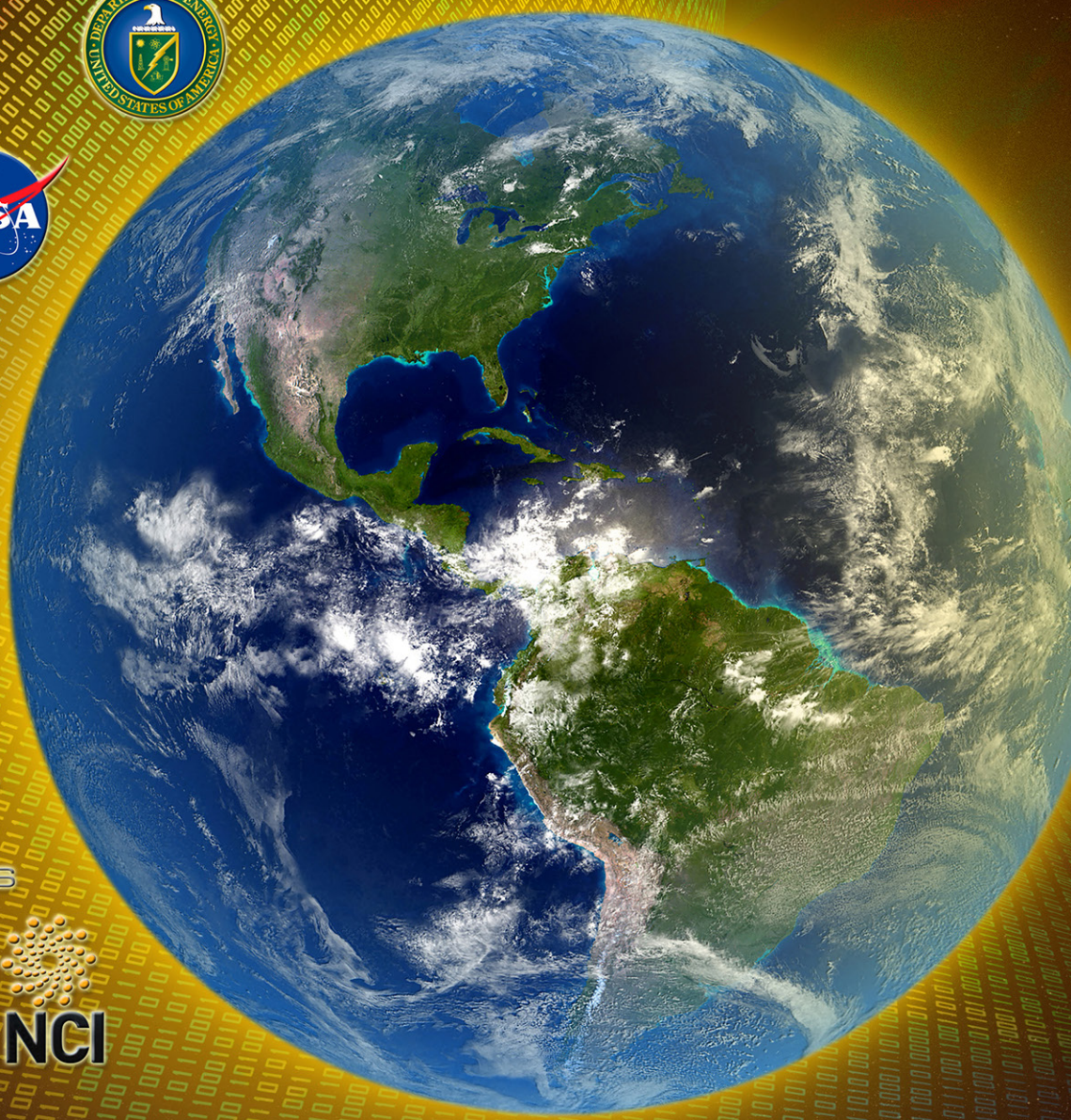
5TH ANNUAL Earth System Grid Federation



DECEMBER 2015

Face-to-Face Conference Report

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 extreme-scale scientific data.

5th Annual Earth System Grid Federation Face-to-Face Conference

December 7–11, 2015
Monterey, California

Convened by

U.S. Department of Energy Office of Science (DOE)
U.S. National Aeronautics and Space Administration (NASA)
U.S. National Oceanic and Atmospheric Administration (NOAA)
U.S. National Science Foundation (NSF)
Infrastructure for the European Network for Earth System Modelling (IS-ENES)
Australian National Computational Infrastructure (NCI)

Workshop and Report Organizers

Dean N. Williams (Chair, DOE)
Michael Lautenschlager (Co-Chair, German Climate Computing Centre)
Sebastien Denvil (Institut Pierre-Simon Laplace)
Luca Cinquini (NASA/NOAA)
Robert Ferraro (NASA)
Daniel Duffy (NASA)
Cecilia DeLuca (NOAA)
V. Balaji (NOAA)
Ben Evans (NCI)
Claire Trenham (NCI)

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 also was undertaken in collaboration with international government-funded organizations, namely from Australia, Germany, the United Kingdom, France, China, Japan, Netherlands, and Italy, along with other U.S. government agencies including NASA, NOAA, and NSF.

Suggested citation for this report: ESGF. 2016. *5th Annual Earth System Grid Federation Face to Face Conference Report*. Earth System Grid Federation.

5th Annual Earth System Grid Federation Face-to-Face Conference Report

December 7–11, 2015

Monterey, California

Publication Date: February 2016

U.S. Department of Energy Office of Science
U.S. National Aeronautics and Space Administration
U.S. National Oceanic and Atmospheric Administration
U.S. National Science Foundation
Infrastructure for the European Network for Earth System Modelling
Australian National Computational Infrastructure

Contents

Preface by ESGF Executive Committee Chair

The Fifth Annual Face-to-Face Conference of the Earth System Grid Federation (ESGF), a global consortium of international government agencies, institutions, and companies dedicated to the creation, management, analysis, and distribution of extreme-scale scientific data, was held December 7–11, 2015, in Monterey, California.

The conference brought together 90 professionals from 15 countries (see **Table 1**, this page) to share their experiences, learn from one another, and discuss the future development of interagency software infrastructure for the climate and weather communities. The conference was structured around small, facilitated session presentations and town hall–like sessions to allow all participants to enter into practical discussions. Special town hall discussion panels were formed to address the specific needs of the community and to help set and focus future ESGF development and implementation plans. In addition, the panels provided insights into how robust, distributed data, computing libraries, applications, and computational platforms can be used in widely varying community projects.

Table 1. Summary of Conference Attendees

Country	Attendees
1. Australia	2
2. Canada	2
3. China	3
4. Denmark	2
5. France	7
6. Germany	6
7. Ireland	1
8. Italy	4
9. Japan	5
10. Netherlands	1
11. Norway	3
12. Spain	4
13. Sweden	2
14. United Kingdom	4
15. United States	44

The conference, a key activity for ESGF and many external projects integrated within the ESGF software infrastructure, aimed to provide value both to veteran attendees and the new and growing network of professionals interested in extreme-scale federated data approaches. The event also offered a glimpse into the activities of ESGF's many working teams and an opportunity to learn from participants' diverse experience. The conference was attended by a wide range of climate community members, including climate and weather researchers and scientists, modelers, computational and data scientists, network practitioners, and others interested in incorporating interagency federated service approaches into their work and policies.

Feedback from the conference was very positive. Participants greatly appreciated and enjoyed the chance to meet like-minded people from so many countries and organizations, to network and learn from one another, and to explore new and exciting ideas.

On behalf of everyone involved in organizing the conference, I would like to express my sincere thanks to the conference facilitators. When we organized the conference, one of our prime aims was to let people learn from each other. The facilitators played a key role in making this happen. I also want to thank everyone who attended for giving so freely of themselves and their experiences and making the conference a memorable and successful occasion. All of us made new friends and contacts, and this can only contribute to the global development of ESGF.

The 2016 conference, to be held Washington, D.C, is now eagerly awaited.

Dean N. Williams
ESGF Executive Committee, Chair

Executive Summary

The purpose of the Fifth Annual Earth System Grid Federation (ESGF) Face-to-Face (F2F) Conference was to present the most recent information on the state of ESGF's software stack and to identify and address the data needs and gaps for the climate and weather communities that ESGF supports. This was the first meeting at which members of the international community, under international ESGF Executive Committee leadership, came together to host the conference. The organizers believed that the ESGF F2F Conference was an ideal gateway for agencies and science project leads to express their current and future data infrastructure needs to tool developers. The organizers also envisioned that attendees would brainstorm and share enhancement plans for data distribution, analysis, visualization, hardware and international network infrastructure, standards, and utilization of assorted community resources. ESGF plays a critically important role in examining cross-cutting integration solutions to address the full spectrum of data lifecycle issues (e.g., collection, management, annotation, analysis, sharing, visualization, workflows, and provenance) that are foundational to achieving interagency and community scientific mission goals.

