provenance perspective, any native provenance representation requires an API to be written and made available so that provenance submitted to the extraction service can be identified and extracted. From the interchange language perspective, the Extraction Service will provide a common API built around a specific interchange language so that each snippet of native provenance is properly translated in a common form.
3. The ***Interchange Language Translation Service*** is called by the Batch Service for each translation request and carries out the translation by interacting with the extraction service. The creation of optimized provenance interchange languages are needed optimize how provenance is conveyed when monitoring processes need to quickly assess soft errors at runtime, extending this concept bi-directional translators could provide a means to convey meaning in post-mortem analysis.
4. ***Load Service***. If the provenance is to be persisted, the Load Service is called after a Batch Service request returns translated provenance. The Load Service provides interfaces to different storage solutions to support the widest range of solutions for the user community.

In addition, ProvEn offers a scalable provenance store and a range of services to interact with the collected provenance, from basic provenance report preparation (how was this data set created) to more in-depth queries and provenance correlations. Overall ProvEn will offer through ESGF a range of 'canned' queries and the ability to directly interrogate the provenance store.

Table 15. *Expected ESGF provenance capture for 2015.*

Timeline of Milestones for ESGF provenance capture
April 2015
• Provide the provenance service and requirement document • Establish the many native forms of provenance relating to data sets coming from the many forms and sources (e.g., workflows, people or instruments recordings, software, etc.)
June 2015
• Each type of native provenance format must be identified by provenance language (if applicable), version number (or version number of the software)/or date time stamp • Original or native provenance needs to be immutable and retained as a digital object or a reference within the design
August 2015
• Decide on a common provenance format that is a community standard and possible Resource Description Framework based • Develop prototype for provenance instances that categorize the native file types, version, and indexes.

December 2015
• Demonstrate lightweight provenance component with verification services

7.2 UV-CDAT Technical Development

7.2.1 Distributed Analytics Application Programming Interface

The primary charge to the ESGF Compute Working Team (CWT) has been to allow ESGF users to execute analysis tools on high-end compute clusters, HPCs, cloud servers, and other forms of compute servers. This team will be designing and developing a general API that will allow users to interface into any prescribed back-end compute platform. The team will also develop a general API for ESGF's multiple analysis and visualization tools (UV-CDAT, Ferret, NCO, CDO, etc.). These efforts will allow ESGF's products and services to provide a range of data analysis and visualizations on many compute platforms. Depending on the compute platform, the ESGF compute software will select default characteristics for plotting and analyzing results. In addition, the ESGF node will be able to refine analysis and visualization characteristics according to user requests sent via analysis scripts. Access to the ESGF compute capabilities will be accessed through scripts, thick client graphical user interface widgets, and smart client Web applications.

The goal is for each national and international agency to help identify specific use cases to help with defining the needed APIs. This will also aid in addressing gaps needed for specific compute requirements for simulations and observational data holdings. As part of this missing critical assignment, the team will address the following questions:

- What priority analytics jobs needed by the community (i.e., CMIP, obs4MIPs, other MIPs, CORDEX, climate change impacts community, etc.)?
- Where will the data be located (i.e., local, remote, combination of the two)?
- What sort of requirements do jobs have in terms of memory usages and parallelism?
- What type of compute complexity can we expect if we use computing resources spread across the federation?
- How do we let only authorized users query, access, and use secured levels of computing resources?
- How to advertise dynamically available services along with capacity (scheduler to advertise computing nodes services)?

Before coding begins, we need to find the best technology available to fit ESGF community needs for server-side computing. With that, there are three layers we will consider when addressing our charge:

1. **API**: Well-defined interface to large-scale, distributed, data-proximal analytics and visualization capabilities for data accessible through ESGF. This API must allow for "operations" and "query capabilities" to be distributed across multiple ESGF sites and allow for the underlying analytic operations and systems to enable distributed analytics—for example, an average across multiple scenarios stored at multiple locations. In addition, the API should have the capability for adaptive construction of the underlying analytical operations to allow for more community engagement.
2. **Analytics Operations** (canonical operations): A set of analytical operations that can be accessible through the API for server-side and distributed analytics, such as subsetting, averaging, variation and anomaly calculations, etc. We should think of these operations as potentially an assembly language of operations upon which the community can create much higher level of operations that can be used to create workflows, stored for provenance as persistent services, and released back to the general community for use. Note that these operations may have different operations on different back-end compute platforms. Everything from MPI to MapReduce to databases must be explored.
3. **Compute Platform**: Back-end HPC platforms to allow for both server-side processing capabilities and the ability to perform distributed analytics. The back-end systems must therefore be able to not only compute local data, but to be able to provide the ability to quickly and efficiently pull analytical results

from other platforms to perform higher level analytical operations across multiple, distributed data sets. There will be a tendency for centers to gravitate to traditional high performance, tightly coupled computing systems, but we need to consider different types of options. These resources include HPC clusters, Hadoop, Object Stores, Cloud computing, etc.

In all cases, there will be costs associated to using distributed compute resources; therefore we must regulate secure access and availability. This means that designated ESGF projects, groups, users, etc. will have prioritized and restricted access levels to compute resources.

There are all kinds of dependencies, but we need to parallelize the effort as best as we can. If we break down the three layers, we can consider the following steps for each, which can be teased apart.

API

- Identification of use cases
- Identification of existing or similar APIs
- Requirements
- Definition
- Development

Analytics Operations

- Identification of Canonical Operations to be exposed to the API
- Identification of any existing or similar operations
- Requirements
- Definition
- Development

Compute Platforms

- Candidate platforms
- Requirements
- Proof of concepts
- Prototypes

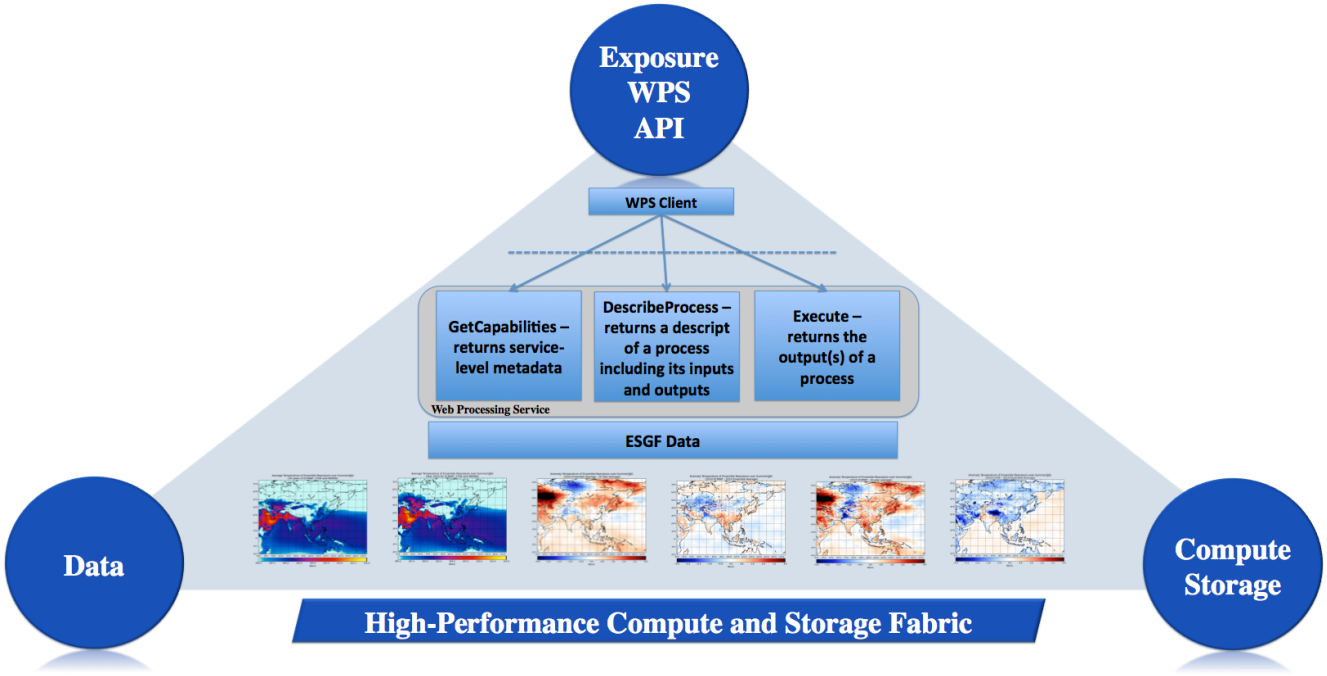

Figure 15. *The API exposes methods by which the analytics are transferred to where the data resides. The data is stored within a high-performance compute and storage fabric using a combination of HPC and cloud computing capabilities. The same data may be accessible through the ESGF along with a number of other web services.*

Table 16. *Expected ESGF-CWT milestones for 2015.*

Timeline of Milestones for ESGF-CWT
February - March 2015
• Finalize Web Processing Service (WPS) API definition for the prototype use case (anomaly) and simple canonical operations (such as, subsetting, average, maximum, minimum)
April - June 2015
• Definition of a standard set of climate data to be initially exposed • Definition of unit tests, data, input, and output to be used to verify the implementation of the WPS API • At least one proof-of-concept implementation of the WPS API at a single site (not distributed at this point) • First implementation of API and analysis for simple multi-model averaging use case
July – December 2015
• Testing of the proof of concept implementation of the WPS API using the unit tests and the climate data • Second proof-of-concept implementation of the WPS API at a second site using a different mechanism to perform the server side analytics • Expansion of use cases through conversations with potential end users and a prioritization of additional capabilities to be exposed by the API
January – June 2016
• Focus on federated analytics to extend the API to act upon data at two locations • Expansions of the capabilities within the API based on the priorities set within the science discussions
July – December 2016
• Expand and elevate the proof of concept to prototype at least the two initial locations • Continued expansion of the capabilities within the API

7.2.2 Ultrascale Climate Data Visualization and Analysis

UV-CDAT comes with its own visualization package, the Visualization Control System (VCS). VCS is tightly integrated with CDMS2 and understands CF metadata, making it one of the best tools for climate data visualization in terms of ease of use, graphics quality, and customization and features. It can understand and plot cdms2 read data "as is" without any user input.

During the upcoming year, UV-CDAT will support and enhance VCS 1D, 2D, and 3D capabilities, respond to bug reports and feature requests from UV-CDAT users, and complete the integration of the xgks-based VCS into VTK workflows. Animation will be further optimized. Unusual projections will be augmented in terms of quality and automation. An editor mode will be added, and geographic information system (GIS) capabilities will be added and possibly integrated with other data projects.

Table 17. *Expected UV-CDAT VCS 2D milestones for 2015.*

Timeline of Milestones for VCS
February 2015
• Optimized 2D animation using VTK workflow
April 2015
• Finish full re-implementation of XGKS-based VCS features via VTK back-end visualization package (e.g., isolines labels, streamlines, text rotation consistent across VTK output backend, Taylor diagrams, etc.)
June 2015
• Self-cleaning of climate data representing many projections
September 2015
• Integrate OpenGIS
December 2015
• GIS-based geo-data plotted together with climate data

7.2.2.1 VCS 2D

LLNL is developing an interactive layer for VCS. The purpose of this component is to give users access to the many advanced plotting capabilities of VCS without having to leave their plot. It will allow users to swap out projections on the fly, allow drag and drop positioning of existing secondary objects (markers/text areas/fill areas), place new secondary objects, edit appearance of secondary objects, change color map for plots and secondary objects, select colors visually for secondary objects, control animation of plots, reposition and resize plot elements, and save out custom templates for later use.

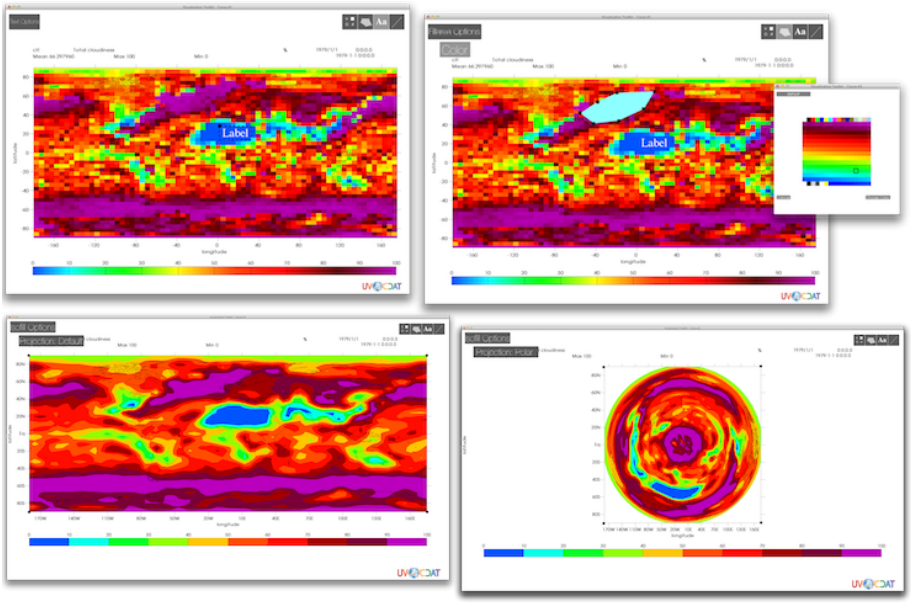

Figure 16. *VCS 2D interactivity examples.*

Adding these features will greatly enhance the flexibility of VCS both inside and outside of UV-CDAT; rather than requiring effort prior to plotting things in UV-CDAT (configuring custom templates, plots, and secondary objects), or requiring users to manually configure these things in the console, users will be able to directly configure them in place. It will reduce the effort required to place objects on a plot, and expose many of the more powerful features of VCS to users who may not have been aware of them. This will allow users to have an easier time of creating the exact plot they desire, highlighting the exact information they want to, and discover the insights they are looking for.

Table 18. *Expected UV-CDAT VCS 2D milestones for 2015.*

Timeline of Milestones for VCS 2D Plots
February 2015
• First implementation of interactivity layer. To be released with UV-CDAT version 2.2
April 2015
• Integrate into Web-based server-side application
July – December 2015
• Develop and integrate additional vcs2D plots and services

7.2.2.2 VCS 3D

The NASA Center for Climate Simulation (NCCS) is developing vcs3D, the 3D visualization component of VCS. This fully integrated UV-CDAT component leverages many UV-CDAT features to provide workflow interfaces, interactive 3D data exploration, hyperwall and stereo visualization, automated provenance generation, parallel task execution, and streaming data parallel pipelines. It enables exploratory analysis of diverse and rich data sets from various sources including ESGF, providing user-friendly workflow interfaces for advanced visualization and

analysis of climate data at a level appropriate for scientists. VCS3D's integration with CDAT's Climate Data Management System (CDMS) and other climate data analysis tools enables scientists to run analyses that were previously intractable due to the large size of the data sets, and using VCS3D, seamlessly couple these analyses with advanced visualization methods.

During the upcoming year, NCCS will support and enhance VCS3D by responding to bug reports and feature requests from UV-CDAT users. VCS3D will be integrated into the UV-CDAT web deployment, installed and made available on the NCCS science cloud, and integrated with the NCCS climate data analysis services.

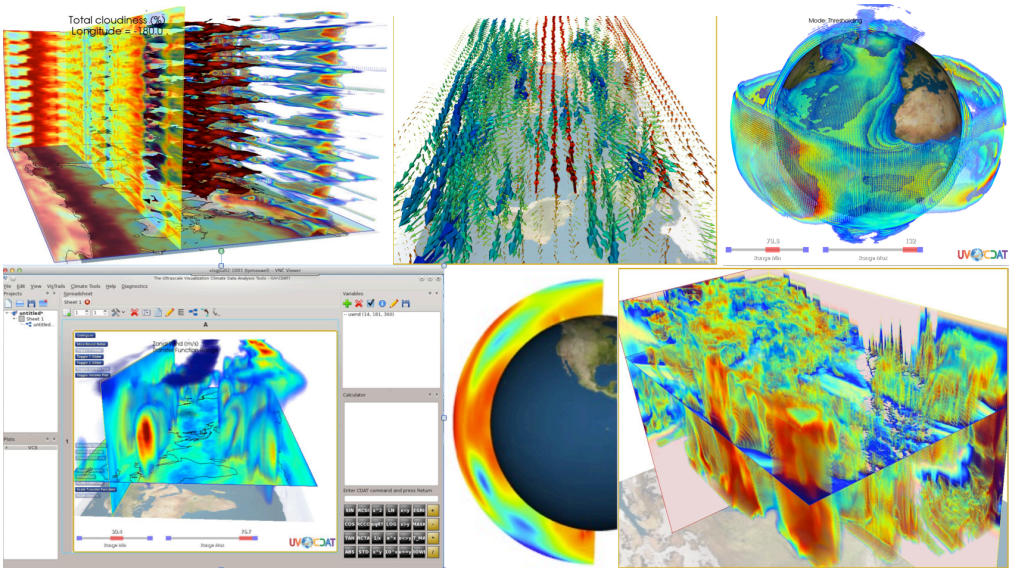

Figure 17. *Sampling of Vcs3D plots.*

Table 19. *Expected UV-CDAT VCS 3D milestones for 2015.*

Timeline of Milestones for VCS 3D Plots
June 2015
• Update version of VCS3D will be released with each public release of the UV-CDAT desktop • Integrate into Web-based server-side application
July 2015
• Integrate VCS3D into the NCCS science cloud
August 2015
• Included additional 3D plots and services

7.2.2.3 CDATWeb

The overreaching objecting of CDATWeb is to bring UV-CDAT analysis and visualization capabilities to the web environment. UV-CDAT on the web will bring about an installation-free, collaborative environment for climate scientists. Not only will it facilitate data sharing, it will ultimately speed up climate science research. We initially developed a prototype of the CDATWeb that uses the ParaViewWeb as the underlying system, to bring VCS2D and VCS 3D visualization in a web browser. The prototype we developed also demonstrated integration with diagnostics.

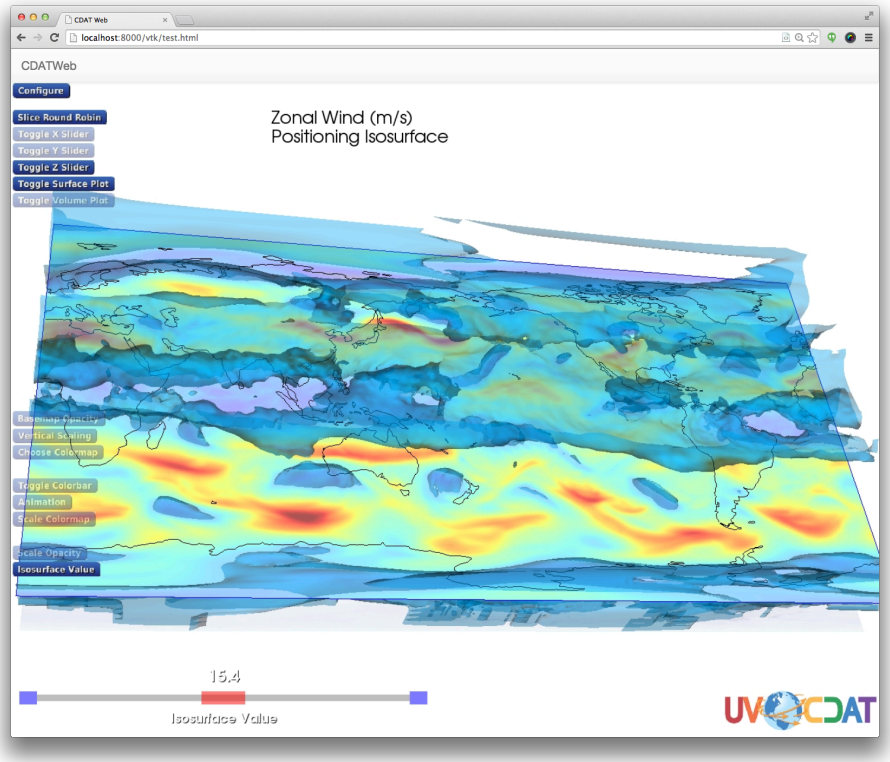

Figure 18. *CDATWeb instance showing interactive VCD 3D visualization in a Web browser.*

During the upcoming year, we will enhance and improve the CDATWeb, making it production-quality software that will be released with the UV-CDAT releases. Some of the work will include improvements to the user interface, better handling of back-end process launching, and RESTful API for creating and shutting down visualization sessions.

Table 20. *Expected CDATWeb milestones for 2015.*

Timeline of Milestones for CDATWeb
March 2015
• Streamlined launching of visualization process • Support for interactions via mouse and keyboard
May 2015
• CDATWeb release 0.1 with documentation • Multi-view support
August 2015
• Integration with diagnostics and other components • First official release of CDATWeb 0.2 with UV-CDAT • Begin integration with sever-side analysis
December 2015
• CDATWeb 0.2 release with cloud and HPC support

7.2.3 UVCDAT Diagnostics and Metrics, and Backend Workflow Support

UV-CDAT includes a diagnostics and metrics package, which is a post-processing system for climate model output. Typically output of a climate model is compared with observations or output of another model. The user chooses the variables to consider, as well as other parameters such as a season or region of interest. The diagnostics package will compute these variables from the model or observation data, and plot them. This

computation may be a simple dimensionality reduction or an involved calculation starting with several output variables. The diagnostics may be run interactively in the UV-CDAT GUI, or from a script, e.g., as part of a larger workflow system.

In addition to the diagnostics computational package, there are several support scripts to facilitate scientists' use of the diagnostics. One script uses the CDAT library to compute climatologies, which can then be used by the diagnostics or for further interactive visualization. A second script generates individual diagnostic plots based on user-specified parameters. A third script generates a series of diagnostic plots meant to duplicate a series of plot sets that are familiar to climate scientists. Additionally, this script is used to generate new plot sets moving forward. The design of the script was influenced by user feedback and should enable scientists to quickly add additional capabilities to the workflow. Last, we will work on a web-based GUI throughout the year, as time permits. Most software will run on a server, so the user will need no more than a web browser.

In 2015, we plan to complete a basic implementation (including the most interesting variables) of all types of atmosphere and land plotting, which were done in an older, NCL-based diagnostics package familiar to experienced users. Even at this point, our package's flexibility will let it support a much wider range of variables and parameters than any older diagnostics package. Ocean diagnostics may take longer to develop, but a prototype should be available at the end of the first quarter of 2015, and some usable diagnostics by the end of the year.

Throughout the year we will concentrate our efforts on integrating the diagnostics system with the ACME workflow system, computing more variables, and adding other capabilities requested by model developers and climate scientists. We will continue our program of guiding development through testing by scientists with real-world problems.

In particular, we will make sure to supply the additional Tier 1b and 1c diagnostics as the ACME atmosphere group develops their specifications. Additionally, we will focus on performance improvements, which have already been requested.

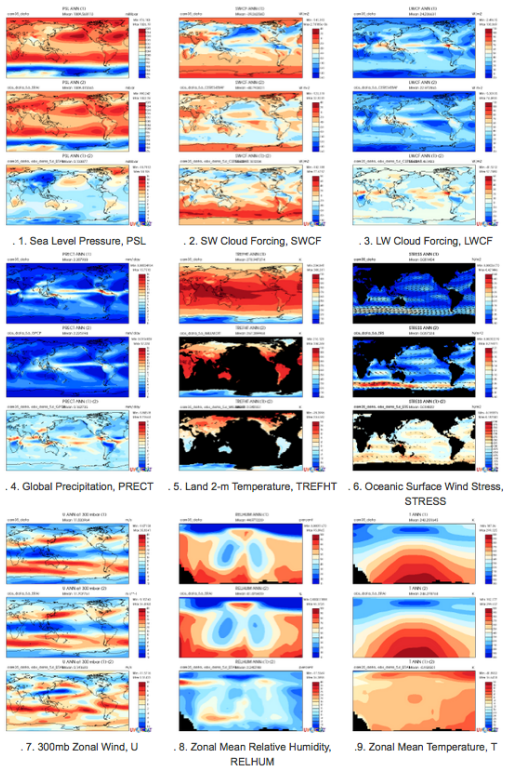

Figure 19. *Atmosphere Tier 1a Diagnostics/Metrics example plots.*

Table 21. *Expected diagnostics and metrics milestones for 2015.*

Timeline of Milestones for ESGF Search Services
June 2015
• All existing atmosphere and land plot set supported • Minimal set of ocean diagnostics supported • CMIP metrics supported for model intercomparison
July 2015
• GUI-based diagnostics and metrics supported
December 2015
• Web-based UI implementation for diagnostics and metrics • Integrated into ESGF CoG UI interface
December 2015
• Significant performance improvements for climatology and diagnostic plot generation

7.2.4 Exploratory Analysis and Web Informatics Infrastructure

The traditional Community Earth System Model (CESM) diagnostics package provides scientists and modelers a way to quickly analyze the effectiveness of a climate model by computing climatological means of the large-scale simulations and producing hundreds of plots and tables in a variety of formats. While the modeling community has successfully utilized this toolkit for a number of years, it is incompatible with modern workflows and methodologies.

Over the past year, Oak Ridge National Laboratory has unveiled the "classic" diagnostics viewer, a novel technology that presents an improved interface for diagnostics. The classic viewer is a key component of the Exploratory Analysis toolkit in the ACME post-processing workflow, which utilizes UV-CDAT for efficient computation of key metrics and the ESGF for data archiving and cataloging. Specifically, we have provided the classic viewer with some base capabilities, which include:

- Harnessing the ESGF user authentication software for access restriction;
- Navigable views for both land and atmosphere diagnostics figures;
- Ability to scan figures quickly; and
- The ability to populate the viewer with easy to use scripts in the post-processing stages of the model workflow.

There are several features that will be added to the classic viewer (and the Exploratory Analysis Toolkit in general) in the coming year. First, the workflow for generating figures will be further refined and streamlined, including fast transfers for both the climatology files and raw data output files from models. Second, using the first stage as a prerequisite, we will add support for ESGF publication and file download directly from the viewer itself. Third, we will improve the user experience by utilizing some of the recommendations by the user community. An example is the ability to see figures side by side for direct comparison. Finally, we plan to append the other views (diagnostics tree, heatmap, etc.) to the framework.

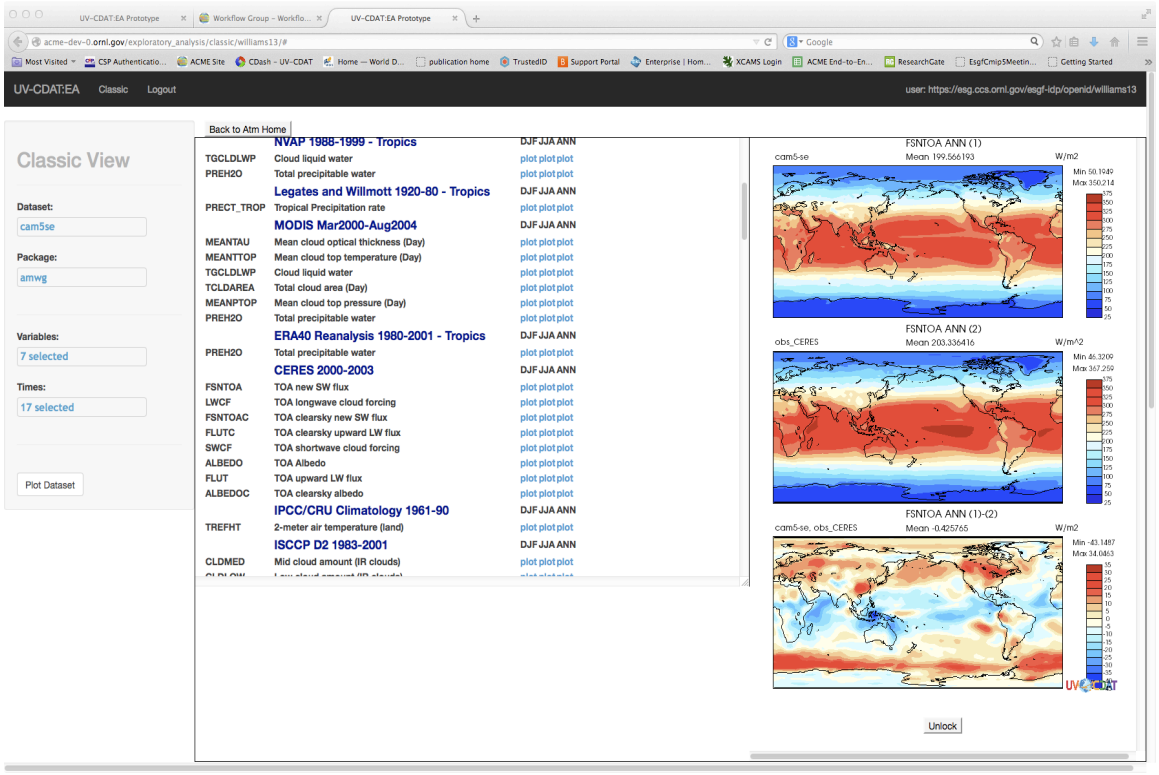

Figure 20. *Classic viewer to display diagnostics and metrics output via Web browsers.*

Table 22. *Expected diagnostics and metrics milestones for 2015.*

Timeline of Milestones for Classic Viewer
February 2015
• Finish all atmosphere diagnostics sets • Download capability for figures and climatologies
June 2015
• First release of tree viewer
October 2015
• First release of heatmap

7.2.5 Remote Data Processing and Exploration

The size of climate data sets is already large enough to be a burden that impedes analysis tasks due to the storage space, computing power, and network bandwidth available to the client. The next generation of data sets will be even larger, so developing methods to facilitate remote analysis are necessary for small and medium institutions to continue to be able to participate in comparative climate study. The ViSUS Server provides coarse-to-fine remote streaming of even disparately located data sets. It is based on the hierarchical z-order IDX format, a hierarchical data reordering that preserves spatial locality without increasing the original data size.

Utilizing ViSUS streaming facilitates exploratory data analysis and visualization of remote data sets, as well as providing dynamic processing functionality. This enables rapid, interactive exploration of specific data without requiring cumbersome downloads of often irrelevant variables or costly computation. ViSUS can be used to efficiently explore various combinations of experimental runs and analyses in order to select the most effective combination of data to be used later in a more comprehensive offline analysis.

Since ViSUS relies on data in the IDX format, a project to provide on demand data reordering of requested data volumes to IDX is currently underway. This system will provide a suitably sized cache of converted data on the

ESGF nodes, as well as a ViSUS server. The initial request for a new volume variable or time step will trigger an immediate conversion of that portion of the data only. The conversion of only a small portion of data at one time will be fast for the initial user, and successive users will experience no delay at all as a benefit of the cached data.

The implementation will proceed in the following steps. A new button will be added to the ESGF search page that request the data to be provided in the ViSUS format. Requesting data in this format will trigger in the background the creation of an empty IDX container for the selected data set. This relies on the CDMS xml volume description. Though construction of this metadata can be time-consuming, it is anticipated that providing these descriptions for all volumes will be advantageous for other uses besides data reordering. The new volume will then be included in the ViSUS server. When a client requests some particular variable and time step, that specific request is converted on the fly and sent immediately to the client. The granularity of this conversion can even be reduced to a specific region of data.