The conference also helps shape the direction of future ESGF development activities. Nearer-term activities are addressed in **Appendix G**, p. 85, while roadmaps and long-term activities will be described in the 2016 ESGF Implementation Plan that is under development. This plan will describe how the ESGF system will be deployed, installed, and operationally transitioned for several community projects. The plan also will contain an overview of the system, a brief description of the major tasks and overall resources needed to support the implementation effort (such as hardware, software, networks, facilities, and personnel), and any site- and project-specific implementation requirements. In some cases, the plan will be a revision of the current development phase for many of the ~20 ESGF components (or working teams). It includes an integration and testing phase, which will be used for guidance during the implementation phase.

As an integral and important resource for a number of climate and weather initiatives, ESGF supports the Intergovernmental Panel on Climate Change Coupled Model Intercomparison Projects (CMIPs), the Coordinated Regional Climate Downscaling Experiment project, national and international *in situ* and remote-sensing observational projects, reanalysis projects, and dozens of other community-driven projects. The use of ESGF over the past decade has helped generate a virtual explosion of data information and exchanges that have spawned more than 2,000 peer-reviewed papers. Driven by a rapid increase in climate data and new advances in data science, there is now a growing body of ESGF developers (more than 100 worldwide) customizing the processes of climate data activities, performing common analyses, and automating routine data operations. Their work frees researchers to concentrate on scientific understanding and knowledge gathering rather than the tedious chores of data management, movement, and manipulation. The scientific relevance of ESGF has ensured that the F2F conferences are translational in nature and has led to tremendous financial support from a wide range of international and interagency partnerships, both large and small. The primary partners include the U.S. Department of Energy, the Infrastructure for the European Network for Earth System Modelling, the U.S. National Aeronautics and Space Administration, the U.S. National Oceanic and Atmospheric Administration, and the Australian National Computational Infrastructure.

The 2015 ESGF F2F Conference format (see **Appendix A**, p. 41) was very similar to previous ESGF conferences, workshops, and meetings, encompassing a combination of oral and poster presentations based primarily on submitted abstracts. While the abstract-driven program continues to keep scientific presentations at the cutting edge and foster considerable communication, a relatively small number of speakers (4 to 8 per session) were invited to provide overviews and add context to the discussions. Of the roughly 60 oral presentations and posters derived from submitted abstracts (see **Appendix B**, p. 47), at least 25% of them, as expected, were from early-career researchers. As usual, these early-career participants were included in their appropriate scientific context rather than isolated

in a separate session. There was also an ancillary career opportunities session for early-career individuals in which participants from various agencies presented their thoughts on how to succeed in a research career, focusing on climate and data science opportunities.

Previous conferences and their reports (esgf.llnl.gov/reports.html) have been the source of crucial ESGF announcements, directional changes, and implementation plans. The 2015 F2F conference was no different, with the significant announcements of the next full-version release of ESGF (version 2.2), new mandatory software security scans (see **Appendix F**, p. 79), and newly supported projects and their integrated requirements. The variety of data research and technology efforts presented were aimed at meeting the needs of new and ongoing partnerships. Organizers believe that the conference was a catalyst for "big idea" thinking that will help advance ESGF into the future, assist in extreme-scale data management and integration, and support interdisciplinary data science in general.

This report captures participants' innovative thinking by recognizing the community infrastructure investments that support and enable analysis of massive, distributed scientific data collections and that leverage distributed architectures and compute environments designed for specific needs. The report addresses this trend by first recognizing the scientific challenges in the form of diverse and disparate use cases. These use cases capture and emphasize the need for data services, data centers, interoperable services and resources, advanced computational environments, data analytics, data monitoring, multiagency collaboration, and the evaluation of existing tools and services for potential reuse. Conference discussion of ESGF infrastructures primarily revolved around the requirements of the scientific use cases, with CMIP being the most prevalent one. Other use cases focused on observations, model runs, and other model intercomparison projects. While these discussions may cover most use case requirements, they lack the generality to address all the use cases that ESGF will need to support. Therefore, further required enhancements are needed to fulfill ESGF's full scientific vision.

Along with these use cases and infrastructure investment descriptions, this report highlights a number of core findings by conference participants concerning ESGF development needs:

1. **Data storage, server-side analysis, and application programming interface (API):** Develop additional storage and computing resources and a common API that conforms to established federation-wide standards to meet multiple community project goals.
2. **Replication:** Develop precise tool requirements and procedures to ensure successful development, implementation, and use of the automated replication feature. These factors will impose data management constraints for supported projects, such as CMIP and the Accelerated Climate Modeling for Energy project.
3. **Networks, data transfers, and movement:** Support the high-speed transfer of large-scale data sets among ESFG data centers through the implementation of Globus and GridFTP data transfer tools, ubiquitous high-speed Internet between sites, and optimization of local environments.
4. **Data compression:** Implement NetCDF4 data compression for CMIP6 to reduce storage and transfer costs and ensure that critical tools work with NetCDF4.
5. **Performance metrics:** Collect, make available, and use system performance metrics to help ensure end-user and project satisfaction.
6. **Software test suite:** Ensure that each software stack component undergoes efficient but thorough regression, integration, and performance testing and well-formed unit tests. Well-tested software is a matter of community trust for ESGF and its users.
7. **Data security access:** Implement two access security schemes (one standard and the other more lightweight) because only some data and project resources require restricted access.
8. **Software security scans:** Develop and implement a strategy for more frequent and comprehensive security scans of the software stack to detect and resolve vulnerabilities more quickly.
9. **Search:** Develop and implement a search service (including a controlled vocabulary) that can perform customized and saved or shared searches across ESGF nodes.

10. **Operations:** Support node configuration (allowing data providers the ability to select resources for delivery to their community), handler specifications (catering to specific community data sets), data publications (answering questions pertaining to data set visibility), and system and software problem resolution (for porting and operating the ESGF software stack).
11. **Cloud:** Use heterogeneous hardware systems, including emerging cloud technologies and platforms, to help address greater storage and more computing power demands from projects supported by ESGF.

Sections 1–10 of this report briefly summarize conference discussions at a high level, including the findings, which are further elaborated in S**ection 3**, p. 7. For more-detailed information on conference proceedings, see **Appendices A–M**, beginning on p. 39.

1. Background and Introduction

The Earth System Grid Federation (ESGF) was established to address the needs of modern-day climate data centers and climate researchers for interoperable discovery, distribution, and analysis of large and complex data sets, with an emphasis on the use of progressive, outwardly disruptive technologies. As an international consortium of agencies [e.g., the U. S. Department of Energy (DOE), Infrastructure for the European Network for Earth System Modelling (IS-ENES), U.S. National Aeronautics and Space Administration (NASA), U.S. National Oceanic and Atmospheric Administration (NOAA), and Australia's National Computational Infrastructure (NCI)], ESGF's mission is to create an open-source software and infrastructure that powers the study of climate science. Institutions and agencies in other countries also contribute to the development, operations, and success of ESGF including, among others, the Beijing Normal University in China, the Japanese Agency for Marine-Earth Science and Technology, and the Geophysical Institute at the University of Bergen in Norway. These internationally or globally federated and distributed data archival and retrieval capabilities were established under the ESGF banner.

Over the past decade, ESGF work has resulted in the production of an ultrascale data system, empowering scientists to engage in new and exciting data exchanges that ultimately could lead to breakthrough climate-science discoveries. The ESGF distributed archive holds the world's premier collection of simulations, observations, and reanalysis data to support analysis of simulations, including the most important and largest data sets in the global climate simulation community [i.e., the Coupled Model Intercomparison Project (CMIP)], making the archive the leading source for today's climate model data holdings. Through this effort, ESGF was able to achieve its strategic goals[1]:

1. Sustain successful ESGF system operations for multiple climate projects and communities (supports over 40 projects).
2. Address projected scientific needs for data management and analysis.
3. Differentiate between scientific project data management and the overall ESGF data infrastructure governance.
4. Extend ESGF to support the World Climate Research Programme's (WCRP) multiple Model Intercomparison Projects (MIPs), including Phase 6 of CMIP (CMIP6).
5. Support scientific activities locally at individual data nodes (see Appendix E, p. 69), such as the Accelerated Climate Modeling for Energy (ACME) project at DOE Office of Advanced Scientific Computing Research facilities including the National Energy Research Scientific Computing Center (NERSC), Argonne Leadership Computing Facility (ALCF), and Oak Ridge Leadership Computing Facility (OLCF).
6. Support *in situ* and remote sensing observational and reanalysis data activities for future climate applications and projects.

Developing the capabilities to manage and understand massive amounts of global atmospheric, land, ocean, sea-ice, and coupled model data generated by increasingly complex computer simulations and driven by ever-larger qualitative and quantitative observations remains one of climate science's most difficult challenges. Because of rapid increases in technology, storage capacity, and networks and a growing demand for information sharing, research communities are providing access to federated open-source collaborative systems that everyone (e.g., scientists, students, and policymakers) can use to explore, study, and manipulate large-scale data. ESGF stands out from these emerging collaborative climate knowledge systems because of the amount of data it provides (tens of petabytes), the number of global participating sites (more than a few dozen), the number of users (over 25,000), the amount of data delivered to users (over 4 petabytes), and the sophistication of its software capabilities. ESGF is thus considered the leader in both current and future climate data holdings.

[1]Williams, D. N., et al. 2015. "Strategic Roadmap for the Earth System Grid Federation." In *Proceedings of the 2015 IEEE International Conference on Big Data*. Santa Clara, Calif., Oct. 29–Nov. 1, 2015, pp. 2182–90. DOI:10.1109/BigData.2015.7364005.

ESGF's predecessors [i.e., Earth System Grid (ESG), ESG-I, ESG-II, ESG-Center for Enabling Technologies] were critical to the successful archiving, delivery, and analysis of CMIP3 data for scientific publications for the Fourth Assessment Report (AR4) of the Intergovernmental Panel on Climate Change (IPCC). ESGF proved equally important in meeting the data management needs of CMIP5, which provided petascale data information for scientific publications for IPCC's 2013 AR5.[2] Now ESGF is poised to meet the next set of data challenges for CMIP (i.e., CMIP6), with data archival size projected to be around 50 petabytes. Although ESGF has been indisputably important to CMIP, its current and future climate impact is limited not only to this high-profile project. Other scientific projects have benefitted from the archiving, intercomparison, and dissemination of instrument data sets including:

- DOE's Atmospheric Radiation Measurement Best Estimate (ARMBE) and Carbon Dioxide Information Analysis Center (CDIAC).
- NASA satellite data sets such as CloudSat, Microwave Limb Sounder (MLS), Multi-angle Imaging SpectroRadiometer (MISR), Atmospheric InfraRed Sounder (AIRS), and Tropical Rainfall Measuring Mission (TRMM).
- NASA and NOAA reanalysis data sets such as the Modern Era Retrospective-Analysis for Research and Applications (MERRA) and the Cloud's and Earth's Radiant Energy System (CERES), respectively.

In addition to hosting data for a number of other climate projects, ESGF has been used to prototype data delivery and intercomparison services for science communities such as biology, hydrology, and astronomy.

To meet the needs of science communities, ESGF requires integration of software, hardware, and network resources used, created, and maintained by various geographically dispersed research institutions. This ambitious endeavor is being achieved through close group collaborations and marathon national and international face-to-face conferences, teleconferences, and coding and debugging sprint sessions. The 2015 yearly subtask accomplishments and development efforts of each ESGF working team are described in **Section 7**, p. 23.

The climate science community has made large investments in ESGF and its component tools listed in **Section 7**, p. 23. Integrating use case capabilities into the ESGF tool suite through familiar interfaces will further reduce the barriers for large-scale adoption. In addition, relatively simple interfaces are needed for other target audiences (e.g., adaptation researchers, students, and policymakers). With this type of environment, the broad community of researchers and modelers will be able to access and compare popular data products in a highly transparent manner.

[2] Williams, D. N., et al. 2015. "A Global Repository for Planet-Sized Experiments and Observations," *Bulletin of the American Meteorological Society*, early online release. DOI:10.1175/BAMS-D-15-00132.1.

2. Scientific Challenges and Motivating Use Cases

In response to an ESGF Executive Committee request for a list of CMIP and other supported project requirements, the following collection of use cases, largely gleaned from the CMIP5 experience, has been compiled. Many of the issues considered here already have been enumerated in earlier documents, some going back a decade (see **Appendix C**, p. 65). Substantial efforts are being made to address these use cases, and strategies and implementation plans are in place for meeting many of the CMIP6 requirements. The Working Group on Coupled Modeling (WGCM) Infrastructure Panel (WIP) has developed position papers describing how to meet many of the needs in documents available at: www.earthsystemcog.org/projects/wip/resources/. For additional use cases spanning a larger collection of user experiences, see the DOE Working Group on Virtual Data Integration report.[3]

In this section, the focus is on use cases for the Model Intercomparison Project (i.e., primarily CMIP) that underlie the requirements. MIPs capture many of the other project requirements discussed during the science driver town hall session. The order in which the use cases are listed in each subsection indicates their relative importance. In each use case, brackets set off the community involved (i.e., user, node manager, publisher, data provider, CMIP panel, or ESGF). The font color indicates the priority level: **High** – **Medium** – **Low**.

Appendix M, p. 121, contains definitions of some of the identifying terms used in the use cases. The town hall panel used a small number of use cases to help identify the most significant needs in terms of a system to support what the scientific community does and what it produces. The use cases presented here are examples. They have enough detail to motivate specific, actionable requirements for ESGF but are not intended to cover the entire programmatic scope or range of capabilities that might be demanded of operations.

[3]U.S. DOE. 2016. *Working Group on Virtual Data Integration: A Report from the August 13–14, 2015, Workshop*. DOE/SC-0180. U.S. Department of Energy Office of Science. DOI:10.2172/1227017.

2.1 CoG Search

(See **Appendix D**, p. 67, for suggested implementations.)

1. [**User**]: I am only interested in those models that have performed all four Diagnostics, Evaluation, and Characterization of Klima (DECK) experiments. From this subset of models, I want to select those that have performed certain Paleoclimate Modelling Intercomparison Project (PMIP) experiments.
2. [**User**]: I am interested in all the experiments performed as part of PMIP (including any experiments not in the proposal endorsed by the CMIP panel and including all models, whether or not they have met the criteria for participation in CMIP6, as long as they have been recognized by PMIP).
3. [**User**]: I am interested only in models that qualify as CMIP6 models. I want to examine their piControl runs.
4. [**User**]: I am interested in downloading all data sets satisfying my search criteria that have been published since Dec. 8, 2015 (when I last searched the archive).
5. [**User**]: Sometimes I want to download *all* data sets satisfying my search criteria, but most of the time I want to limit my search to only the latest version of each data set.
6. [**User**]: Under various circumstances I would like to select data to download based on search categories such as experiment, variable, frequency, model, and realm. I want to select multiple categories with success determined by "and" logic, and within each category I want "or" logic to apply.
7. [**User**]: I have fast Internet connections to the British Atmospheric Data Centre (BADC) and Institut Pierre Simon Laplace (IPSL) and for now would like to limit my search to only those nodes.

2.2 Errors in Data Sets

1. [**Node manager and publishers**]: An error has been found in one of my previously published

CMIP data sets, and I want to withdraw it (and perhaps subsequently replace it with a corrected data set). For this and related use cases, I want to

- Inform users who have downloaded my data (and registered for notification services) that the data have been withdrawn.
- Make sure that future users do not mistakenly download the withdrawn data.
- Make sure all replicas of my withdrawn data are also withdrawn.
- Enable users to cite my data in a way that makes clear to everyone which version of my data they used in their study.

2. [**Node managers and users**]: I need to know which error checks and quality assurance (QA) tests have been applied to data sets of interest to me. How can I find out if the data
   - Appear to be complete and error free.
   - Have been withdrawn because nontrivial errors were found in data or coordinate information.
   - Have been withdrawn because errors were found in metadata.
   - Have been flagged (but not withdrawn) because "nonfatal" errors were found in the metadata.
   - Also, how can I learn which QA tests have been performed on the data?
3. [**User**]: Six months have passed since I downloaded all available data satisfying certain search criteria (and I know the date when I last checked availability). I would like to enter the same search criteria and the date that I last checked the archive to learn which new data sets have become available and which data sets have been withdrawn.
4. [**User**]: I have identified and downloaded data satisfying certain search criteria. I want to be notified if those data are subsequently found to be in error, or if any of the files I have downloaded have been withdrawn or replaced.
5. [**User**]: I have a list of tracking IDs from all the files I have downloaded (but I do not know when I downloaded those files). I want to find out whether any of the files subsequently have been withdrawn and, if so, why they were withdrawn.
6. [**Data provider**]: I have too few resources to comprehensively check all the data I produce (for compliance with CMIP standards and to discover possible errors introduced in my post-processing procedures). I would like Earth System Grid Federation (ESGF) to check my data and notify me of any errors discovered by users.