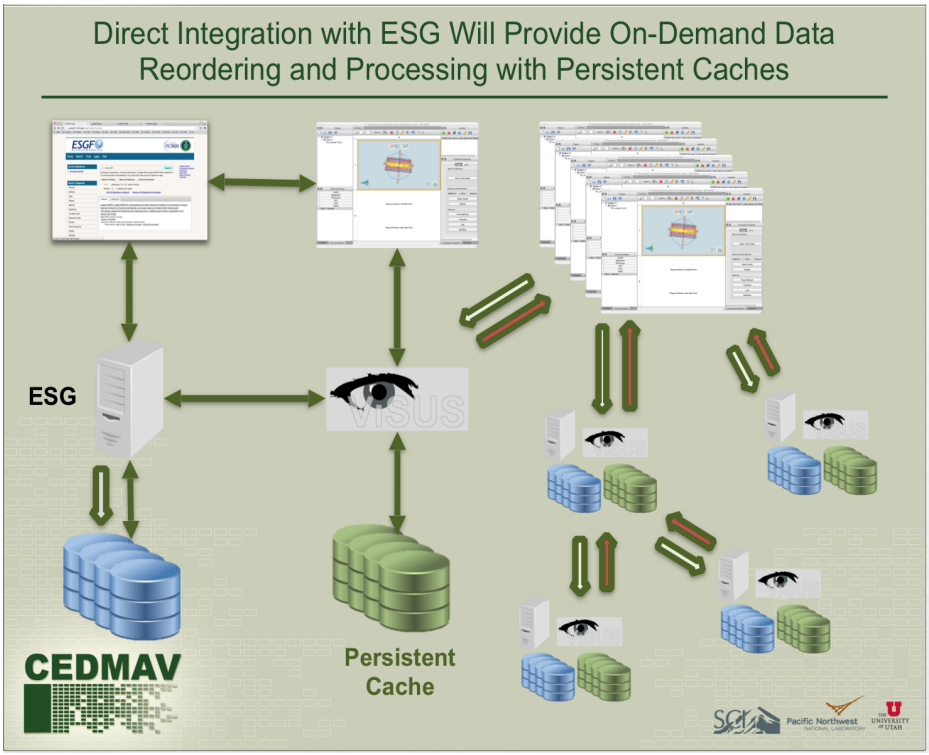

Figure 21. *Figure showing the visual exploration speedup using ViSUS.*

Table 23. *Expected diagnostics and metrics milestones for 2015.*

Timeline of Milestones for Classic Viewer
February 2015
• Enable ViSUS Data Access through LLNL's ESGF Node
April 2015
• Dynamic data conversion and storage as streaming IDX
November 2015
• Data Management

7.2.6 LAS-ESGF Analysis and Visualization

The LAS-ESGF analysis and visualization system is a connection between the power and user friendliness of the Live Access Server (LAS) and the ESGF data search and distribution system. By making this connection we allow users of ESGF the ability to visualize and compare ESGF data sets, perform basic analyses, and in some

instances download specialized subsets of the data. During the coming year. we have planned three phases of improvements to the LAS-ESGF integration.

The first improvement will be the integration of the latest LAS user interface into the LAS-ESGF distribution. This user interface contains many usability improvements, especially related to the controls when making comparison plots between data sets. This phase of the integration will also include improvements to the selection of multiple variables from the data cart provided by the Django-based ESGF user interface. This will also allow us to eliminate the duplicate search UI widgets in the LAS-ESGF interface itself.

The second phase of the improved integration will provide LAS-ESGF access to curvilinear 2D horizontal data grids commonly used in ocean models, such as the tri-polar grid. This improvement will be phased in with improvements to the publisher to provide better metadata about the latitude and longitude extents of the curvilinear grids.

The third planned integration will be implementation of the CWT web-processing service protocol for computations. In the initial phase, we will implement the first use-case identified by the compute working team by building bridge software that will allow us to take in the WPS request and create the necessary scripts and commands to fulfill the request using our existing server-side computation engine the Ferret-THREDDS Data Server.

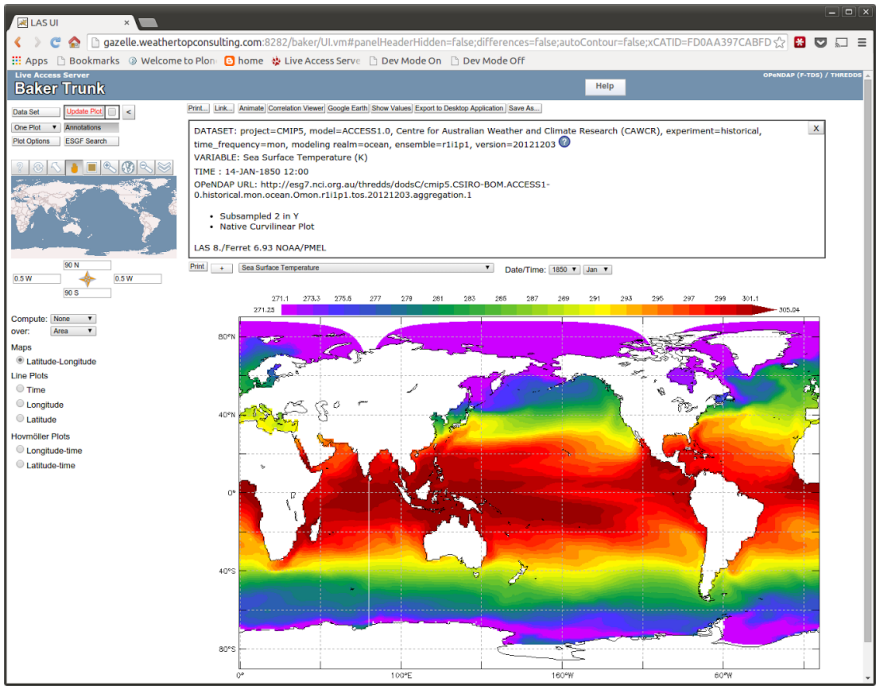

Figure 22. *This image shows a native grid plot of an ocean realm tri-polar grid data set from Centre for Australian Weather and Climate Research. Plots on these curvilinear grids will be available after the next phase of the ESGF-LAS integration.*

Table 24. *Expected LAS analysis and visualization milestones for 2015.*

Timeline of Milestones for ESGF Search Services
February 2015
• Working with the ESGF-CWT, integration of the new user interface that is under development
May 2015
• Working with the ESGF-PWT, integration of curvilinear grids using new publisher metadata
October 2015
• Working with the ESGF-CWT, implementation of WPS server-side analysis use case

7.2.7 Enabling GIS Operations and Visualization in UV-CDAT using OpenClimateGIS

The overall goal of this calendar year's technical development will be to establish proof-of-concept visualization of GIS file formats (i.e. ESRI Shapefile, GeoJSON) in UV-CDAT to identify appropriate development pathways for enabling additional GIS capabilities. Example capabilities include spatial subsetting with arbitrary geometric boundaries, such as a watershed and writing to common GIS formats such as ESRI Shapefile. Incorporating GIS into UV-CDAT will provide it with a set of unique capabilities for localized climate analysis allowing users to more easily operate on areas-of-interest suitable to their analysis domains. The Python-based software OpenClimateGIS will provide the GIS API for UV-CDAT.

OpenClimateGIS is a NOAA-funded open source Python software library designed for geospatial manipulation, subsetting, computation, and translation of climate data sets stored in local netCDF-CF files or files served remotely via OPeNDAP. The software is optimized to work with high-dimensional climate data sets allowing a number of common GIS operations (e.g. intersects, clip, spatial averaging, coordinate system transformations) to be performed on a file storage format notoriously difficult to manipulate in modern GIS software. OpenClimateGIS shares a number of structural similarities to UV-CDAT, most importantly the Python programming language and NumPy-based array storage. These similarities will simplify the integration of OpenClimateGIS's geospatial functionality with the UV-CDAT library.

This year's development tasks will continue the integration plan outlined in calendar year 2014. The next OpenClimateGIS release (v1.1), scheduled for the February 2015 timeframe, will include functionality to support the April 2014 milestone. The April 2014 milestone described work-related to adding ESRI Shapefile as a data format equivalent to netCDF-CF in its ability to be read and altered by OpenClimateGIS. It also included demonstration code in UV-CDAT using the new OpenClimateGIS shapefile API. We successfully brought OpenClimateGIS into the UV-CDAT build system (March 2014 milestone) last year. In summary, the March 2014 and a portion of the April 2014 milestone were completed. Milestones following April 2014, which included development required to achieve an alpha version of GIS capabilities in UV-CDAT (i.e. implementation plan, coding, testing), were not accomplished.

The milestones for 2015 represent a focusing of the timeline for bringing GIS capabilities into UV-CDAT. Many of the planned milestones in 2014 were not accomplished because the development demands on both OpenClimateGIS and UV-CDAT proved too ambitious in the end. The narrowing of goals for 2015 is intended to identify a succinct set of tasks that will move integration forward without severely interrupting other software priorities.

Technical development this year will focus primarily on the interoperability layer in OpenClimateGIS for reading, writing, and converting GIS files into visualization formats acceptable to UV-CDAT (e.g. VTK). It is assumed that a visualization format suitable for UV-CDAT may also be accessed for information such as variables, dimensions, and geometries and may be retroactively used as a data source for input to GIS operations. This requires the UV-CDAT–compatible format be a fully qualified data object in OpenClimateGIS that may be used, for example, to subset other geospatial data (i.e. netCDF-CF). There are two main tasks required to accomplish the 2015 development milestones:

1. Correspondence with the UV-CDAT development team to identify an appropriate visualization and analysis data structure to use in UV-CDAT for interacting with GIS-based data formats.
2. Implement data access routines in OpenClimateGIS for the identified format followed by demonstration code in UV-CDAT to, at the very least, visualize the format through the UV-CDAT GUI.

With regard to interoperability with geospatial data formats, it is also important to mention the integration of OpenClimateGIS with an additional package used by UV-CDAT called ESMPy. ESMPy is the Python interface to the ESMF regridding framework and is the default grid-remapping package used by UV-CDAT. This integration work has focused on simplifying the conversion of geospatial data into ESMPy data objects including the UGRID convention for representing unstructured grids. Initially the two-dimensional flexible mesh format has been targeted, as it is ideal for storing geospatial data such as high-resolution watershed catchments. The goal of

this development has been to simplify access to high performance conservative regridding operations exposed in the ESMPy interface.

Table 25. *Expected GIS milestones for 2015.*

Timeline of Milestones for ESGF Search Services
February 2015
• ESRI shapefile data interoperability within OpenClimateGIS providing the Framework for conversion to VTK and UGRID
June 2015
• Identify appropriate visualization and analytic data format for input into UV-CDAT; this will likely be some form of VTK or UGRID file leveraging existing APIs where possible
November 2015
• Provide proof-of-concept GIS data visualization in UV-CDAT

8 Community Developments

8.1 ES-DOC and Controlled Vocabulary

Earth System Documentation (ES-DOC) (http://es-doc.org) is an international project supplying tools, APIs, and consultancy in support of earth system documentation creation, analysis, and dissemination. It nurtures a sustainable, standards-based, documentation ecosystem that aims to be integrated into exascale data set archives, for instance, the ESGF system. At the core of this ecosystem is a set of documentation tools that support documentation creation, search, viewing, comparison, and visualization. All tools are built on top of the ES-DOC API. Documentation creation is supported via both a command line client and an online questionnaire.

The ESGF system already contains links to the ES-DOC documentation viewer from within its CoG search front end. ES-DOC will collaborate with ESGF in the efforts to design, develop, and deploy a robust set of simple controlled vocabulary services. Such services will be leverageable by various components within the ESGF stack. The set of controlled vocabularies include earth system model component hierarchies for the various MIPs, as well as DRS terms, CF names, and the like.

ES-DOC will also exploit the faceted search capabilities of the ESGF search API in order to build graphical archive visualizations that help users better understand the archive from multiple viewpoints.

Table 26. *Expected ES-DOC and Controlled Vocabulary milestones for 2015.*

Timeline of Milestones for ESGF Search Services
January 2015
• pyesdoc validation and publication
February 2015
• EU workshop
March 2015
• Simple controlled vocabulary service design
• Questionnaire development sign-off
May 2015
• Simulation comparator
September 2015
• Extend viewer to support PDF formatting
October 2015
• Build visualization data sets ESGF search facets
November 2015
• Full support for CMIP6 simulation documentation

8.2 CF Conventions

The CF Conventions are a set of conventions for the organization of self-describing files containing climate and other scientific data. The conventions describe metadata and a standard vocabulary, which are used to describe how the data itself is organized and what it means. These conventions are an essential prerequisite for automated interpretation and post-processing of climate data and hence are an important tool for scientific sharing and collaboration. All data on ESGF is expected to follow the CF Conventions. The CF Conventions allow for significant flexibility. Other conventions, such as the CMIP5 "data reference syntax," add more requirements so that automated postprocessors can do still more. An international group of CF Conventions users develop additional features as needs arise.

The CF Conventions are defined in a document on a website. The website offers additional documents, such as a "conformance document" that describes how an automated system can check a file for conformance with the conventions. In the last year, the computer hosting the web site had a major hardware failure, necessitating the construction of a new site. Further, the "issue tracker" used to develop and plan changes to the Conventions had to be migrated to a new host computer on two occasions. The whole system is up and running again.

Last year, we began working on a new version of the CF Conventions. Now that the web site and issue tracker are stable, we will finish this version, 1.7, and begin working on Version 1.8.

Table 27. Expected CF milestones for 2015.

Timeline of Milestones for ESGF Search Services
February 2015
• CF Convention session at the 2015 GO-ESSP meeting
June 2015
• Release CF Conventions 1.7
December 2015
• Release CV Conventions 1.8

8.3 Preparing CMOR for CMIP6 and other WCRP Projects

CMOR2 was written and in some cases hard-wired to meet CMIP5 modeling needs. In order to prepare CMOR for CMIP6, we need to upgrade it to Version 3. In this version, the following should be addressed:

- Some of the attributes written by CMOR do not apply to observations (e.g., model name, experiment name).
- The input tables for each new MIP need to be simplified and reconfigured. The input table needs to be modularized by separating project-specific metadata from variable information.

Once the code has been reformatted, the following should occur:

- Generate new CMOR tables containing additional required global attributes and recognizing CMIP6 controlled vocabulary (or any project actually).
- Provide more complete QC information to CMOR (e.g., valid max and min for each field) in these tables.

In the coming year, CMOR plans to first integrate some existing development on GitHub, implement the code changes required for CMIP3, and apply them to CMIP5 as a first cut. Then we will need to create the CMIP6/CMOR3 tables; this will probably require more code tweaking, as special cases will show up. After an initial phase of testing with "friendly users," most of the summer will be dedicated to early/beta testing, with a release planned for Fall 2015 followed by a period of tutoring for new and migrating users.

Table 28. Expected CMOR milestones for 2015.

Timeline of Milestones for ESGF Search Services
February 2015
• Cleanup existing branches on GitHub
• Layout clearly with changes what are to be addressed for CMIP6
April 2015

• Implement changes and update CMIP5 tables to reflect the new structure for CMIP6 • Create CMIP6 alpha tables • Test with friendly users
May 2015
• Create CMIP6 beta tables • Test with friendly users
June 2015
• Official CMOR beta release • Continue to test beta release
September 2015
• Official release CMOR 3.0.0
October – December 2015
• Create CMOR user support • Create documentation

8.4 Ophidia: A Big Data Analytics Framework for eScience

Ophidia is a research effort addressing Big Data challenges in the scientific domain. It executes data-intensive analysis and input/output (I/O) exploiting advanced parallel computing techniques and smart data distribution methods. It includes the Parallel Data Analysis Service (PDAS), which provides parallel I/O and data analysis functionalities, an array-based storage model, and a hierarchical storage organization to partition and distribute multidimensional scientific data sets. Since the storage model does not rely on any scientific data set file format, it can be exploited in different scientific domains and with very heterogeneous sets of data.