2.3 Reducing Effort in Downloading Data

1. [**User**]: I plan to carry out a variety of research and want to minimize how much data I download. To perform the studies I am planning, I need
   - Climatological monthly mean precipitation rates.
   - Several surface variables over North America for the historical period.
   - Zonal mean winds for the historical period.
   - Global mean temperature for the historical period.
   - Zonal wind fields of 200 hPa for the historical period.
   - Note: In the future, I might need to regrid all the model output to a common grid or to generate multimodel ensemble means, but those capabilities are lower priority.
   - Only data from the year 1980 of the historical period.
   - For practical reasons, I would prefer the data reduction implied by the above be performed server side to minimize the downloaded volume.
2. [**User**]: I want to be able to set up a chron job–like procedure, whereby I can periodically query the CMIP archive, identify data satisfying certain search criteria, and download any data that I have not already downloaded. Also, I want to flag data sets from my local holdings that have been withdrawn.
3. [**User**]: My connectivity to certain data nodes is poor (either because the node is often "down" or the downloads are really slow), but I need data from those nodes for my research. I would like to be able to download data from all models at high speeds.
4. [**User**]: I am performing a study of how climate models have evolved both in formulation and

performance. I would find it convenient to access CMIP3, CMIP5, and CMIP6 model output and documentation through a single access point.

5. [**User**]: I have found better response from certain data nodes, and I would like to specify my order of node preference in selecting which replica versions of data sets I wish to download.

6. [**User**]: I do not want to be bothered with upgrading my local NetCDF library from 3 to 4. I would like ESGF to convert NetCDF4 files to NetCDF3 files before I download them.

2.4 License Restrictions, Data Citations, and Usage Tracking

1. [**User**]: I am publishing an article based on CMIP6 data. To satisfy a WGCM requirement to credit modeling groups, I need citable references for the data sets I used (e.g., preliminary persistent identifiers [PIDs] assigned at the time of publication and digital object identifiers (DOIs) assigned once the data have been determined to be valid and properly documented and archived). The PID does not need to be uniquely associated with a particular version of the data set cited, but the ESGF should record information that enables recovery of the data sets used based on the PID and the download date. Researchers should be able to determine if any of the files I used in my study have been withdrawn and why (but it is not necessary for the withdrawn data to be preserved unless a DOI has been assigned to it).

2. [**Data provider**]: I am concerned that output from my model might find its way into "dark" archives (outside ESGF) and might be subsequently redistributed. I want to ensure that license restrictions that I have placed on my data are clearly communicated to those downloading my data from anywhere.

3. [**CMIP panel**]: To help gauge the impact of CMIP and plan for future phases, we would like to monitor and make available statistics concerning quantity of data accessed by users, broken down by variable, modeling center, and simulation.

4. [**Data provider**]: It will help me secure ongoing funding in support of CMIP if I can highlight its impact with the following (both for the collection of models and for my model alone):

   - Summary of archive size (including a page showing "data availability").
   - Amount of data downloaded (cumulative and rate).
   - Map of the world showing locations of users who have downloaded data.

5. [**CMIP panel**]: For future planning, the panel would like to know which variables collected in the past have never been analyzed. Also, for which variables have output from an ensemble of simulations by a single model been used? It would be useful to record the number of downloads per variable across all models for each experiment.

2.5 Managing Data Nodes and the CMIP Archive

1. [**Node manager**]: I have had a disk failure and expect recovery of my data to take several months. My local backup data might also have been lost. Can ESGF bail me out?

2. [**Node manager**]: My site will be down for maintenance for two days. I want ESGF to cover for me, so users can (1) learn that I have temporarily unavailable data of interest to them and (2) access a replica hosted elsewhere.

3. [**Node manager and node publisher**]: I need step-by-step instructions (a "recipe") explaining how to stand up a data node and publish CMIP6 output. I also need to know what to do if I find errors in my data and want to withdraw and possibly replace some of it. I may have questions about publication procedures and need access to a help desk or similar service.

4. [**ESGF and CMIP panel**]: A node manager, publisher or provider modifies an already published file (to correct a problem), but the manager fails to change the tracking ID and to follow the correct procedures for publishing a new version of the data set containing that file. Furthermore, the manager fails to notify others in ESGF about the actions taken. To discover issues like this, an automated procedure needs to be put in place to periodically

check whether the check sums on files match the tracking IDs as recorded in the ESGF catalog.

5. [**CMIP panel**]: A model has been listed by the CMIP panel as qualifying for contributing to CMIP6 but has failed to perform certain mandatory experiments or conform to critical aspects of the experiment protocol. The CMIP panel needs to remove this model from those that appear as qualifying for CMIP.

2.6 Miscellaneous

1. [**User**]: I need an overview of CMIP, including information about the participating models; the experiments; the output produced and its limitations; and, most importantly, how I should go about selecting data to download and then downloading that data.
2. [**User**]: I have downloaded model output from 15 models and 2 different CMIP experiments. Where can I find documentation about those model runs?
3. [**User**]: I have encountered difficulties with ESGF and need help.
4. [**User**]: To compare models to observations, I would like the data I download to be reported on a common grid, or I would like access to tools that facilitate regridding model output to a common grid.
5. [**User**]: I want to determine which models have completed each of a certain set of experiments and have produced the model output I need for my research (including which years are available from each simulation). Having web access to a summary like the one produced for CMIP3 (www-pcmdi.llnl.gov/ipcc/data_status_tables.htm) or information similar to that available at the Swiss "dark" archive (iacweb.ethz.ch/staff/beyerleu/cmip5/) would be sufficient.

3. Conference Findings

The 2015 ESGF F2F Conference brought together international experts from the government, public, and private sectors to address four objectives: (1) suggest recommendations for multiple project requirements that can enhance climate and weather data dissemination at international, national, and regional scales; (2) clarify policy and development directions and interventions that guide efforts toward the mission of sustaining interdisciplinary data projects; (3) provide funding agencies with concrete, actionable development recommendations attuned to current science driver challenges; and (4) identify ways in which ESGF can position itself to reap the full benefits of community collaborations and build on its comparative data delivery advantages as a whole.

The science driver presentations at the meeting described a range of data requirements needed by multiple projects, including short- and long-term storage, tracking, quality, and provenance capturing. Discussions during the town hall science driver session centered on how to more easily integrate these and future disruptive technologies into ESGF and to schedule their integration in a way that maximizes their utility for all current and upcoming projects. Some of these technology imperatives are fairly independent. Rather than integrating them thoroughly as part of a monolithic system, they can be modularized and linked into the ESGF software stack as separate components. Thus, an important architectural finding is that by compartmentalizing various modules, the community can pick and choose only those components required by their organization or project. This approach reduces the impact of security concerns and increases ESGF's build and operational flexibility.

By the end of the conference, many such findings were noted and associated with the projects driving the needs. Additional findings also came out of conference presentations, posters, workshop reports, expert testimonials, use cases, and associated discussions. Data-intensive activities are increasing at data centers and high-performance computing facilities, and interoperable services are critical in enabling these activities. Key findings from the conference attendees are summarized briefly in **Table 2**, this page, from the perspective of identifying the sponsor-funded investments most likely to positively impact the mission and science goals of multiple community projects.

Table 2. Summary of Conference Findings

Topic	Finding
Application Programming Interface	Generalizing the current operational ESGF infrastructure into a template architecture is important so that each implementation layer is hidden behind a well-defined application programming interface (API). Thus, different communities may decide to adopt or swap any single part (or component). If the component changes, the system will handle it as long as the API remains consistent. The API will enable users to determine which version of a component is in use and exchange that component with an older version, as operations dictate. The API also will allow competition between components (e.g., enabling users to decide which computing component to invoke to derive an outcome). The API will make the components independent and ESGF more flexible.
Archive Size, Timeline, and Expected Requirements	To provide the necessary development efforts and resources for operations, the ESGF community must receive the anticipated archive sizes, timelines, and minimum requirements for project success. Currently, reliable estimates are not available to ESGF for archive sizes, timelines, or expected requirements for CMIP6, ACME, Collaborative REAnalysis Technical Environment Intercomparison Project (CREATE-IP), or other large-scale projects that ESGF supports. For greater project success, the ESGF community must have concrete archival numbers, timelines, and expected requirements as soon as possible for system and operational adjustments. For CMIP6, some groups already have conducted DECK experiments and are waiting to begin data processing and ESGF publication. A year to 18 months is an unacceptable time period for integrating services and asking user groups to wait for CMIP6 data.