PDAS comes with an extensive set of primitives to operate on n-dimensional arrays (i.e. on the arrays contained in fragments). To achieve flexibility requirements, primitives are designed as dynamic libraries in order to be plugged in different I/O servers with no effort. Furthermore, plugins can also be nested into each other to enable a more complex task. Currently available array-based functions allow data subsetting, data aggregation (i.e. maximum, minimum, average), array concatenation, algebraic expressions, predicate evaluation, and compression routines (i.e. zlib, xz, lzo). Core functions of well-known numerical libraries (e.g., GSL, PETSc) and tools have been included into the primitives.

The PDAS also provides many parallel operators to work on a whole data cube—that is, on all fragments associated with a data cube. Some examples include data cube subsetting (slicing and dicing), data cube aggregation, array-based primitives at the data cube level, data cube duplication, data cube pivoting, and netCDF file import and export. Other operators can also handle more than one data cube at time allowing intercomparison.

The most relevant data analytics use cases implemented in national and international projects with CF convention compliant CMIP5 data in netCDF format target fire danger prevention, sea situational awareness, interactions between climate change and biodiversity, climate indicators and remote data analysis, and large-scale data analytics.

A design study to include Ophidia as a compute engine into the ESGF middleware is being performed. The most relevant activity in 2015 will be related to the involvement into the ESGF CWT to implement the needed interface for data processing/analysis in the ESGF. Security aspects will be also analyzed and dispatched during the year. Interoperability will represent a major requirement to be addressed through the adoption of standards at different levels (e.g., security, server interface, scientific data formats, and related conventions). A key point will also relate to a substantial refactoring related to the metadata management support.

Table 29. Expected Ophidia milestones for 2015.

Timeline of Milestones for Ophidia Search Services
March 2015
• Software release as RPM and VMs (it includes a new metadata management support) • Analysis of requirements to include Ophidia as compute engine into ESGF

June 2015
• Prototype implementation of the compute node interface (e.g., WPS) based on ESGF-CWT recommendations
October 2015
• Alpha release-level implementation of the server interface (e.g., WPS) based on ESGF-CWT recommendations
December 2015
• Production-level release compliant with ESGF-CWT recommendations • First release available for inclusion into the ESGF software stack

9 Planned Development and Integration for Projects' Success

As important as all the projects described in Section 4 are to the climate research community, equally important is ensuring that the data ecosystem and infrastructure is optimally designed to enable the development of new, validated, and verified capabilities with proven technology. The ecosystem environment will provide the data and computing infrastructure for rapid development and assessment of new scientific modules and provide a testing-to-production environment for simulation and evaluation (i.e., metrics, diagnosis, and intercomparison) with observational and reanalysis data. Development of the necessary software tools to accomplish this is driven by the scientific requirements and a diverse set of climate use cases to develop and use the overall enterprise and the individual components as standalone systems. While some of the tools are specific to a particular project, wherever possible the development teams have identified common methods and similar APIs across component models and have coupled systems together to foster synergistic developments that satisfy the requirements of many projects.

To achieve the goals of individual projects and the community, the team will continue to build upon and enforce standards and promote the sharing of resources, such as in the case of netCDF, CF conventions, ESGF, UV-CDAT, ES-DOC, DRS, Globus, and many others. These open-source projects have growing recognition and use in segments of the research community, and the tools and experience resulting from these sponsored projects will provide the foundation on which the data ecosystem infrastructure will be based. By building upon and integrating these existing technologies, open standards, and community expertise, we are building a unique complete yet flexible framework suitable for supporting model development and experimental requirements, such as integrated data dissemination, workflow and provenance, analysis and visualization, and automated testing and evaluation in secure environment.

10 Presentation Abstracts

As a community, we are building an integrated data ecosystem and workflow as shown in **Figure 1**. The presentations presented at the conference cover a wide span of complex data generating systems ranging from LIDAR detectors, satellites, high-performance computers, etc. Together, we are feeding these large-scale data into the collective ESGF data management and dissemination system to be analyzed where the data are located. We eventually want to do provenance to repeat and reproduce workflow capturing by the community at large. In addition, the presentations reflect the need to capture hardware and network resources for compatible and integrated use. The ultimate goal is get data and knowledge information in hands of humans so they can make intelligent decisions on climate change. Analytical modeling of the entire ecosystem is underway which will help predict how long it takes to do various parts of the end-to-end data workflow. These along with many other ecosystem components were presented at the conference. Below are the presentation abstracts along with presenter(s) information. All presentations are available on the ESGF and UV-CDAT websites:

- http://esgf.llnl.gov/facetoface.html
- http://uv-cdat.llnl.gov/facetoface.html

Day 1: Tuesday, 9 December 2014
User Feedback and Project Requirements

Session Two: ESGF Governance

Title and Presenter	Abstract
ESGF Governance ***Justin Hnilo (DOE BER CESD, Justin.Hnilo@science.doe.gov)***	Although initiated by DOE, ESGF has become a multi-agency, international collaboration critical to the success of archiving, delivering, and analyzing well-known climate data, including the Working Group on Coupled Modeling's CMIP data used for the Intergovernmental Panel on Climate Change assessment reports. Now that the ESGF infrastructure has been adopted by many other projects and supporting CMIP activities, there is a need to establish a more formal governance structure that ensures resource implications for users of disparate data as well as funding agencies. The presentation and discussion will touch upon the status of the ESGF governance in respect to the current funding agencies: DOE, EU Commissions, NASA, NOAA, and the Australian National Computational Infrastructure..

Session Three: Project Feedback and Requirements

Title and Presenter	Abstract
Significance of the User Support Process in ESGF **(Session's Keynote)** ***Hashim Chunpir (DKRZ Hamburg, chunpir@dkrz.de)***	ESGF infrastructure is a collection of technologies, systems, people, policies, practices, processes, and relationships that interact with each other in a federated environment. The ESGF infrastructure came into being as a result of a need, and the need was to fulfill data-driven climate community research. As the ESGF reached its production level and became a commodity, it started serving diverse user communities, particularly from the domain of Climate Science. The key to envisioning a usable and a healthy e-infrastructure today is to streamline the user support process and to try to address the communication and standardization challenges that arise between the stakeholders of the ESGF. This talk will briefly describe the current user support scenario for ESGF based on the empirical findings collected in the past and highlight the challenges that we face today that must be addressed in future. Addressing user and usability issues within the ESGF infrastructure will not only attract the user communities but also will provide user satisfaction, cost efficiency, sustainability of e-infrastructure, and promote innovation and development of the infrastructure, resulting in a boom in data-driven e-research.
WGCM Infrastructure Panel Requirement ***Karl Taylor (DOE/LLNL, taylor13@llnl.gov)***	The newly formed WGCM Infrastructure Panel (WIP) is charged with attempting to articulate and set priorities for what is needed in the way of climate modeling infrastructure from the perspective of the modeling groups and the WGCM. The goal is to establish standards and policies for sharing climate model output that ensure consistency across WGCM activities. The WIP is currently developing four white papers describing key aspects of the infrastructure needed to support CMIP6, and these will be discussed with a focus on those components of particular interest to the ESGF community.
CMIP6 and MIPs ***Karl Taylor (DOE/LLNL, taylor13@llnl.gov)***	CMIP has been restructured to enhance its scientific impact while reducing the burden placed on modeling centers and the modeling infrastructure. CMIP now calls for an ongoing small set of benchmark Diagnosis, Evaluation, and Characterization of Klime Experiments (DECK) that will be performed as part of the model development cycle at each modeling centers. Building on these, CMIP6 will offer a smorgasbord of additional MIPs that address specific science questions. Modeling groups will be free to tailor their participation in CMIP6 according to their resource limits and scientific interests. In the context of the new CMIP6 design, implications for modeling infrastructure will be discussed.
High–End Computing Program Manager and Weather Focus Area Program Scientist ***Tsendgar Lee (NASA HQ, tsengdar.j.lee@nasa.gov)***	Earth system scientists are facing significant data analysis challenges as observation and model output data become bigger, which is caused by the spatial, temporal, and spectral resolutions becoming higher and the data records becoming longer. The information systems developed and architected in the past will need to be upgraded and enhanced to face the Big Data challenges. In this talk, we will discuss how the space agencies are confronting the challenges by putting together a next-generation system architecture that couples data, compute, storage, and tools. An open-source strategy has been established to enable open development and open collaboration.

Title and Presenter	Abstract
Collaborative Reanalysis Technical Environment-Intercomparison Project and ana4MIPs: Enhancing Access to Reanalysis Using ESGF ***Jerry Potter (NASA/GSFC, gerald.potter@nasa.gov)***	NASA/GSFC is gathering gridded reanalysis data from major weather forecast centers around the world and saving them side by side on ESGF to better understand the patterns responsible for such phenomena as heat waves, droughts, and floods and ultimately improve climate model predictions. The Collaborative Reanalysis Technical Environment-Intercomparison Project has created a repository of reanalyses (essentially re-forecasts of past weather using the latest forecast models) that can help improve weather and climate forecasts by studying the differences and similarities among various reanalysis efforts. Participating organizations include NASA, NOAA's National Centers for Environmental Prediction and Earth System Research Laboratory, the European Centre for Medium-Range Weather Forecasts, and the Japan Meteorological Agency. In addition to distributing the data on ESGF, NASA's Climate Model Data Services group is partnering with the NASA Center for Climate Simulation to develop tools that will bring to bear the massive computing power of the latest supercomputers along with large data storage facilities to make analysis and comparison of complex model output faster and more efficient for climate scientists.
ESGF Functionality Needed for obs4MIPs and Other Data Sets ***Robert Ferraro (NASA/JPL, robert.d.ferraro@jpl.nasa.gov)***	Obs4MIPs is delivering satellite observations formatted in the same manner as CMIP5 model outputs via ESGF, but there are differences between the observations' attributes (metadata) and the model attributes. These differences do not mesh well with either the DRS or the standard search capabilities, and the agencies that support these data sets are asking for reporting of usage statistics. As obs4MIPs grows and other projects like ana4MIPs are added, some evolution of the ESGF search and data delivery capabilities will be needed to support a more diverse user base and a wider variety of file types.
CORDEX ***Sébastien Denvil (IPSL/IS-ENES2, sebastien.denvil@ipsl.jussieu.fr)***	GCMs are at the basis of climate change science and of the provision of information to decision-makers and a large range of users. Within Europe, the European Network for Earth System Modeling (ENES) gathers together the European climate/Earth system modeling community, which is working on understanding and prediction of future climate change. ENES, through IS-ENES (phases 1 and 2), promotes the development of a common distributed modeling research infrastructure in Europe in order to facilitate the development and exploitation of climate models and better fulfill the societal needs with regards to climate change issues. IS-ENES2 gathers 18 partners from 10 European countries and includes the 6 main European GCMs. IS-ENES combines expertise in climate Earth system modeling , computational science, and climate change impact studies. This talk will highlight efforts undertaken under the IS-ENES2 coordination to support the dissemination of CORDEX on the ESGF. In many aspects, this efforts prefigure what will be done for CMIP6.
The climate4impact Portal: Bridging the CMIP5 and CORDEX Data Infrastructure to Impact Users ***Maarten Plieger (KNMI/IS-ENES, plieger@knmi.nl), Wim Som de Cerff, Christian Page, Natalia Tatarinova, Ronald Hutjes, Fokke de Jong, Lars Barring, and Elin Sjökvist***	The aim of climate4impact is to enhance the use of climate research data and to enhance the interaction with climate effect/impact communities. The portal is based on 17 impact use cases from five different European countries and is evaluated by a user panel consisting of use case owners. It has been developed within the European projects IS-ENES and IS-ENES2 for more than five years, and its development currently continues within IS-ENES2. The focus is mainly on the scientific impact community due to the community's breadth. This work has resulted in the ENES portal interface for climate impact communities and can be visited at www.climate4impact.eu. The climate4impact is connected to ESGF nodes containing GCM data from CMIP5 and regional climate model data from the Coordinated Regional Climate Downscaling Experiment. This global network of climate model data centers offers services for data description, discovery, and download. The climate4impact portal connects to these services using OpenID, and offers a user interface for searching, visualizing, and downloading GCM data and more. A challenging task was to describe the available model data and how it can be used. The portal tries to inform users about possible caveats when using climate model data. All impact use cases are described in the documentation section, using highlighted keywords pointing to detailed information in the glossary. During the project, the content management system Drupal was used to enable partners to contribute on the documentation section. In this presentation, the architecture and following items will be detailed: • Visualization: Visualize data from ESGF data nodes using ADAGUC Web Map Services. • Processing: Transform data, subset, export into other formats, and perform climate indices calculations using Web Processing Services implemented by PyWPS, based on NCAR NCPP OpenClimateGIS and IS-ENES2 icclim. • Security: Login using OpenID for access to the ESGF data nodes. The ESGF works in

Title and Presenter	Abstract
	conjunction with several external websites and systems. The climate4impact portal uses X509-based short-lived credentials, generated on behalf of the user with a MyProxy service. Single Sign-on is used to make these websites and systems work together. • Discovery: Faceted search based on e.g. variable name, model, and institute using the ESGF search services. A catalog browser allows for browsing through CMIP5 and any other climate model data catalogues (e.g. ESSENCE, EOBS, UNIDATA). • Download: Directly from ESGF nodes and other THREDDS catalogs. This architecture will also be used for the future Copernicus platform, developed in the EU FP7 CLIPC project. • Connection with the downscaling portal of the University of Cantabria. • Experiences on the question and answer site via Askbot. There are two current main objectives for climate4impact. The first one is to work on a web interface, which automatically generates a graphical user interface on WPS endpoints. The WPS calculates climate indices and subset data using OpenClimateGIS/icclim on data stored in ESGF data nodes. Data is then transmitted from ESGF nodes over secured OpenDAP and becomes available in a new, per user, secured OpenDAP server. The results can then be visualized again using ADAGUC WMS. Dedicated wizards for processing of climate indices will be developed in close collaboration with users. The second one is to expose climate4impact services, so as to offer standardized services which can be used by other portals. This has the advantage of adding interoperability between several portals, as well as enabling the design of specific portals aimed at different impact communities, either thematic or national.
Learning About Data Systems and Data Tools from Practitioner Applications ***Ricky Rood (University of Michigan, rbrood@umich.edu)***	As discussed in Rood and Edwards (2014; http://earthzine.org/2014/05/22/climate-informatics-human-experts-and-the-end-to-end-system/) there are substantial barriers to the usability of climate data by both scientific and non-scientific users. The conclusions in Rood and Edwards were drawn from several sources: experience in the Great Lakes Integrated Sciences and Assessments project, the National Climatic Predictions and Projections Platform, academic literature on usability, and formal evaluations by students in a University of Michigan class, Climate Informatics. A fundamental misconception remains in many data systems and data services in the climate community; namely, that providing accessibility of data is adequate to ensure the broad use of data. The perpetuation of this loading-dock model of data provision assures that data and knowledge are used by those most vested in the communities—e.g., climate scientists and those specifically funded in projects that require the use of the data. Moving data off the loading dock requires services that understand the use cases of the end users and contribute to a chain of tools and services, including training on what to do with the data and how to do it. In numerous interviews with end users, basic barriers such as glossaries that define scientific terms as well as arcane files names were cited as barriers that motivated users to abandon online services in a matter of minutes. Other barriers included data formats unknown in user communities and difficulty in developing interfaces with a community's tools. These barriers in concert with broader issues of tailoring data and knowledge to specific applications led to the need for data systems to support the roles of human intercessory in the chain connecting data providers with data users.
Model Output Evaluation and Data Dissemination for Seasonal and Shorter Time Scales: NMME and HIWPP ***Cecelia De Luca (NOAA/ESRL, cecelia.deluca@noaa.gov) and Eric Nienhouse (NSF/NCAR, ejn@ucar.edu)***	The evaluation of model outputs for seasonal and shorter-term predictions is being coordinated through projects that involve efforts from multiple modeling and forecast centers. These coordinated efforts can introduce high-volume data products and new challenges to distributed data systems such as ESGF. In this talk, we describe two such efforts, the National Multi-Model Ensemble and the High Impact Weather Prediction Project. We will present the projects' use of ESGF and the CoG user interface as a project documentation, information dissemination, data discovery, and data access infrastructure.
The ACME Modeling Project Infrastructure ***Dave Bader (DOE/LLNL, ACME Project Council Chair, bader2@llnl.gov)***	The ACME project is a newly launched project sponsored by the Earth System Modeling program within DOE's Office of Biological and Environmental Research. ACME is an unprecedented collaboration among eight national laboratories, the National Center for Atmospheric Research, four academic institutions, and one private-sector company to develop and apply the most complete, leading-edge climate and Earth system models to the most challenging and demanding climate-change research imperatives. It is the only major national modeling project designed to address U.S. DOE mission needs and efficiently use DOE Leadership Computing resources now and in the future.

Title and Presenter	Abstract
	ACME will achieve its goals through four intersecting project elements: 1. A series of prediction and simulation experiments addressing scientific questions and mission needs; 2. A well-documented and tested, continuously advancing, evolving, and improving system of model codes that comprise the ACME Earth system model; 3. The ability to use effectively leading (and "bleeding") edge computational facilities soon after their deployment at DOE national laboratories; and 4. An infrastructure to support code development, hypothesis testing, simulation execution, and analysis of results. This talk will describe how UV-CDAT and ESGF are essential pieces for the development of the scalable infrastructure to enable science simulation at the exascale. The priority science drivers and resulting three-year experiments were used to define the functionality of the initial simulation system. Initial infrastructure design was based on the requirements to facilitate hypothesis-testing workflows (configuration, simulation, diagnostics, and analysis). The infrastructure element will be continuously evolving. It will maintain a disciplined software engineering structure and develop turnkey workflows to enable efficient code development, testing, simulation design, experiment execution, analysis of output, and distribution of results within and outside the project.
A User's Perspective on Acquisition and Management of CMIP5 Data ***Jennifer Adams (COLA/NOAA, jma@cola.iges.org)***	The complexity, volume, and distributed nature of CMIP5 data collection has left many users struggling to acquire the CMIP5 data they need. This presentation outlines strategies that were developed to overcome the challenges CMIP5 data users face: authentication, searching for published data that match a list of desired experiments and variables, acquisition of Wget scripts, managing Wget script execution and the high Wget failure rate, retention of critical metadata not present in the data files, version control, local data management, and setting up the data for analysis and visualization using GrADS. All these strategies exist in an automated workflow that is completely independent of any browser interface.
The GeoMIP Perspective on Interactions with ESGF ***Ben Kravitz (Pacific Northwest National Laboratory, ben.kravitz@pnnl.gov)***	In this talk, I discuss some of the strengths and weaknesses of ESGF as I have seen through my work on the Geo-engineering Model Intercomparison Project (geoMIP). ESGF has provided an excellent common framework so that all of the climate model output necessary for conducting geoMIP analysis is available in one place in a standard format; such coordinated efforts are necessary for projects of the magnitude of CMIP6. However, the issues we have encountered in both hosting and retrieving climate model output are substantial. Establishing a data node has proven to be costly and time consuming, and transferring output to other nodes for hosting has met with moderate success. Downloading climate model output from ESGF is most reliably done one file at a time due to complications with certificate authentication. Although some nodes are well set up for Wget or Globus Online transfer, we have yet to discover a universal method for downloading large numbers of files from ESGF.

Session Four: Modeling and Data Center Requirements

Title and Presenter	Abstract
Modeling and Data Center Requirements **(Session's Keynote)** ***Sébastien Denvil (IPSL/IS-ENES2, sebastien.denvil@ipsl.jussieu.fr)***	Earth system model simulations are central to the study of complex mechanisms and feedbacks in the climate system and to provide estimates of future and past climate changes. Recent trends in climate modeling are to add more physical components in the modeled system, increasing the resolution of each individual component and the more systematic use of large suites of simulations to address many scientific questions. Climate simulations may therefore differ in their initial state, parameter values, representation of physical processes, spatial resolution, model complexity, and degree of realism or degree of idealization. In addition, there is a strong need for evaluating, improving, and monitoring the performance of climate models using a large ensemble of diagnostics and better integration of model outputs and observational data. At the same time, the Data and Supercomputing Center offers services to several communities and to specific communities (like the climate modeling community). This talk will try to gather common general requirements that must be fulfilled to ensure the Center's acceptance of a system such as ESGF and to ensure its willingness to support and contribute to ESGF.
Australia (ANU/NCI)	NCI provides a high-performance collaborative center for the Australian research community that spans national science agencies and research institutions. A particular initiative has been to provide

Title and Presenter	Abstract
Ben Evans (ANU/NCI, Ben.Evans@anu.edu.au)	a high-performance data and high-performance computing environment that is suited for both modeling and analyzing the whole earth system. To achieve this at a high standard for both research and government outcomes, there have been significant improvements to the breadth of services and their functionality, their integration with the underlying hardware (supercomputing, cloud, visualization, and data storage), the overall manageability and provenance management, and flexibility for ongoing expansion. The core of this includes the data management (both metadata and data), the tools (both standard and collaborative development), data services, and community environments (virtual laboratories). ESGF infrastructure is a significant component of our infrastructure, and a flagship for international collaborative infrastructure, but it is just one part in this dizzying array of components. I will highlight how the ESGF currently fits for us and consider the challenges of infrastructure that is highly connected, relevant to research needs, nationally and internationally trusted, aligned with our world peers, and stays on the critical path to help meet the future needs.
IPCC-DDC (DKRZ) ***Martina Stockhause (WDCC/DKRZ, stockhause@dkrz.de)***	DKRZ hosts the IPCC Data Distribution Center, which provides long-term access to Coupled Model Intercomparison Project data for interdisciplinary (re-)use. Beyond permanent and persistent data access, the Center must provide detailed documentations, a uniform data quality, and DataCite DOI data citations to enable data users to accept or even trust the data and to give credit to data creators.
France (IPSL) ***Sébastien Denvil (IPSL/IS-ENES2, sebastien.denvil@ipsl.jussieu.fr)***	ESGF has been operational in France since April 2011. There are six institutions running ESGF in France. Three of them are modeling groups—IPSL, CNRM, and CERFACS—and three of them are National Supercomputing and Data Centers—TGCC, IDRIS, and CINES. This talk will present how the French partners organized themselves to ensure smooth operations and effective contributions to the ESGF federation.
The Status of ESG-BNU Node in China ***Baogang Zhang, China (Beijing Normal University)***	ESGF is an operational system for serving climate data from multiple locations and sources. The ESG-BNU node, established in 2012, is one of 5 ESGF nodes in China. It is also the only IdP/index node in China. At present, the ESG-BNU node has published 17 experiments of 10,595 GB data set for CMIP5 and 4 experiments of 1,174 GB data set for GeoMIP—both generated by the Earth System Model of Beijing Normal University (BNU-ESM). It has provided over 31.02 TB of data with an average 700 KB/s download speed to scientists all over the world. As an IdP/index node, we try to make replicas of other model centers so that the Chinese data user can access the data sets much faster. We find that over 95% of data requests are from China, the U.S., and Japan. About 75% downloads come from historical, rcp45 and rcp85 experiments and 79% of downloads come from the monthly data sets. As for the upcoming CMIP6, data volume of one model center will reach to over 100 TB. Making replicas of all experiments may be too expensive and not necessary for many ESGF nodes. We hope such download statistical analysis can be helpful for ESGF nodes to decide which experiments should be replicated or be replicated with priority.
Experiences with the ESGF Data Portal at CCCma ***Slava Kharin (Canadian Centre for Climate Modelling and Analysis, Slava.Kharin@ec.gc.ca)***	We present some experiences at CCCma dealing with the ESGF data portal during the CMIP5 exercise and suggest a few areas for improvements.
UK (BADC) ***Phil Kershaw (BADC/IS-ENES2, philip.kershaw@stfc.ac.uk)***	The BADC is one of four data centers operated by CEDA, The Centre for Environmental Data Archival on behalf of the U.K. Natural Environment Research Council. CEDA receives it overarching requirements for engaging ESGF through NERC: to maximize the U.K.'s contributions to the CMIP cycle and exploitation of the data for the user communities it serves. Alongside this, there are number of supplementary requirements related to CEDA's stakeholders to which it is contracted to provide services. These include the U.K. Met Office, U.K. government, IPCC, European climate community and others. Over the past years, we have seen international collaboration as a key to meeting these objectives: engaging with shared software development effort was more likely to result in systems fit for purpose and builds a community upon which to pool resources to create common tools and services. With the growth of ESGF as an operational system, however there is a need to address a number of additional critical factors: how to best integrate ESGF services with the rest of CEDA's evolving infrastructure, support for multiple projects within the federation, and the ongoing operation and maintenance of the system and expected levels of service providers can meet. In this

Title and Presenter	Abstract
	presentation, we will explore these challenges.
NSF NCAR Data Center Requirements: Reducing Barriers to Community Data Products (USA) ***Eric Neinhouse (NSF/NCAR)***	As a national center, NSF-NCAR manages over 5 PB of climate and related data products. Integration of these products with distribution systems such as ESGF removes barriers to scientific data discovery and access. In this talk we will present key requirements and challenges for including these valuable data products in ESGF: • Publication of existing data products • Challenges of data discovery and use • The power of use metrics from distributed systems • Integration with tertiary storage systems • The need for software and content governance
GFDL Model Data Requirements and ESGF (USA) ***Serguei Nikonov* (NOAA/GFDL,** *Serguei.Nikonov@noaa.gov)*	The last two reports from the IPCC drove the architecture of the Geophysical Fluid Dynamics Laboratory (GFDL) data portal and brought it to its current state. During the CMIP5 project, GFDL was running two Data Portal infrastructures—ESGF and GFDL Curator. Running both gave us a good opportunity to compare strong and weak sides of both systems and elaborate requirements to the "ideal" system. For example, a useful feature would be interchangeable modularity of ESGF architecture for possibility to incorporate local model center existing subsystems element into ESGF. Some examples of such GFDL infrastructure elements will be discussed for consideration.

Session Five: Network Requirements

Title and Presenter	Abstract
Network Requirements—ICNWG **(Session's Keynote)** ***Eli Dart (DOE/ESnet, dart@es.net)***	ICNWG has been concentrating on improving data transfer performance for replication activities between five sites: ANU/NCI, BADC, DKRZ, KNMI, and PCMDI/LLNL. This talk will provide an update on the progress of the ICNWG efforts, and will discuss possible future scenarios for high-speed data replication and transfer between major data centers and computing facilities. If time permits, some highlights from the recent Climate CrossConnects meeting in Boulder, CO, will be discussed as well.

Day 2: Wednesday, 10 December 2014
ESGF and UV-CDAT Technical Presentations and Discussions

Session Six: ESGF Technical Development

Title and Presenter	Abstract
Technical Developments for the Community ***Dean N. Williams (DOE/LLNL, williams13@llnl.gov)***	The community of technical team members, consisting of computational and climate scientists, worked across institutional boundaries to develop and integrate software packages and sub-components for facilitating climate research. This effort enabled scientists in their daily activities and aided in the publication of hundreds of scientific articles. The developed partnership among the sponsors, institutions, universities, and private companies used the best human–computer interactions theory-based approach, coupled with pragmatic implementation of the system–user interface and modes of interaction. The scaled Agile development practice took full effect, highlighting individual roles, teams, and activities. This allowed team members to adjust schedules and priorities as necessary to quickly provide new solutions and meet the community's ever-changing needs. The goal of this presentation is to show integration of the many software components and how they relate to existing and future community projects.
ESGF Installation Working Team ***Nicolas Carenton (IS-ENES/IPSL, nicolas.carenton@ipsl.jussieu.fr) and Prashanth Dwarakanath (Linköping University, pchengi@nsc.liu.se)***	The ESGF installation working team was created in March 2014. Its main responsibilities are ESGF releases management, installation tools maintenance as well as node administrators support. One of its most important challenges is to provide an automated installation of a node that can complete in less than an hour. We will present here the work done since the team creation, which includes the recovery of the ESGF build process, the implementation of several distribution mirrors, the improvement of the release management process, and the new test and validation tools. We will also present the major releases of the year and the upcoming work on the installer that will lead to easier installation for administrators.

Title and Presenter	Abstract
CoG: The New ESGF User Interface ***Luca Cinquini (NOAA/ESRL, Luca.Cinquini@jpl.nasa.gov)***	The Earth System CoG is a web interface that organizes data distribution for a multitude of projects in a federated and distributed environment. CoG will soon be replacing the current ESGF web user interface (i.e. the "web front-end" module). Over the past year, the CoG team has worked through the tasks that needed to be accomplished to make the ESGF/CoG merging possible. Major upgrades in CoG functionality for the ESGF user community include a more powerful and flexible search interface, a model for exchanging information among peer nodes, streamlined group registration, and co-location of project data and documentation, just to name a few. This talk will review the status of the CoG development for ESGF adoption, and the last steps needed to execute the switch.
Publication as a Service: Globus Publish to ESGF ***Sasha Ames (DOE/LLNL) and Rachana Ananthakrishnan (U. of Chicago, ranantha@uchicago.edu)***	This talk will describe our progress made in the design and implementation of "publication-as-a-service" for ESGF. Publication-as-a-service takes the "system administration" overhead of data publication out of the hand of scientists, whose goal is to make their data available to the community. We will present the web interface to the service and details the backend processes that show the interaction of the web service with the ESGF publisher infrastructure. Additionally, we will discuss some recent changes to the publisher software that fit with the goals for supporting publication as a service.
Quality Control Working Team: esgf-qcwt ***Martina Stockhause (IS-ENES2/DKRZ, stockhause@dkrz.de) and Katharina Berger (IS-ENES2/DKRZ)***	Within the quality team, we consider all questions of how ESGF could be improved in order to increase the quality of ESGF data services. The main working task is the integration of external information into ESGF, such as information on provenance, quality, and data citation. This implies the storage of data unpublishes events for provenance and the support of data collections (granularity of data citations). Requirements out of the WIP for CMIP6 for this team will be integrated. Close collaborations with the ESGF teams on "Replication and Versioning" and "Publication" are required. We plan to give a demonstration to show the first results of the team.
ESGF IdEA—Developments with ESGF's system for Identity, Entitlement and Access Management ***Phil Kershaw (IS-ENES/BADC, philip.kershaw@stfc.ac.uk) and Rachana Ananthakrishnan (U. of Chicago, ranantha@uchicago.edu)***	We will present an update on progress with the activities identified in the roadmap for the development of ESGF access control system—ESGF-IdEA—set out at this meeting last year. This defined a plan for enhancements and improvements to the system. Since then an ESGF-IdEA working team has been established and has been meeting to co-ordinate work. We will present developments including support for simplified browser-based single sign-on process and authentication with Wget scripts without the need for X.509 certificates. We will also provide an assessment of the priorities for future work in the context of experience with the operational federation over the last year.
ESGF Transfer ***Eric Blau (U. of Chicago, blau@mcs.anl.gov), Rachana Ananthakrishnan (U. of Chicago, ranantha@uchicago.edu)***	During this session, we will provide an update on the transfer capabilities available in ESGF. Updates include using latest GridFTP version of server, simplification of install process, "Science DMZ" friendly install options and improvements to Wget for usability.
Automated Replication and Versioning ***Stephen Kindermann (IS-ENES2/DKRZ, kindermann@dkrz.de) and Tobias Weigel (IS-ENES2/DKRZ, weigel@dkrz.de)***	We will give an overview on past experiences made with replication and versioning and pressing issues. We will present some first concepts on how to tackle these issues, which involve the coherent assignment and use of persistent identifiers across ESGF services. At the technical level, the first step is to embed resolvable identifiers in the netCDF headers of files submitted to ESGF. In a second phase, identical replicas could be identified and individual replica IDs stored in their metadata. This, however, only works if the necessary policies are enforced and services or tools are provided to encapsulate the whole functionality for daily operations. Individual data versions may also be made more accountable if persistent identifiers are assigned to each published set and if metadata is carried along that enabled the user interface to redirect users to the most recent version.
The ESGF Desktop: A Web–Desktop Interface to the ESGF Monitoring Infrastructure	The ESGF Desktop represents the graphical user interface of the ESGF monitoring infrastructure. It exploits the MVC design pattern and it relies on a strong adoption and implementation of Web 2.0 concepts such as mash-up, Google maps, and permalinks. It provides several views at different (hierarchical) granularity levels of the entire federation. In particular, through the ESGF Desktop,