Table 2. Summary of Conference Findings	
Topic	**Finding**
Cloud	Many users hinted at the possibility of partnerships with public cloud facilities such as Amazon, Yahoo, and Google. However, some indicated that the use of clouds in the ESGF environment is too expensive and will have unresolved scalability issues. To complete the search for greater storage and more computing power, ESGF must be able to perform on heterogeneous hardware systems, including emerging cloud technologies and platforms. Through API cloud resources, ESGF can be tailored to specific project requirements to meet growing user demands. Ultimately, the ESGF environment must be amenable to stronger adoption of cloud deployment.
Data Compression	Data compression is important to ESGF in terms of data storage and transfers. Because of the sheer size of ESGF archives, compressing data for storage or transfer considerably reduces overall costs. Therefore, the CMIP community has determined that CMIP6 data will use NetCDF4 data compression. To return NetCDF3 (classic) to users, ESGF must provide a converter from NetCDF4 to NetCDF3 in the retrieval process, which may require additional compute resources from the major CMIP data centers. To estimate how much disk space will be saved under the NetCDF4 data compression, ESGF will compress a sampling of CMIP5 data using NetCDF4. Extrapolating the calculation throughout CMIP5's entire archive (~2 petabytes) will give a good estimate of how much disk space will be saved when compressing CMIP6 data. ESGF must ensure all critical tools work with NetCDF4.
Data Quality	ESGF data quality persists in the form of provenance, quality control (QC) checks, errata, and data citations. Various components help to improve data quality checks in the ESGF publishing process. EzCMOR (Climate Model Output Rewriter) is one such software package that may be connected to ESGF to enable QC checking before publishing to ESGF. In the context of Observations for Model Intercomparisons (Obs4MIPs), ESGF could allow observational experts to more easily submit information pertaining to the data. In this way, QC tools would be brought to the data publisher and could be exploited during automated publication. Within Ultrascale Visualization–Climate Data Analysis Tools (UV-CDAT), ESGF could conduct standard QC tests relevant for all kinds of data. Someone prepping data could use the information even before publishing into ESGF. This process would help to facilitate the QC process.
Data Security	Some federated data cannot be openly accessible to the public. To secure the data, users must register with any of the ESGF nodes and use their assigned OpenID to login via a web browser or client application such as UV-CDAT. For data and projects that do not require secure access to ESGF data and resources, ESGF will install a more lightweight security scheme. Users still will have to authenticate for use metrics purposes, but they will receive unobstructed access to unsecure data and project resources. With construction of the new Node Manager, the security system must be able to access the attribute or authorization service provided by the federation and its new set of requirements.
Data Storage	The sheer size of current and expected future archives makes storage a difficult issue to address. If the expected storage for CMIP6 is over 10 petabytes (with estimates as high as 50 PB), then a uniform storage strategy must be put into place among the major CMIP data center sites—Lawrence Livermore National Laboratory (LLNL), Deutsches Klimarechenzentrum (DKRZ), Centre for Environmental Data Analysis (CEDA), and National Computational Infrastructure (NCI). This includes the purchasing of storage units and possible archiving of data on tape for long-term data preservation. At the storage level, multiple optimizations and ways to approach input/output subsystems are required, including reimplementation of the high-performance storage systems access and retrieval by users. However, having most data accessible from rotating disks is still desirable. All projects stated that they did not have enough computational and data storage resources, with data storage resources being at the top of their needs list. ESGF must access data from distributed heterogeneous storage systems, including mass storage systems, move data efficiently among storage systems, and manage the content of data in the storage spaces.

Table 2. Summary of Conference Findings

Topic	Finding
Data Streaming	Data streaming has the potential to speed up operations as domain sizes grow, wherever they are performed. These data transfer operations will include a set of possibilities ranging from explicit user requests to automated streaming driven by an interactive data exploration process. In addition, specialized data transfer mechanisms for understood data formats (e.g., NetCDF4) will allow progressive streaming of multisolution representation that feeds directly into interactive data exploration components. Visualization is one example of streaming information. For example, Google Maps streams coarse resolution of highly detailed data. Once the user requests specific data, then high–resolution data is streamed for clearest viewing. This approach balances speed with level of detail.
Data Transfers and Movement	A critical requirement is that the ESGF data centers support the transfer of large-scale data sets among their sites for replication. Replication transfer requests are expected to be made automatically and by site administrators. GridFTP servers located on site data transfer nodes (DTNs) will provide download options to facilitate high-speed disk-to-disk data transfers. GridFTP for disk-to-disk download is configured with custom security handlers that allow use of ESGF security services. Globus will be the preferred data movement service for end users. It builds on the GridFTP data transfer mechanism and provides secure and reliable high-performance data transfers from data centers to specified user destinations (e.g., other data centers, laptops, and clusters). This includes auto-tuning transfers to ensure best performance based on transfer size and the number of files per transfer. ESGF must have easy movement of data to data centers from the modeling centers. Therefore, dozens of worldwide modeling centers will also need, in some capacity, to be connected via high-speed Internet to the data center and be GridFTP-enabled for reliable data transfers. From a software architecture perspective, making data transfer and movement function as a separate component keeps ESGF independent of disruptive data transfer technologies and maintains its flexibility.
End-Use Requirements	A better definition of the end user is needed, prepared by the supported ESGF projects, to better capture end-user requirements. However, the projects are not completely sure who the consumers of their data products are (or will be), and the consumer lists vary over time. Another issue is determining the estimated data volume and when the data must be delivered to the community for certain types of user analysis and data evaluation. This issue will become more essential as ESGF expands its ability to perform server-side (i.e., remote) analysis to reduce data movement. Each user community has a different set of requirements. ESGF no longer can make the same early assumption that end users are experienced scientists in the climate community who understand all aspects of the data they are retrieving or manipulating for knowledge discovery. Moving forward, there must be a concerted effort to define the level of competence of each end user.
ESGF	Projects may expect ESGF to do everything in terms of data handling, but that simply is not possible. ESGF is a layered baseline architecture on top of which other components build and integrate within their architectural scheme. With this construct comes a project-specific infrastructure development and management plan that must stay true to the original concept of ESGF's peer-to-peer architecture principles of modular components and standard API protocols. Projects cannot rely on ESGF to do it all. They must balance their own vetting, debugging, operation, and support processes to offset costs. ESGF must focus on executing fewer things flawlessly rather than doing many things merely adequately. This means making sure the requirements keep the failure rate low and the uptime high for the tens of thousands of end users and dozens of diverse projects.

Table 2. Summary of Conference Findings	
Topic	**Finding**
Funding and Management	With joint funding under DOE, IS-ENES, NASA, NOAA, and NCI, research scientists and software engineers apply a diverse base of expertise to develop software and serve data that enable climate research groups worldwide to access, visualize, and analyze large data sets. Segments of the ESGF community have envisioned opportunities for expanding or spinning off ESGF components to meet the needs of projects and subcommunities, and ESGF as a whole needs to support these endeavors and prepare to help meet these demands. This need for support also applies to funding and management of ESGF's diverse base of experienced software developers and climate researchers. All opportunities in computing hardware, networks, software, people, and other resources must be realized and managed constructively in this highly collaborative problem-solving environment. As an example of funding opportunities, the U.S. team has successfully written a DOE proposal (titled "DREAM: Distributed Resources for ESGF Advanced Management") to access large data sets across multiple DOE, NASA, and NOAA compute facilities. This effort will immediately improve climate research as well as numerous other data-intensive applications in the United States, Europe, Asia, and Australia.
Hardware	A cost-benefit analysis is needed for long-term storage. For example, what would it cost to regenerate versus store the data? If storage is cheaper, then what type of storage is appropriate for specific data usage (i.e., tape storage versus rotating disks)? If regeneration of data is more cost-effective, then what type of computing and cloud services will be needed? This option would include determining the benefits of putting additional resources next to high-performance computing (HPC) facilities, and at what cost? What level of access control will the community have over the system? In this scenario, the hope would be to secure free computing and low-cost data storage. Some science communities have aggressively expanded their hardware infrastructure to include cloud technology, which means that ESGF software must be flexible in deploying, expanding, adapting, and managing cloud solutions as part of its vast global ecosystem of hardware, networks, software, and services. In today's pricing, hardware is so expensive that no one project could afford it alone. Thus, the cost would have to be shared among diverse groups and projects.
Bringing ESGF Back Up Efficiently and Effectively	To bring back up and sustain ESGF operations, two Confluence documents were generated: (1) the ESGF Back to Operations Plan and (2) the ESGF Back to Operations Nodes Checklist. Document 1 identifies two operational tiers. Tier 1 consists of LLNL, DKRZ, the British Atmospheric Data Centre, Institut Pierre Simon Laplace, Jet Propulsion Laboratory, and Linköpings Universitet's National Supercomputer Centre in Sweden (LIU). Tier 2 consists of all other ESGF nodes, such as Geophysical Fluid Dynamics Laboratory, Oak Ridge National Laboratory, Argonne National Laboratory, Royal Netherlands Meteorological Institute (KNMI), and Centro Euro-Mediterraneo sui Cambiamenti Climatici (CMCC). All Tier 1 nodes will have an identity provider, while Tier 2 nodes most likely will not. Document 2 is a checklist of action items for verifying the running of an operational node. Both documents are for **internal use only** and are designed to inform ESGF administrators of the operational status of other nodes.