Title and Presenter	Abstract
P. Nassisi (IS-ENES2/CMCC, paola.nassisi@cmcc.it), S. Fiore Aloisio (IS-ENES2/CMCC, sandro.fiore@unisalento.it) and G. Aloisio (IS-ENES2/CMCC, giovanni.aloisio@unisalento.it)	the user has the ability to visualize a set of statistics for both local (node-level) monitoring and global (institution-level and/or federation-level) monitoring. Real-time gadgets are also available. From an implementation point of view, the ESGF Desktop is a pure JavaScript application with a set of RESTful APIs to make all of the metrics available to external applications. In terms of data usage statistics, the ESGF Desktop displays through specific gadgets: • The number of downloads; • The number of downloaded data sets; • The number of users that have downloaded some data; • The amount of data in term of downloaded gigabytes or terabytes; • The most downloaded (e.g. top ten) data sets, variables, and models; and • Data download client distribution. Additional gadgets are more related to multimedia content access (wiki pages, etc.) and desktop customizations.
Monitoring the Earth System Grid Federation Through the ESGF Dashboard *P. Nassisi (IS-ENES2/CMCC, paola.nassisi@cmcc.it), S. Fiore Aloisio (IS-ENES2/CMCC, sandro.fiore@unisalento.it) and G. Aloisio (IS-ENES2/CMCC, giovanni.aloisio@unisalento.it)*	The ESGF Dashboard is a software component of the ESGF stack, responsible for collecting key information about the status of the federation in terms of: • Network topology (peer-groups composition) • Node type (host/services mapping) • Registered users (including their Identity Providers) • System metrics (e.g., round-trip time, service availability, CPU, memory, disk, processes, etc.) • Real-time metrics (e.g. RAM, CPU, etc.) • Download statistics (both at the Node and federation level) The last class of information is related to the data usage statistics, which are very important since they provide a strong insight into CMIP5 experiments. During the presentation, the ESGF Dashboard architecture components (the information provider, the dashboard catalog, and the command line interface) will be presented jointly with a new scalable and configurable back-end storage model for long-term metrics.
Improved Usability and Support: esgf-swt *Matthew Harris (DOE/LLNL, harris112@llnl.gov)*	Documentation: developers do not want to write it and users do not read it. This age-old argument has come to the forefront of ESGF discussions. With an aging and converted wiki, an up, down, and then up again Askbot, and unarchived non-searchable email list, user support needs to be reconsidered and reorganized. The user must be able to define and support multiple software components to gain usage understanding. **Key points:** • What does the term user mean? • Who are our users? • Askbot beta runs. Now production • Creating a achieved email list and making it searchable • Wiki/Wikis standards, cleanup, uses
Revisiting the ESGF Node Manager for Federation Scalability *Prashanth Dwarakanath (Linköping University, pchengi@nsc.liu.se) and Sasha Ames (DOE/LLNL, ames4@llnl.gov)*	The ESGF node manager is the component within the ESGF software stack that: • Gathers metrics • Shares node information across federated nodes • Facilitates user group management The present implementation of the node manager has scalability limitations arising from the peer-to-peer protocol. We present a design for next-generation node manager software and protocol that will address the exiting shortcomings of the current implementation and will offer additional features for evolving software components.
ESGF Metadata Search Evolution: esgf-mswt *Luca Cinquini (NASA/JPL, Luca.Cinquini@jpl.nasa.gov)*	Arguably, one of the most important features of ESGF is the capability to search, in real time, a system of metadata archives that are distributed around the world, and administered independently. Nonetheless, as the ESGF collaboration grows in scale and scope, its search capabilities must evolve to address new challenges, including: • Rigorous validation of metadata according to controlled vocabularies and schemas; • Partition of the search space according to project-specific criteria; • Dynamic configuration of peer circles; • Big Data scalability; and • New search operations such as temporal and geospatial constraints.

Title and Presenter	Abstract
	This talk will discuss some ideas on how to upgrade the search infrastructure for the next generation projects, staring with CMIP6.
Making the Case for the ESGF and Apache: Long-Term Software Stewardship ***Chris Mattmann (NASA/JPL, chris.a.mattmann@jpl.nasa.gov)***	In this talk I will give an overview of the Apache Software Foundation, its setup, meritocracy and governance structure. I will discuss Apache as a home for many long term and widely used software projects including HTTPD, Tomcat, and more recently de facto big data platforms like Spark, Mesos, Hadoop, Tika, Lucene, and others. I will also identify Apache as a modern home for science-driven OSS, including the Apache OODT, and Open Climate Workbench projects. I will finally propose Apache as a potential home for ESGF software stewardship and code base.

Session Seven: UV-CDAT Technical Development

Title and Presenter	Abstract
ESGF Compute Working Team (ESGF-CWT): Distributed Analytics Application Programming Interface (API) **(Session's Keynote)** ***Dan Duffy (NASA Center for Climate Simulation, daniel.q.duffy@nasa.gov)***	The model output from the IPCC-AR6 is estimated to create four to five times more data than is currently in the AR5 distributed archive. It is clear that data analysis capabilities currently available across ESGF will be inadequate to allow for the necessary science to be done with AR6 data—the data will just be too big. A major paradigm shift from downloading data to local systems to perform data analytics must evolve to moving the analysis routines to the data and performing these computations on distributed platforms. In preparation for this need, the ESGF has started a CWT to create solutions that allow users to perform distributed, high-performance data analytics on the AR6 data. The team will be designing and developing a general API to enable highly parallel, server-side processing throughout the ESGF data grid. This API will be integrated with multiple analysis and visualization tools, such as UV-CDAT, netCDF Operator, and others. This presentation will provide an update on the ESGF CWT's overall approach toward enabling the necessary storage proximal computational capabilities to study climate change using the AR6 extreme-scale distributed data archive. An update on the API will be provided, along with a survey of the overall computational approaches being reviewed and studied by the members of the ESGF CWT.
ESGF Compute Node API ***Charles Doutriaux (DOE/LLNL, doutriaux1@llnl.gov)***	With each new round of CMIP/IPCC, the volume of data served has grown about two orders of magnitude. CMIP6 will be no exception and is expected to generate four to five times the amount of data of CMIP5. With such volumes, it is not only impractical for any organization to hold all the data locally, but the end user will most likely not be able either to download locally the subset of data needed for his/her research both in terms of disk space and download bandwidth. One solution to this is to stop bringing the data to the scientists, but rather to bring their codes to the data. With this in mind, the ESGF CWT was established. The primary charge to team is to allow ESGF users to execute analysis tools on high-end compute clusters, high-performance computers, cloud servers, and other forms of compute servers. In this talk we describe the state of the group, the decision taken so far, and the directions the groups is exploring.
Diagnostics: acme-dwt ***Jeff Painter (DOE/LLNL, painter1@llnl.gov), Jim McEnerney (DOE/LLNL, mcenerney1@llnl.gov), and Brian Smith (DOE/ORNL, smithbe@ornl.gov)***	UV-CDAT diagnostics are one of the more important tools for climate scientists and code developers to compare model output with observations or another model. Data computed from model output or observations is plotted or sometimes tabulated. We have improved this diagnostic tool to offer twelve sets of plot for atmosphere and five for land; often with hundreds of possible plots per plot set. We can make plots for the UV-CDAT GUI, where they may be changed interactively; or we can write them as static image files for viewing with traditional tools such as web browsers. The diagnostic tools can be run from a GUI or with one of three command-line scripts.
Diagnostics and Web Informatics Support Infrastructure ***Brian Smith (DOE/ORNL, smithbe@ornl.gov)***	We have developed a number of tools to make viewing output produced by UV-CDAT-based scripts easier for the end users. Part of that work has been an extensible Django-based back end and several front-end scripts to produce results that are similar to the NCAR diagnostics yet flexible enough to allow additional diagnostics and visual analysis tools to use them. In this talk, the "meta" diagnostics master script and some of the Django back-end components will be discussed.
ACME Exploratory Analysis and Diagnostics Viewer ***John Harney (ORNL/DOE, harneyjf@ornl.gov)***	The traditional CESM diagnostics package provides scientists and modelers a way to quickly analyze the effectiveness of a climate model by computing climatological means of the large-scale simulations, and producing hundreds of plots and tables of the mean climate in a variety of formats. While the modeling community has successfully utilized this toolkit for a number of

Title and Presenter	Abstract
	years, it is incompatible with modern workflows and methodologies. In this talk, we introduce the novel "Classic" diagnostics viewer, an improved interface for diagnostics. The Classic viewer is a key component of the Exploratory Analysis toolkit in the ACME post-processing workflow, which utilizes UV-CDAT for efficient computation of key metrics and ESGF for data archiving and cataloging. We will explore the various features offered by the Classic viewer, as well as key features currently under development.
On Demand Data Reordering for Remote Data Processing and Exploration ***Timo Bremer (LLNL/DOE, bremer5@llnl.gov)***	The ViSUS client integrated UV-CDAT allows interactive processing and visualization of large-scale ensembles of both local and remote data sets. However, it requires a data reordering to enable a progressive, out-of-core data access. In this talk we will introduce a new server infrastructure currently under development that provides a transparent re-ordering of data stored in an ESG node. This will enable any data severed by an ESG to be accessed through ViSUS and will allow an extensive processing and exploration of remote data.
ESGF Analysis and Visualization: Challenges and Opportunities ***Roland Schweitzer (NOAA/PMEL, roland.schweitzer@noaa.gov) and Kevin O'Brien (NOAA/PMEL, kevin.m.o'brien@noaa.gov)***	The Live Access Server from NOAA's Pacific Marine Environmental Laboratory is an independent web application that has a decades-long record of providing analysis and visualization of climate data. As such, we were fortunate to join the ESGF project during the run-up to the release of the CMIP5 data collection. During the initial planning, we developed an ambitious set of goals for a deeply integrated UI with a custom data set selection, region selection and other UI controls as a direct part of the Gateway Node software stack. We further envisioned multiple backend engines, namely NCL, CDAT, and Ferret, producing products as desired by the installer of the system. However, having a custom and integrated UI proved to be too ambitious. In fact, to date, it seems the majority of ESGF use has been focused on discovering data to download for local use. Offering services like those envisioned at the outset of the CMIP5 cycle requires a clear understanding of where and how the interaction with data set will take place. In some cases, that interaction takes place in environments that lack the flexibility for easy integration of authentication and authorization mechanisms. In this presentation, we will discuss opportunities that we envision to provide better visualization and subsetting capabilities in the ESGF context, while also recognizing the challenges that have prevented this from becoming a reality.
Web-based Visualization: Overview of Client and Server Side Techniques and their Use in GeoJS and CDATWeb ***Aashish Chaudhary (Kitware, aashish.chaudhary@kitware.com)***	With the power of Web 2.0 at the fingertips of the developers, it is now possible to create high-performance visualization for climate and geospatial domain experts in a web browser. Currently, this can be achieved in many ways. In this talk, we will present various technologies that can provide practical solutions and how we are incorporating them in the development of CDATWeb, a web-based visualization framework for UV-CDAT.
Ultrascale Climate Data Visualization and Analysis Using UV-CDAT and DV3D ***Thomas Maxwell (NASA/GSFC, thomas.maxwell@nasa.gov) and Jerry Potter (NASA/GSFC, gerald.potter@nasa.gov)***	In collaboration with the UV-CDAT development consortium, NASA National Center for Climate Simulation is developing climate data analysis and visualization tools for UV-CDAT. These tools feature workflow interfaces, interactive 3D data exploration, hyperwall and stereo visualization, automated provenance generation, parallel task execution, and streaming data parallel pipelines. NASA's DV3D is a UV-CDAT package that enables exploratory analysis of diverse and rich data sets from various sources including ESGF. DV3D's user-friendly interface places many complex visualization operations, previously sequestered within the exclusive domain of visualization specialists, at the fingertips of climate scientists. DV3D's tight integration into UV-CDAT seamlessly couples a wide range of high-performance climate data analysis operations with a rich palette of interactive visualization methods.
Workflow and Testing for Modern Software ***Aashish Chaudhary (Kitware, aashish.chaudhary@kitware.com)***	Software testing is an essential component in making sure that software meets a certain standard; open source or otherwise. To ensure the software quality for the modern software, it is important to setup a workflow that strengthens software testing for tools using a distributed code repository and code contribution from developers from various geo-locations. In this presentation, we will talk about modern software processes and testing for desktop- and web-based projects.
GIS Capabilities in UV-CDAT ***Ben Koziol (NOAA/ESRL, ben.koziol@noaa.gov)***	Following the ESGF/UV-CDAT F2F meeting in December 2013, a plan was initiated to bring low-level geographic information system capabilities into UV-CDAT (i.e. subsetting via ESRI Shapefile) through functionality provided by OpenClimateGIS (OCGIS). OCGIS is an open source Python package designed for geospatial manipulation, subsetting, computation, and translation of climate data sets stored in local netCDF files or files served through OpenDAP data servers. In addition to an overview of OCGIS, this presentation will provide a description of the proposed

Title and Presenter	Abstract
	integration plan with UV-CDAT and discuss the current state of development. Furthermore, an overview of OCGIS development with ESMPy (the Python interface to the Earth System Modeling Framework [ESMF]) will be described. The ESMP-OCGIS connection is an additional pathway for bringing GIS-like operations into UV-CDAT. ESMP, the previous version of the ESMF Python interface, is currently used by UV-CDAT for regridding operations.

Day 3: Thursday, 11 December 2013
ESGF and UV-CDAT Technical Presentations and Discussions
Technical Interoperability Discussions

Session Eight: Other Related Community Contributions

Title and Presenter	Abstract
ES-DOC ***Mark Greenslade (IS-ENES/IPSL, momipsl@ipsl.jussieu.fr)***	During 2014, the ES-DOC project (http://es-doc.org) extended its ecosystem by deploying significant upgrades of its documentation creation, search, viewing, and comparison tools. The ES-DOC questionnaire and the pyesdoc scripting library will serve as the pathway for generating CMIP6 documentation. Underpinning these tools is a publically available web service API in support of documentation search, publication, and comparison. ES-DOC continues to support and leverage the emergent Metafor CIM documentation standard and is proactively assisting other projects to do likewise. Furthermore, the project has improved its own software development process via streamlined deployments and deepened automated testing.
Towards a Controlled Vocabulary Service ***Mark Greenslade (IS-ENES/IPSL, momipsl@ipsl.jussieu.fr)***	DKRZ hosts the IPCC Data Distribution Center, which provides long-term access to CMIP data for the interdisciplinary (re-)use. Beyond permanent and persistent data access, the center needs to provide detailed documentations, a uniform data quality, and DataCite DOI data citations in order to enable data users to accept or even trust the data and to give credit to data creators.
System for Offline Data Access (SODA) ***Prashanth Dwarakanath (Linköping University, pchengi@nsc.liu.se)***	The System for Offline Data Access is one of the activities being carried out under the Climate Information Platform for Copernicus FP7 EU project, under the task "Dynamic Tape-archive extraction and post-processing." The deliverable of this exercise is a generic system that allows ESGF users to search and retrieve data stored on different kinds of offline tape systems. This work targets the publication of data from the EURO4M project present on the MARS system at SMHI-LIU as a demonstrator.
Climate and Forecast (CF) Conventions ***Karl Taylor (DOE/LLNL, taylor13@llnl.gov), Jeff Painter (DOE/LLNL, painter1@llnl.gov), Matthew Harris (DOE/LLNL, harris112@llnl.gov)***	The CF Conventions define a standard structure for netCDF files. They are widely followed in the climate community because they provide the machine-readable semantics that software needs to process climate data. Most of the data on ESGF follows the conventions, and the requirements for CMIP Phases 3–6 build on the CF Conventions. In the last year we have updated the CF Conventions document with some of the settled issues in our Trac issue-tracker system. We will make more such changes soon and then release it as CF Conventions version 1.7. Then CF Conventions 1.8 will contain a new chapter, describing the Gridspec specification, a structured way to describe unstructured grids. Most of our effort in the last year has been devoted to dealing with a hardware failure in the web server. We built a new GitHub-based web server for the CF Conventions site, and moved the Trac system first to one new server and then another. As these new hosts grow more stable, we will be able to shift back to work on the document itself.
Preparing CMOR for CMIP6 ***Charles Doutriaux (DOE/LLNL, doutriaux1@llnl.gov) and Karl Taylor (DOE/LLNL, taylor13@llnl.gov)***	One of the key components to the success of the MIPs, and in particular CMIP, has been the availability of multiple models output generated using common standard, which makes them easy to analyze and compare. CMIP5 came with a full set of documents describing in details which variables should be stored and how they should be made available back to the community, be it format, naming conventions, description conventions, etc. In addition to these many documents, CMIP5 provided a multi-language tool to help scientists producing conforming data. This tool is the Climate Model Output Rewriter (CMOR). This talk describes the state of the CMOR software

Title and Presenter	Abstract
	and what steps are envisioned to get it ready for CMIP6. In particular we will look at what needs to be added to conform to the new CMIP6 requirement and also to adapt to data that are not necessarily issued from the modeling community (observation, reanalysis, etc.).
Ophidia: A Big Data Analytics Framework for eScience ***G. Aloisio (IS-ENES2/Univ. of Italy, giovanni.aloisio@unisalento.it), S. Fiore (IS-ENES2/CMCC, sandro.fiore@unisalento.it)***	The Ophidia project is a research effort on big data analytics–facing scientific data analysis challenges in the climate change domain. It provides parallel (server-side) data analysis, an internal storage model, and a hierarchical data organization to manage large amount of multidimensional scientific data. The Ophidia analytics platform provides several data operators to manipulate multidimensional data sets. Some relevant examples include: 1) data subsetting (slicing and dicing), 2) data aggregation, and 3) data analysis. Additionally, the Ophidia framework provides about 100 primitives to perform time series analysis, subsetting, and data aggregation on large arrays of scientific data. Multiple primitives can be also nested to implement a single more complex task (e.g., aggregating by sum a subset of the entire array). The entire Ophidia software stack has been deployed at CMCC on 24 nodes (16-cores/node) of the Athena HPC cluster. A comprehensive benchmark and test cases are being defined with climate scientists to extensively test all of the features provided by the system. Preliminary experimental results are already available and have been published on scientific research papers.
JASMIN: Cloud Computing System ***Phil Kershaw (IS-ENES/BADC, philip.kershaw@stfc.ac.uk)***	Cloud computing is a disruptive technology that presents some significant opportunities and challenges for successful exploitation by the scientific community. It promises the ability to provide near-limitless computing resources on demand, pooling of resources between different groups and activities, and broad network access to support widely distributed communities of users. However, the specialist requirements for scientific workloads can be at odds with the commodity-based infrastructure typically provided by public clouds. One solution is to deploy a private cloud customized to meet these needs, but in doing so there are technical and operational challenges to be tackled. In the case of climate science and earth observation applications, the increasing volumes of data mean that access to large capacity storage with performant I/O to compute is a critical factor for effective processing and analysis. In this presentation we will explore experiences tailoring cloud computing technology and service models for a private cloud for JASMIN. JASMIN is a large storage and analysis facility for the UK climate science community and its international partners. It was first established two years ago and funded through the UK Natural Environment Research Council. Over the past year, a second phase of development has been underway expanding the system to over 3,000 processing cores and around 12 PB of disk-based storage. In addition the scope of the service has been broadened to the big data analysis needs of the entire UK environmental science community. JASMIN in its first phase rapidly demonstrated the benefits of its global file system and high performance I/O between storage and compute. One of the challenges in this new phase is to broaden access to the resource to better meet the needs of the so-called long tail of science. This area is being addressed directly with some of the first tenancies on the new cloud service. Among these, a service hosting Linux desktop environments has been customized with applications and libraries for an individual community's needs. Similarly, the popular IPython Notebook application was hosted. In each case users obtain access to interfaces and tools they are familiar with but underpinned by the extensive compute and storage resources of a large centralized facility. This presentation will explore the technical and organizational aspects involved in developing and deploying infrastructure for an operational cloud service.
ESGF in an OpenStack Cloud ***Ben Evans (ANU/NCI, Ben.Evans@anu.edu.au)***	Cloud systems provide a flexible platform for supporting a range of complex services that can easily scale in resources, be managed and upgraded, and is able to be linked with other inter-operating functionality and as the capabilities are developed and released. ESGF is a typical example of a software environment that is well suited to this. In this talk, I will discuss some of the key components of our installation, and some of the infrastructure issues that have been addressed, and issues that need to be addressed in future deployment of the ESGF system.
Requirements for a Biology node on ESGF ***Patrik D'haeseleer (DOE/LLNL, dhaeseleer2@llnl.gov) and Sasha Ames (DOE/LLNL,***	The federated database platform provided by ESGF can be of use in other disciplines as well, especially in biology, where there are hundreds of existing databases for specific purposes, and a chronic need for more integration and interoperability between them. One key opportunity lies in incorporating epidemiological data on ESGF. Disease outbreaks are inherently spatiotemporal, so this type of data should mesh well with the existing climate modeling infrastructure. Some diseases are known to exhibit seasonal, weather, and climate variations, so it should be possible to predict the likelihood of disease outbreak based on weather patterns, how the endemic range of diseases such as dengue is expected to shift with climate change, and perhaps

Title and Presenter	Abstract
ames4@llnl.gov)	even which mutations in pathogens are associated with climatic adaptations. We will focus on two key types of data: epidemiological data describing disease outbreaks over time in different locations, and sequence data, which is one of the most fundamental data types for modern-day biology.

11 Glossary

Acronym	Meaning and Website
ACME	Accelerated Climate Modeling for Energy: DOE's effort to build an Earth system modeling capability tailored to meet the climate change research strategic objectives
AGU	American Geophysical Union (http://sites.agu.org/)
AIMS	Analytics and Informatics Management Systems project responsible for all LLNL's climate software, including ESGF, UV-CDAT, and the DOE ACME Test Bed (http://aims.llnl.gov/)
Akuna	Advanced Simulation Capability for Environmental Management (ASCEM) developed AKUNA framework (http://akuna.labworks.org/)
AMS	American Meteorological Society (https://www.ametsoc.org/)
ana4MIPs	A pilot activity to make reanalysis products more accessible for climate model intercomparisons (http://www.wcrp-climate.org/documents/ezine/WCRPnews_17082012.pdf)
ANU/NCI	Australian National University/National Computational Infrastructure (http://www.nci.org.au)
API	Application Programming Interface (http://en.wikipedia.org/wiki/Application_programming_interface)
AR5	Fifth IPCC Assessment Report, published in 2013 (http://www.ipcc.ch/report/ar5/#.UwVGOCTm6Gg)
ARM	Atmospheric Radiation Measurement is a U.S. Department of Energy scientific user facility, providing data from strategically located in situ and remote sensing observatories around the world (http://www.arm.gov/)
BADC	British Atmospheric Data Centre (http://badc.nerc.ac.uk/)
BER	Office of Biological and Environmental Research under the DOE Office of Science (http://science.energy.gov/ber/)
CA	Certificate Authority is an entity that issues digital certificates (http://en.wikipedia.org/wiki/Certificate_authority)

Acronym	Meaning and Website
CDMS	11.1.1.1.1 Climate Data Management System
CESM	Community Earth System Model
CF	Climate and Forecast metadata convention, for processing and sharing netCDF data files (http://cf-pcmdi.llnl.gov/)
CFSR	NOAA's Climate Forecast System Reanalysis (http://cfs.ncep.noaa.gov/cfsr/)
Client-Server	Relationship between two computer programs, where the client program makes a service request, which the server program fulfills (http://en.wikipedia.org/wiki/Client-server)
CMCC	Euro-Mediterranean Center on Climate Change (http://www.cmcc.it/)
CMIP5	Coupled Model Intercomparison Project Phase 5, sponsored by WCRP/WGCM, and related multi-model database for the IPCC AR5 (http://cmip-pcmdi.llnl.gov)
CMIP6	Coupled Model Intercomparison Project Phase 6, sponsored by WCRP/WGCM, and related multi-model database planned for the IPCC AR6
CoG	A web environment that enables users to create project workspaces, connect projects into networks, share and consolidate information within those networks, and seamlessly link to tools for data archival, reformatting and search, data visualization, and metadata collection and display (https://earthsystemcog.org/)
CORDEX	Coordinated Regional Climate Downscaling Experiment, providing global coordination of Regional Climate Downscaling for improved regional climate change adaptation and impact assessment (http://wcrp-cordex.ipsl.jussieu.fr/)
CLIPC	11.1.1.1.2 Climate Information Platform for Copernicus
CMOR	11.1.1.1.3 Climate Model Output Rewriter
CPU	11.1.1.1.4 Central processing unit
CREATE-IP	11.1.1.1.5 Collaborative REAnalysis Technical Environment-Intercomparison Project
CVs	Controlled Vocabularies.
CWT	11.1.1.1.6 Compute Working Team
Data Node	Internet location providing data access or processing (http://en.wikipedia.org/wiki/Node-to-node_data_transfer)
DECK	11.1.1.1.7 Diagnosis, Evaluation, and Characterization of Klime Experiments
DKRZ	German Climate Computing Center (http://www.dkrz.de/)
DOE	Department of Energy, the U.S. government entity chiefly responsible for implementing energy policy (http://www.doe.gov/)

Acronym	Meaning and Website
DOI	11.1.1.1.8 Digital object identifier
DRS	11.1.1.1.9 Data Reference Syntax
DV3D	3D Climate data visualization using python and VTK (http://portal.nccs.nasa.gov/DV3D/)
EDEN	Exploratory Data analysis ENvironment is a visual analytics tool for exploring multivariate data sets. (http://cda.ornl.gov/projects/eden/)
ENES	11.1.1.1.10 European Network for Earth System Modeling
EGU	11.1.1.1.11 European Geosciences Union
EOS	NASA's Earth Observing System (http://eospso.gsfc.nasa.gov/)
ES-DOC	Earth System-Documentation, an international effort to develop metadata services for a set of climate and related projects (http://earthsystemcog.org/projects/es-doc-models/)
ESGF	Earth System Grid Federation, led by LLNL, a worldwide federation of climate and computer scientists deploying a distributed multi-petabyte archive for climate science (http://esgf.org)
ESMF	Earth System Modeling Framework is software for building and coupling weather, climate, and related models (http://www.earthsystemmodeling.org/)
ESnet	DOE Energy Science Network (https://www.es.net/)
ESRL	NOAA Earth System Research Laboratory (http://www.esrl.noaa.gov/)
Exascale	Computer processing capabilities of order 10^{18} operations per second (http://en.wikipedia.org/wiki/Bit)
F2F	Face-to-Face: being in the presence of another (http://www.thefreedictionary.com/face-to-face)
FUNET	Finnish university and research network (http://www.csc.fi/hallinto/funet)
Gbps	Gigabit per second, 10^9 bits of information (http://en.wikipedia.org/wiki/Data_rate_units)
GB	Gigabyte, 10^9 bytes of information (http://en.wikipedia.org/wiki/Gigabyte)
GCM	11.1.1.1.12 Global climate model
GeoMIP	Geo-engineering Model Intercomparison Project (http://climate.envsci.rutgers.edu/GeoMIP/)
GFDL	11.1.1.1.13 Geophysical Fluid Dynamics Laboratory

Acronym	Meaning and Website
GIS	11.1.1.1.14 Geographic information system
GO	Globus Online is an open-source software toolkit used for building grids (https://www.globus.org/)
GridFTP	Is an extension of the standard File Transfer Protocol for high-speed, reliable, and secure data transfer (http://toolkit.globus.org/toolkit/docs/latest-stable/gridftp/)
GUI	11.1.1.1.15 Graphical user interface
HPC	High-Performance Computing (http://en.wikipedia.org/wiki/Supercomputer)
ICNWG	International Climate Network Working Group, formed under the Earth System Grid Federation (ESGF), is to help set up and optimize network infrastructure for their climate data sites located around the world (http://icnwg.llnl.gov/)
IdEA	11.1.1.1.16 Identity, Entitlement, and Access Control
INTERNET2	U.S. research and education network (http://www.internet2.edu/)
I/O	11.1.1.1.17 Input/output
IPCC	Intergovernmental Panel on Climate Change, a scientific body of the United Nations, periodically issues assessment reports on climate change (http://www.ipcc.ch/)
IPSL	Institut Pierre-Simon Laplace (http://www.ipsl.fr/)
IS-ENES2	Infrastructure for the European Network of Earth System Modelling (https://verc.enes.org/ISENES2/)
JANET	U.K. research and education network (https://www.ja.net/)
JPL	NASA Jet Propulsion Laboratory (http://www.jpl.nasa.gov/)
KNMI	Royal Netherlands Meteorological Institute (http://www.knmi.nl/index_en.html)
LANL	Los Alamos National Laboratory, sponsored by the DOE (http://www.lanl.gov/)
LAS	11.1.1.1.18 Live Access Server
LBNL	Lawrence Berkeley National Laboratory, sponsored by the DOE (http://www.lbl.gov/)
LLNL	Lawrence Livermore National Laboratory, sponsored by the DOE (https://www.llnl.gov/)
LUCID	Local Urban Climate Model and its Application to the Intelligent Design of Cites in the U.K. (http://www.homepages.ucl.ac.uk/~ucftiha/index.html)
MERRA	NASA's Modern Era Retrospective-analysis for Research and Applications (http://gmao.gsfc.nasa.gov/merra/)