Table 2. Summary of Conference Findings	
Topic	**Finding**
Network	ESGF requires the ability to control the timing of data- and network-intensive replication operations for large climate data sets. For many projects, the ability to move large amounts of data from end to end (disk to disk) at high speeds for globe-spanning distances is critical. For this effort, the International Climate Network Working Group (ICNWG) project evolved to develop and test end-to-end capabilities for next-generation networks via the ESGF data nodes. One of ICNWG's main goals is to enhance data transport technology based on application requirements to ensure that the climate community is ready for the next generation of high-bandwidth networks. The challenge is to invoke the GridFTP transfer protocols to transfer multiple files concurrently as well as use parallel transmission control protocol streams for replication operations among the major CMIP data centers. For higher transfer throughput (for both network and storage), transfer queues and concurrent connections must be managed. When a data set has a wide range of file sizes, continuous data flow from the storage into the network can be achieved by prefetching data from storage onto the transfer queue of each concurrent transfer connection. Setting up the optimal level of concurrency is important, especially for an application with varied file sizes. Regression and performance testing need to be better integrated into the ICNWG effort. Additionally, the Globus transfer application must be integrated with DTNs and the ESGF software stack.
Operations	Operational support is needed to sustain the numerous ESGF nodes operated by simulation, observation, and reanalysis projects. Such support includes (1) node configuration, enabling data providers to pick and choose resources to be delivered to their community; (2) handler specifications, catering to specific community data sets; (3) data publications, answering questions pertaining to the visibility of data sets to end users; and (4) system and software problem resolution, for porting and operating the ESGF software stack. As the ESGF team transitions from version 2.x development activities to production and operations, it is tasked with making data available to all users seeking to understand, process, extract value from, visualize, and communicate the information to others. This ongoing effort, though daunting in scope and complexity, greatly magnifies the value of numerical climate model outputs (e.g., CMIP5 and CMIP6) and climate observations (e.g., Obs4MIPs) for future national and international climate assessment reports.
Performance Metrics	Performance metrics must be included as part of ESGF operations. The goal is to have displayable and well-understood performance metrics to track and monitor the overall system and to gather data transfer performance metrics among major CMIP data center sites. Performance operations also must include overall system robustness monitoring and functionality benchmarking to ensure end-user and project satisfaction. Performance must be tied into additional management tools such as the dashboard and allow for performance-tuning knobs wherever and whenever possible for finer adjustment of ESGF nodes.
Persistent Identification	Persistent identifiers (PIDs) for digital resources could play a vital role in identifying and tracking the usability of data in ESGF. PIDs also will incorporate associated services such as modeling group, resolution, and other metadata bindings. Exactly what will be tracked by the different projects has yet to be determined. These identifiers must be persistent, locatable, and actionable in the long term. Thus, this process could prove to be costly over time and will require an overall digital preservation strategy.

Table 2. Summary of Conference Findings	
Topic	**Finding**
Provenance Capture	Provenance capturing is necessary for reproducing complex analysis processes at various levels of detail in a shared environment. For ESGF, a comprehensive provenance infrastructure is needed to maintain detailed history information about the steps followed and data derived in the course of an exploratory task. ESGF will need to maintain provenance of data products and the workflows used to derive these products and their executions. Although this information could persist in several forms—including JSON and XML files or in a relational database—it must allow users to intuitively navigate workflow versions, undo changes without losing results, visually compare different workflows and their results, and examine the actions that led to a result. The provenance component also must gather information derived in the scientific discovery process and enable users to manage and exploit this information to understand relationships, anomalies, trends, and features contained in specific climate data. Because provenance contains useful information that can help users understand previous analyses as well as construct new ones, these mechanisms for querying provenance are essential.
Replication	Before ESGF can properly prepare for automated data replication and distribution, supported projects must address issues that affect the underlying replication development and operation process. Key issues include the granularity of file duplication (i.e., do users want to replicate every file every time?) and the amount of data replicated (e.g., if the projected CMIP6 archive is 50 petabytes, then replication of all the data will be impossible). ESGF already operates a data replication vehicle, but the automated process for replication is still under development. It is expected to be ready for CMIP6. Clear guidelines for the replication process are needed, along with a replication organization strategy that will help coordinate and manage file replication among the CMIP data center sites. The latter requirement is particularly important if ESGF is to manage the replication of a 50-petabyte archive that probably will be distributed across multiple centers (i.e., no single site will have all the data but may have, at most, up to 20 petabytes). Such a situation would require a highly coordinated effort for data movement to ensure easy access, security, and backup. This coordinated strategic effort will need documentation, training, and online tutorials for ESGF administrators to prevent them from improperly replicating or removing data. Provenance information is essential for data that must persist to ensure proper and consistent data handling by administrators.
Resource Management	By managing resources (e.g., hardware, network, storage, and software), users can direct and adapt the intended ESGF services for short- and medium-term effects on targeted experiments when needed most. The long-term objective is for ESGF climate projects (e.g., ACME and CMIP6) to adopt resource management services into production and successfully establish test beds for other supported climate projects. While developing ESGF and testing it for the community, the ESFG team will gather feedback in a participatory process for setting community resource requirements and priorities accepted by targeted groups and other stakeholders. These resource requirements will further help ESGF formulate objectives as the project moves forward to meet scientific challenges. Currently, management of processing resources is done on the same server as the processing, using Open-source Project for a Network Data Access Protocol (OPeNDAP), a tool that weighs quite heavily on ESGF data resources. Movement to a more distributed resource management scheme will better leverage cloud environments and HPC processing power.
Searches and Controlled Vocabulary	Developing and exploiting a service (including a controlled vocabulary) that supports unified user-defined data searches across ESGF nodes will let users customize, save, and share their searches. The ESGF search service is based on the latest version of Solr—a search engine used by many commercial-grade websites and applications. Many applications connect to ESGF via its Solr API. For climate modelers, the CMIP6 community is working to standardize a set of controlled vocabularies that will need to be integrated into ESGF.

Table 2. Summary of Conference Findings	
Topic	**Finding**
Server-Side Analysis (and Derived Data Sets)	The size of some data sets makes moving most of the needed data to the end user's home institution infeasible. Data analysis therefore must be performed remotely. By running analyses entirely remotely, server-side HPC, cloud, and cluster resources can handle larger data sets independent of local resource quality. In addition, remote visualization and analysis capabilities will enable researchers to perform their work on the data sets where they exist, minimizing the need to transfer and store multiple copies of large data sets. Simulation and observation data sets are rarely at a single site, so analysis will require remote access to and seamless retrieval of data and tools from multiple data centers, archives, and investigator ESGF sites. Each site may have subtle or even major differences in protocols and calibrations, affecting their use in any validation or analysis. Integration of such data will require robust metadata and tools that can effectively use the metadata to perform any necessary conversions, renormalizations, or other transformations.
Software Security Scans	While ESGF cannot prevent a dedicated hacker from trying something new, the project is coming up with new strategies to scan for known code vulnerabilities. The latest software security breach has necessitated an inventory of all software in the ESGF software stack, and ESGF developers have coordinated component development to combine and share information about existing vulnerabilities that may affect secure ESGF operations. This security event has kicked off the new Software Security Scanning Working Team to develop a strategy for quickly detecting code security violations. Led by NASA, the ESGF software stack will be scanned routinely every quarter in addition to the twice-a-year code scanning conducted prior to each major ESGF version (i.e., code) release. (The ESGF software stack is scheduled for only two major releases a year.) These additional security scans are expected to detect and resolve software vulnerabilities more quickly, thus minimizing the impact on ESGF operations by allowing ESGF to recover rapidly if there is a security issue. If a security issue does arise, coordinated notification will be sent to each ESGF administrator (at each ESGF node site) and the ESGF Executive Committee via PGP-encrypted email.
Software Test Suite	Many ESGF components are maintained by a broader development community and contain a rich suite of testing frameworks and methodologies. For example, the ESGF data publishing utility offers a variety of data publishing tests and confirms that the configuration of data resources is being published correctly. However, other components of the ESGF software stack are not being tested for proper installation or use. To address resiliency in the ESGF software stack, each component requires *regression*, *integration, and performance* testing in addition to well-formed unit tests. These tests must be coupled with facility software stack implementation to ensure that ESGF is installed properly and that data can be downloaded and manipulated properly. Increased robustness is needed in all areas of ESGF software testing to ensure proper data and component lifecycle and use. From conference discussions, these tests are proving to require more time and resources than anticipated; therefore, ESGF and the community at large must carefully and cleverly manage them.
Training and Documentation	Training is important to ensure proper data use and dissemination. ESGF already provides useful installation workflow, documentation, and training videos, but they must be maintained as ESGF evolves to meet the demands of its supported communities. Training support on how to access and use the data is ongoing and costly. There have been several attempts at automation with websites and Internet forums, but none have worked as well as direct e-mail responses to questions. Work is still in progress to find an up-to-date resolution for administration and customer training and documentation of ESGF.
Use Metrics	Use metrics help projects know how the community is using their hardware, software, network, data, and other resources. Metric information such as number of users will serve as base metrics for various data and services within ESGF. Service-specific metrics also should be defined to measure the usage and adoption of specific capabilities and to evaluate their usefulness. Another important metric is identification of the number of software packages provided by other institutes accessible via ESGF. As server-side services come online, the ability to measure the usability of computational and visualization tools for specific user tasks will become imperative. The ESGF dashboard capability is already in effect, providing data on users, downloads, and published data for designated projects.