Acronym	Meaning and Website
Metadata	Data properties, such as their origins, spatio-temporal extent, and format (http://en.wikipedia.org/wiki/Metadata)
NARCCAP	North American Regional Climate Change Assessment Program (http://www.narccap.ucar.edu/)
NASA	National Aeronautics and Space Administration (http://www.nasa.gov/)
NCAR	National Center for Atmospheric Research, sponsored by the National Science Foundation (http://www.ncar.ucar.edu/)
NCI	National Computational Infrastructure at the Australian National University (http://www.nci.org.au/)
NCCS	NASA Center for Climate Simulation (http://www.nccs.nasa.gov/)
netCDF	A machine-independent, self-describing, binary data format (http://www.unidata.ucar.edu/software/netcdf/)
NERC	Natural Environment Research Council
NOAA	National Oceanic Atmospheric Administration, an agency of the U.S. Commerce Department (http://www.noaa.gov/)
NSF	National Science Foundation, an agency of the U.S. (http://www.nsf.gov/)
NYU	New York University (http://www.nyu.edu/)
obs4MIPs	Observation for Model Intercomparison Project, a pilot activity to make observational products more accessible for climate model intercomparisons (http://obs4mips.llnl.gov:8080/wiki/)
OCGIS	OpenClimateGIS is a Python package designed for geospatial manipulation, subsetting, computation, and translation of climate data sets stored in local netCDF files or files served through THREDDS data servers (https://earthsystemcog.org/projects/openclimategis/)
OPeNDAP	Open-source Project for Network Data Access Protocol is a data transport architecture and protocol widely used by Earth scientists (http://www.opendap.org/)
OpenID	Allows users to use an existing account to sign in to multiple websites, without needing to create new passwords (http://openid.net/)
ORNL	Oak Ridge National Laboratory, sponsored by the DOE (http://www.ornl.gov/)
ParaView	An open-source, multi-platform data analysis and visualization application. (http://www.paraview.org/)
PB	Petabyte, 10^{15} bytes of information (http://en.wikipedia.org/wiki/Petabyte)
PCMDI	Program for Climate Model Diagnosis and Intercomparison, located at LLNL (http://www-pcmdi.llnl.gov/)

Acronym	Meaning and Website
PDAS	Parallel Data Analysis Service
PID	11.1.1.1.19 Persistent identifier
PKI	Public Key Infrastructure is a set of hardware, software, people, policies, and procedures needed to create, manage, distribute, use, store, and revoke digital certificates (http://en.wikipedia.org/wiki/Public-key_infrastructure)
RCD	11.1.1.1.20 Regional climate downscaling
RVWT	11.1.1.1.21 Replication and Versioning Working Team
SODA	11.1.1.1.22 System for Offline Data Access
SURFnet	Netherlands network (http://www.surf.nl/)
SWT	Support Working Team
TAMIP	Transpose Atmospheric Model Intercomparison project (http://www.metoffice.gov.uk/hadobs/tamip/)
TB	Terabyte, 10^{12} (a trillion) storage bytes (http://en.wikipedia.org/wiki/Terabyte)
THREDDS	Thematic Real-time Environmental Distributed Data Services (https://www.unidata.ucar.edu/software/thredds/current/tds/)
UI	11.1.1.1.23 User interface
UV-CDAT	Ultrascale Visualization Climate Data Analysis Tools, provides access to large-scale data analysis and visualization tools for the climate modeling and observational communities (http://uv-cdat.llnl.gov)
VCS	Visualization Control System (http://uv-cdat.llnl.gov)
VisTrails	An open-source system that supports data exploration and visualization (http://www.vistrails.org/)
VO	Virtual Organization is one whose members are geographically apart, usually working by computer e-mail and groupware while appearing to others to be a single, unified organization with a real physical location (http://en.wikipedia.org/wiki/Virtual_organization)
VTK	Visualization ToolKit, an open-source, freely available software system for 3D computer graphics, image processing, and visualization (http://www.vtk.org/)
WCRP	World Climate Research Programme, which aims to facilitate analysis and prediction of Earth system variability and change for use in an increasing range of practical applications of direct relevance, benefit, and value to society (http://www.wcrp-climate.org/)
Web portal	A point of access to information on the World Wide Web

Acronym	Meaning and Website
	(http://en.wikipedia.org/wiki/Web_portal)
WGCM	Working Group on Coupled Modeling
Wget	The non-interactive network downloader is describe as a computer program that retrieves content from web servers and is part of the GNU Project (https://www.gnu.org/software/wget/)
WIP	11.1.1.1.24 Working Group on Coupled Modeling (WGCM) Infrastructure Panel
Workflow	A sequence of operations, performed by person(s), organization(s), or mechanism(s) (http://en.wikipedia.org/wiki/Workflow)

12 Participants and Contributors to the 2014 Report and Conference

Figure 23. *Group photo of the 2014 international conference attendees.*

12.1 Attendees and Contributors

Last Name	First Name	Affiliation
01. Adams	Jennifer	IGES/Cola
02. Aloisio	Giovanni	CMCC
03. Ames	Alexander	DOE/LLNL
04. Ammann	Casper	NSF/NCAR
05. Bader	David	DOE/LLNL
06. Baker	Jamie	Amazon Web Services
07. *Barrie	Daniel	NOAA HQ
08. Berger	Katharina	DKRZ
09. Blau	Eric	Argonne National Laboratory
10. Bremer	Peer-Timo	DOE/LLNL
11. Canada	Curtis	DOE/LANL
12. Carenton-Madiec	Nicolas	IPSL
13. Carriere	Laura	NASA/GSFC
14. Chaudhary	Aashish	Kitware, Inc.
15. Cheng	Hua Qiong	BNU China
16. Christensen	Cameron	SCI Institute/University of Utah
17. Chunpir	Hashim	DKRZ
18. Cinquini	Luca	NASA/JPL and NOAA/ESRL
19. Dart	Eli	DOE/ESnet
20. DeLuca	Cecelia	NOAA/ESRL
21. Denvil	Sébastien	CNRS/IPSL
22. D'haeseleer	Patrik	DOE/LLNL
23. Doutriaux	Charles	DOE/LLNL
24. Drach	Robert	Independent
25. Duffy	Daniel	NASA/GSFC/Center for Climate Simulation
26. Dwarakanath	Prashanth	National Supercomputer Centre
27. *Evans	Ben	NCI/Australian National University
28. Ferraro	Robert	NASA/JPL
29. Fiore	Sandro	CMCC
30. Frazier	Tim	DOE/NIF & Photon Science
31. Fries	Samuel	DOE/LLNL
32. Haley	Mary	NSF/NCAR
33. Harney	John	DOE/ORNL
34. Harris	Matthew	DOE/LLNL
35. Hester	Mary	DOE/ESnet
36. *Hnilo	Jay	DOE BER Headquarters
37. Hoang	Anthony	DOE/LLNL
38. Jefferson	Angela	DOE/LLNL
39. Kershaw	Philip	NCAS/BADC
40. Kharin	Slava	CCCma, Environment Canada
41. Kindermann	Stephan	German Climate Computing Center, DKRZ
42. Kleese-Van Dam	Kerstin	DOE/PNNL
43. Kostov	Georgi	NOAA/NCDC
44. Koziol	Benjamin	NOAA/ESRL NESII/CIRES
45. Krassovski	Misha	CDIAC, Oak Ridge National Laboratory
46. Lacinski	Lukasz	University of Chicago
47. *Lee	Tsengdar	NASA Headquarters
48. Levavasseur	Guillaume	IS-ENES/IPSL
49. Mattmann	Chris	NASA/JPL
50. Maxwell	Thomas	NASA/Goddard
51. McCoy	Renata	DOE/LLNL
52. McEnerney	James	DOE/LLNL

53. Mendes	Wanderley	INPE
54. Nassisi	Paola	CMCC
55. Nienhouse	Eric	NSF/NCAR
56. Nikonov	Serguei	NOAA/GFDL and Princeton University
57. O'Brien	Kevin	NOAA/PMEL University of Washington/JISAO
58. Oh	Jaiho	Pukyong National University
59. Painter	Jeffrey	DOE/LLNL
60. Peterschmitt	Jean-Yves	LSCE/IPSL
61. Plieger	Maarten	KNMI
62. Pobre	Alakom-Zed	NASA/GSFC/NCCS
63. Potter	Gerald	NASA/GSFC
64. Ruthkoski	Traci	Amazon Web Services/Scientific Computing Team
65. Schweitzer	Roland	Weathertop Consulting, LLC
66. Sénési	Stéphane	CNRM
67. Smith	Brian	DOE/ORNL
68. Stockhause	Martina	German Climate Computing Centre (DKRZ)
69. Taylor	Karl	DOE/LLNL/PCMDI
70. Wei	Min	National Meteorological Information Center, China Meteorological Administration
71. Weigel	Tobias	Deutsches Klimarechenzentrum
72. Williams	Dean	DOE/LLNL

*** Funding agency program manager**

12.2 Online Attendees and Contributors

Last Name	First Name	Affiliation
01. Ananthakrishnan	Rachana	University of Chicago
02. Ambrose	Stephen	NASA/GSFC
03. Balaji	V.	NOAA/GFDL and Princeton University
04. Burrows	Susannah	DOE/PNNL
04. Cheng	Huaqiong	Chinese Academy of Meteorological Sciences
06. Curri	Endrit	University of Hamburg
07. Greenslade	Mark	IS-ENES2/IPSL
08. Kharin	Slava	CCCma, Canada
09. *Koch	Dorothy	DOE/BER/CESD Program Manager
10. Kravtiz	Ben	DOE/PNNL
11. Levavasseur	Guillaume	IPSL/LSCE
12. Marra	Osvaldo	CMCC
13. McEnerney	James	DOE/LLNL
14. Nadeau	Denis	NASA/GSFC
15. Palav	Vinayak	Palav
16. Ramthun	Hans	DKRZ
17. Rathmann	Torsten	DKRZ
18. Rood	Richard	University of Michigan
19. Rutledge	Glenn	NOAA/NCDC
20. Schuster	Douglas	NSF/NCAR
21. Sénési	Stéphane	MeteoFrance/IS-ENES
22. Thiede	Annemarie	
23. Toussaint	Frank	DKRZ
24. Wu	Qizhong	Beijing Normal University
25. Zhang	Bao	Beijing Normal University
26.	Marco	

*** Funding agency program manager**

12.3 Conference and Report Organizer

Dean N. Williams, LLNL

12.4 Program Managers in Attendance

Dr. Justin Hnilo, DOE BER HQ
Dr. Tsendgar Lee, NASA HQ
Dr. Daniel Barrie, NOAA HQ
Dr. Ben Evans, NCI HQ
Dr. Dorothy Koch, DOE BER HQ

12.5 Joint International Agency Conference Committee

Luca Cinquini, NASA/JPL, ESGF
Chris Mattmann, NASA/JPL, CMAC
Sébastien Denvil, IS-ENES2/IPSL, ESGF
Ben Evans, ANU/NCI, ESGF
Cecelia DeLuca, NOAA/ESRL, CoG
Giovanni Aloisio, IS-ENES2/CMCC, ESGF

13 Awards

13.1 External Awards

In 2013, the ESGF community won the Federal Laboratory Consortium (FLC) technology transfer award for outstanding partnership. This year, 2014, the UV-CDAT community won the FLC technology transfer award for outstanding partnership and technical achievement. UV-CDAT's development was fueled by exponential increases in the computational and storage capabilities of high-performance computing platforms and the evolution of climate simulations toward high numerical fidelity, complexity, and volume. These technological advances are coming at a time of explosive growth in climate data; experts estimate that tens or hundreds of exabytes of climate data will be created by 2025. The primary partnership that brought UV-CDAT to life consists of Lawrence Livermore, Lawrence Berkeley, Los Alamos, and Oak Ridge national laboratories; NASA's Goddard Space Flight Center; NOAA's Earth System Research Laboratory; New York University; the University of Utah; Kitware Inc.; and Tech-X Corporation.

13.2 Internal Awards

Every year, climate's software engineering community comes together to determine who has performed exceptional or outstanding work in the successful development of community tools for the acceleration of climate science in the data science domain space. These awards recognize dedicated members of our community who are contributing nationally and internationally to our efforts. Recipients of the awards capture and display the best of our community spirit. Recognition of their efforts through these awards is but a small token of our appreciation.

13.2.1 2014 Award Winners for the Success of ESGF

Prashanth Dwarakanath, Linköping University, and Nicolas Carenton-Madiec, IPSL, won awards for their outstanding contributions and leadership in the continued development of the ESGF Node installer. Through the

development of the installer, they led the release of ESGF 1.7 and 1.8 and are currently integrating other components, such as CoG and UV-CDAT, into the ESGF node for the upcoming release of ESGF 2.0.

Luca Cinquini, NOAA/ESRL, won an award for the design and implementation of the new ESGF CoG user interface. This effort will enable many disparate projects to create their own project workspaces while seamlessly integrating their data archives with other science and project domains.

Torsten Rathmann, DKRZ, won an award for his ongoing commitment to helping the user community, through both technical and scientific assistance. We recognized that technologies such as ESGF are useless without the commitment of those like Torsten to help in answering user and community questions on a daily basis. Over the year, he has proven himself to be a technically competent and professional ESGF consultant.

Stephen Pascoe, BADC, won an award for his years of ESGF leadership excellence in the EU and abroad. He was also recognized for his significant contributions to many areas of ESGF development, such as user support, replication, versioning, publication, and installation, to name only a few.

13.2.2 2014 Award Winners for the Success of UV-CDAT

Charles Doutriaux, DOE/LLNL, won an award for his outstanding contributions in the redesign and porting of the backend graphics to VTK. In addition, he was also recognized for his leadership in the continued development of CDAT and other science related integration tools, such as the Climate Model Output Rewriter. These tools are staples of the UV-CDAT framework and the community at large.

Aashish Chaudhary, Kitware Inc., won an award for his contributions to the building, installation, and testing of the UV-CDAT overall product, which went above and beyond expectations. His implementation and utilization of Kitware projects such as CMake, CTest, CDash, and VTK and his leadership contributed to the successful release of UV-CDAT 2.0 for the community.

Thomas Maxwell, NASA/GSFC, won an award for his outstanding contributions to 3D visual displays. These visuals are invaluable to the knowledge discovery process and play an important role in addressing data complexity.

Claudio Silva and the VisTrails Team, NYU, won an award for the development and integration of the VisTrails workflow and provenance capture tools into UV-CDAT. In addition, they were also recognized for their development of UV-CDAT's Graphical User Interface.

Acknowledgments

The conference organizers wish to thank the national and international funding agencies for providing travel funding for attendees to join the conference in person, LLNL for hosting the annual event, and presenters for their contributions to the conference and to this document. The organizers also wish to acknowledge Angela Jefferson for her help in processing endless paperwork, finding the conference location, and arranging many other important logistics. We also wish to acknowledge and thank the LLNL video/media services group for their setting up and breaking down of presentation equipment and the Technical Information Department technical writers for taking detailed conference notes.

The development and operation of ESGF and UV-CDAT continue to be supported by the efforts of principal investigators, software engineers, data managers, projects (i.e., CMIP, ACME, CORDEX), and system administrators from many agencies and institutions worldwide. Primary contributors to these open-source software products include: ANL, ANU, BADC, CMCC, DKRZ, ESRL, GFDL, GSFC, IPSL, JPL, Kitware, NCAR, NYU, ORNL, LANL, LBNL, LLNL (leading institution), Tech-X, and the University of Utah. There are

many other organization and institutions that have contributed to the efforts of ESGF and UV-CDAT. Apologies if you are not among the listed. The DOE, NASA, NOAA, IS-ENES2, and the Australian National Computational Infrastructure provide major funding for the community software efforts.