4. ESGF Data Center Challenges and Motivating Use Cases

At the ESGF F2F Conference, five of the major CMIP5 ESGF data centers presented their views on ESGF challenges and motivating use cases for climate (model) data infrastructure developments. A summary of these challenges and use cases is presented in Table 3, this page. (See **Appendix E**, p. 69, for more details.)

Table 3. Summary of Challenges and Use Cases Presented by CMIP5 ESGF Data Centers

Institution	Country	Challenge	Relevant Use Case
Lawrence Livermore National Laboratory	**United States**	• Optimizing network performance • Providing compute resources for data site petabyte storage, data processing, and end-user data analysis tools	• Accelerated Climate Modeling for Energy (ACME) • CMIP6
National Computational Infrastructure	**Australia**	• Migrating to transdisciplinary data • Creating an integrated scientific computing environment • Synchronizing data • Organizing multipetabyte data archives	• National Environmental Data Interoperability Research Platform • CMIP6
Climate Computing Centre	**Germany**	• Merging compute and data services into one system • Supporting long-term data archiving and citation • Integrating an improved version of the CMIP5 quality assurance software • Implementing persistent identifiers based on the Corporation for National Research Initiatives' (CNRI) Handle Server	• National MIP data analysis platform • CMIP6
Institut Pierre Simon Laplace	**France**	• Comparing data between climate models, ground observations, and satellite observations • Providing a national academic platform to analyze CMIP6 outcomes • Adapting ESGF developments to the CMIP6 and IPCC AR6 timelines	• Operating the national data analysis environment • CMIP6
Centre for Environmental Data Analysis	**United Kingdom**	• Maintaining data quality • Supporting end users • Handling a variety of product requests • Handling "big data" volume, velocity, and variety	• European Space Agency's Climate Change Initiative • OPTImisation environment for joint retrieval of multi-sensor RADiances (OPTIRAD) project for container technology • CMIP6

5. ESGF Data Center Requirements and Findings

5.1 Needs

All five data centers represented in **Table 3**, p. 15, identified data replication, computing, and integration of facilities close to the archive as central ESGF development needs. Further, all centers develop applications on their ESGF data nodes to serve national user communities. These applications seem to be closely related to national funding streams. Serving the CMIP6 global data federation is common to all five of these data centers.

CMIP6 and ESGF data management requirements include (1) publication of national CMIP6 data and replica from international contributions, (2) data quality assurance, (3) early data citation for ESGF data publication, (4) digital object identifier (DOI) minting for DataCite data publication and CMIP6 data archiving, and (5) PID-based data services such as data management services (for versioning and replication). ESGF must make certain that released infrastructure is robust and well-documented to support the stable, consistent operation of ESGF data nodes (preferably 24/7); provide training; ensure quality assurance across the federation; and inspire the confidence of both users and data sites in ESGF operations. Achieving these aims involves a stronger focus on end-user requirements, such as end-analysis tools and compute resources at ESGF (replication) data nodes.

No key technology gaps or problems were identified for existing technology beyond the limited person power for ESGF development and maintenance. This limitation necessitates a careful prioritization of future development items in the ESGF implementation plan.

5.2 Missing Information

CMIP6 data management constraints: Although the basic principles are clear from position papers of the Working Group on Coupled Modelling Infrastructure Panel (WIP) and discussions between WIP and ESGF's Executive Board, concrete specifications were not discussed at the ESGF F2F conference. For example, specifications of mandatory attributes for CMIP6 NetCDF files must be finalized before the implementation of constraints into the ESGF data publisher.

Final timeline of CMIP6/Intergovernmental Panel on Climate Change Sixth Assessment Report (IPCC-AR6): Knowing these dates is necessary for the ESGF implementation plan and integration of CMIP6 requirements.

Quantity of CMIP6 data: CMIP6 is expected to include more data than CMIP5, both in total size and number of data entities. Data are likely to increase from CMIP5 to CMIP6 by a factor of 20 to 50. For CMIP5, ESGF contained 1 to 2 PB of climate model data, which comprised 5 million individual files. If the data increase by a factor of 20, then CMIP6 would include 30 PB of data and 100 million files; if a factor 50, then 75 PB of data and 250 million files. Whatever the increase, the total data amount is too large to store in one place and the number of files too large to manually correct any inconsistencies in the archive. These vast amounts of data files demand that data and data responsibilities be shared among the ESGF replication nodes in CMIP6.

5.3 Replication and Storage

Based on these numbers, there was an intensive discussion about ESGF data replication, focusing on automated replication, replication strategy, and storage media. The automated replication process requires workflows and tools that run on the ESGF replication data nodes and take into account local infrastructure constraints. ESGF normally does not have its own infrastructure; it runs as part of an already existing legacy environment. In this case, the technical aspect is to optimize the existing network paths from DTN to DTN. Experience from past optimizations show that the most critical bottlenecks are in the local compute environments, which normally are not optimized to transfer large amounts of data at network peak performance.

The replication strategy answers the question of which replication node archives which data. This strategy takes into account the available storage and network resources at each ESGF replication node

to minimize the data transfer. Another aspect in the strategy is data security. An option might be to store two or three data sets across the CMIP6 federation.

Storage media might help if nearline storage on tape libraries such as high-performance storage systems (including disk cache and data storage on spinning disks) is taken into account. Nearline storage is an intermediate type of data storage that represents a compromise between online storage (supporting frequent, very rapid access to data) and offline tape storage or archiving. Usage of nearline storage requires that scientific CMIP6 data management decide on data storage priorities. High-priority data would be stored on spinning disks and lower-priority data in nearline storage.

6. Computational Environments and Data Analytics

ESGF F2F Conference presentations from the Compute Working Team (CWT) described their motivation for and approach to integrating HPC data analytics capabilities with ESGF. Team members discussed the necessary components for creating these computational environments, types of exposures to the analytics, and potential architectures.

6.1 Motivation

The amount of data being created throughout the scientific community, especially within climate research, has grown dramatically since the last IPCC assessment report (AR). For instance, the NASA Goddard Institute for Space Studies (GISS) team created about 80 terabytes of data for AR5. The same team wants to double the resolution of their code, which will result in an eightfold increase in the amount of data to be produced for AR6—that is, one model alone will contribute more than half a petabyte of data to the next assessment. Using this metric, a reasonable expectation is an eight- to tenfold increase in the amount of data stored throughout ESGF in the next few years.

The sheer size of these data sets makes moving the data to local end-user environments for analytics infeasible. Therefore, data analytics must be performed where data sets reside; the results would then be transferred back to end users. To accomplish the integration of data analytics as a service within ESGF, three critical components are required:

1. **Relevant data sets:** The data sets within ESGF are assumed to be of extreme relevance to the climate community and future climate research. Without data valued by the research community, ESGF would not exist in the first place.
2. **Exposure:** In addition to the data, the exposure of analysis routines through a standard API is necessary. Using the API, end users can perform standard analysis of data sets while also building up more complex analytics across very large data sets.
3. **High-performance computing and storage:** To execute the analysis services exposed through the API in a timely fashion, a combination of high-performance computing and storage is required.

6.2 Exposure

ESGF has begun creating a web processing service (WPS) API to expose analytics. WPSs are used heavily throughout the open geospatial community. Based on relatively simple use cases (described in **Section 2**, p. 3), such as averages and anomalies, an initial WPS specification has been created. **Table 4**, p. 20, shows an example of the specification and how users would begin to address such complexities as the variable, domain, and axes of interest along with output formats.

This is just a subset of parameters that could be used to specify analytics. To perform analytics over multiple data sets, data must be regridded. Therefore, regridding methods also must be exposed through the API and the back end equipped with sufficient compute and data resources to perform these computationally intensive operations. **Table 5**, p. 20, gives an example of how the API will begin to expose regridding options to the end user.

6.3 High-Performance Computing and Storage

Many ESGF member sites have experience with HPC-based data analysis of model output and high-performance storage designed for large-scale climate models. However, the combination of these two activities creates a major strain on traditional HPC systems.

Models generate huge amounts of data during a large-scale climate simulation. Most HPC environments thus have data systems optimized for large streaming writes and reads. However, analytics typically will slice through the data in a number of different ways, resulting in small-block, random input/output. These two competing requirements place significant strain on traditional HPC environments and necessitate different practices.

In addition to these constraints, security is a major factor in selecting a method for large-scale data analytics exposure to ESGF. For example, HPC centers probably

Table 4. Web Processing Services Parameter Exposed Through the ESGF Compute Working Team's API

Parameter	Syntax

Variable	{
	"uri":"<address to the data file|collection://collection_name>",
	"id":"<variable name in the file>[:user defined unique identifier]",
	"domain":"<domain id specified below>"
	}

Domain	{
	"id":"<user defined domain name identifier>",
	"<level|time|latitude|longitude|axis_name(s)>":
	{
	"start":<int/float/string>",
	"end":<int/float/string>,
	"step":<int>,
	"crs":"<indices|values>"
	}
	}

Axes	"<x|y|z|t|(axis_name)>"
Format	"<opendap|netcdf|png>"

Table 5. Regridding Options Exposed Through the API

Parameter	Syntax
GridderTool	"<ESMF>" Default: ESMF
GridderMethod	"<conservative|linear|nearestneighbor>" Default: linear
Grid	"<string|[:user defined unique identifier]>" Default: T85

will not want to expose the same compute and storage fabrics that are running the models to an ESGF analysis service. Potential approaches for accommodating these competing requirements were discussed during the ESGF F2F Conference, including:

1. **Dedicated HPC resources:** Creating a smaller, separate high-performance cluster dedicated to ESGF analytics will resolve the resource contention issue for model runs and will ease security concerns. Computing centers know how to create these environments and can even use older systems to save money. Moreover, traditional cluster-based file systems (e.g., POSIX®) can be used to expose analytics using existing tools, such as UV-CDAT. NASA Goddard is working on exposing UV-CDAT capabilities through a traditional HPC cluster using a climate data analytics framework. One benefit of this approach is that data can be used in their native format and do not have to be altered.
2. **Shared-nothing environment:** Riding the wave of the emerging "big data" analytics stack from other industries, NASA Goddard has investigated using

the Apache™ Hadoop® ecosystem with climate data. An initial service has been created and primitives are exposed through a Python™ API so that users do not have to write MapReduce routines. Industry and universities have made major contributions to and investments in this ecosystem, which could play a vital role for certain types of analytics. While it is typical to sequence data into a Hadoop file system (which means changing the data), Goddard is exploring the use of both native scientific data and HPC shared file systems, such as Lustre and General Parallel File System.

3. **Array-based storage model:** Ophidia is a research effort by the Euro-Mediterranean Center on Climate Change designed specifically to address large-scale climate data analytics. At the core of this approach is an array-based storage model using data cube operators. Scientific data do have to be rewritten when stored in Ophidia. Beyond storage, Ophidia provides a robust environment for data management, complex workflows, interfaces to climate analysis operators and primitives, and programmatic access via C and Python.
4. **Web-based services:** IS-ENES has created a WPS API and a set of web-based services integrated with ESGF to help users explore climate data and perform analytics. This web-based approach uses traditional file system- and server-based processing capabilities but brings them together in a robust environment for users to search, visualize, and perform calculations across large data sets.

The ESGF community realizes that no one method for high-performance storage and analytics will fit everyone's needs. Therefore, it is possible, and expected, that many of the above architectural approaches will be enabled throughout ESGF, all consistently exposed through a standard API. Further, not all sites will have compute capabilities, but some sites within ESGF are expected to have a significant amount of resources to perform server-side analytics by CMIP6. However, given the trends in storage and computing, all ESGF sites quite possibly could have some type of analytic capabilities in the future.

7. Technology Developments

Presentations from the leads of several ESGF working teams constituted a central part of the ESGF F2F Conference. Each team described its development progress over the past year and its intended roadmap for 2016. This section gives a brief overview of accomplishments and plans across all teams, with further details provided in **Appendix G**, p. 85.

For the ESGF community, 2015 was dominated by the June security incident, which prompted node managers to take all ESGF nodes offline, bringing system operations to a halt. While the incident was an obvious setback to the ESGF brand, it was an opportunity for developers to work on hardening, improving, and upgrading the software stack. All ESGF modules have been subjected to dynamic and static security scans and protected against all known common vulnerabilities and exposures (CVEs). The ESGF team has formulated a plan for executing these scans periodically, including before every ESGF major release is deployed into production (see **Appendix F**, p. 79). The team also has installed the latest versions of the underlying software libraries and engines used by ESGF (e.g., Postgres, Apache, Tomcat, Java, Python, Django, Solr, and OpenSSL) and implemented a process to maintain and keep them current moving forward. Other critical upgrades relate to the ESGF software itself. For example, because all data had to be republished anyway, the new metadata archives are based on the newest Solr version, which offers better performance and can be upgraded as Solr evolves. Additionally, ESGF made the decision to cease use of and support for the old ESGF "web front-end" user interface and instead brought the system back up directly with CoG as the new web user interface.

The new software stack has been named "ESGF 2.0," as it incorporates a drastic overhaul of security and functionality. Perhaps the most important conference outcome was the formulation of a plan for ESGF 2.0 deployment across all nodes in the federation and for bringing the system back to full operations by spring 2016, in time for ESGF to support the management and distribution of CMIP6 data worldwide. This plan is currently being executed and is on schedule but may slip if unforeseen issues occur.

Looking forward into 2016 and beyond, ESGF developers will be fully engaged in a wide range of activities that will provide expanded functionality, reliability, and performance to the climate community for years to come. Working teams and their tasks follow:

- **Computing Working Team:** Designing a powerful API, based on the Open Geospatial Consortium/WPS standard, for executing remote computations on climate data distributed across the federation.
- **User Interface Working Team:** Supporting the deployment and federation of CoG web portals and enhancing the application as new requirements emerge from ESGF administrators and users.
- **Dashboard Working Team:** Implementing the next-generation architecture for gathering and analyzing usage metrics across the federation. The new architecture will be ready for deployment by early spring.
- **Data Transfer Working Team:** Interfacing with Globus services to enable faster, more reliable movement and downloads of data across the system.
- **Installation Working Team:** Upgrading the installer to keep pace with the evolution of the rest of the ESGF software stack, as well as enabling scalable deployments of ESGF node types in new configurations.
- **International Climate Network Working Team:** Collaborating with the major CMIP6 climate centers to set up a new data node architecture that decouples data movement and replication among centers, from data downloads to end users, for increased overall performance.
- **Identity, Entitlement, and Access Management (i.e., Security Access) Working Team:** Planning a full infrastructure upgrade as they transition the system to OpenID Connect and OAuth, to provide redundancy and fail safes for some of the most critical security services.
- **Metadata and Search Working Team:** Providing full support for CMIP6 data search and discovery and for evolving the architecture to scale to the much larger metadata volumes expected in the future.

- **Node Manager Working Team:** Developing a next-generation, peer-to-peer engine to maintain an up-to-date registry of available services throughout the federation.
- **PID Services Working Team:** Developing new paradigms for querying and tracking ESGF data objects through their lifecycle.
- **Provenance Capture Working Team:** Implementing a framework for capturing detailed information through data publication, analysis, and generation of derived products.
- **Publishing Working Team:** Working on finalizing metadata requirements in support of CMIP6, as well as enabling a new publishing service to support "long-tail" data providers.
- **Quality Control Working Team:** Defining the process for validating and flagging CMIP6 data sets as they are processed through several stages of the data processing pipeline.
- **Replication and Versioning Working Team:** Setting up an infrastructure to execute massive automatic data transfer and publishing across major climate centers, using tools such as Synda and Globus.
- **Software Security Working Team:** Supporting ESGF software security scans, which are critical for minor and major releases of the ESGF software stack.
- **User Support Working Team:** Redefining tools and services that can be better integrated with the new ESGF 2.0 software stack.

Again, please see **Appendix G**, p. 85, for a much more detailed report of the activities of each ESGF working team.

8. Technology Integration of Interoperable Services

Since ESGF's inception, its architectural philosophy has been to develop a highly modular and configurable software system, rather than a single monolithic system. ESGF software stack development began with the integration of popular open-source components and engines—including Postgres, Apache httpd, Tomcat, Solr, THREDDS Data Server (TDS), Live Access Server (LAS), and Ultra-scale Visualization Climate Data Analysis Tools (UV-CDAT)—with services and interfaces developed by ESGF's developer community (e.g., for data publishing, security, searching, and user interfaces). Today, ESGF continues to develop custom components and integrate promising open-source tools into the stack to meet the needs of an increasingly diverse and rapidly expanding user community.

The latest software architecture for an ESGF node is shown in **Fig. 1**, this page. Software components and services are grouped into four broad areas of functionality, called node "types." At installation, ESGF node administrators can choose which node types to install, depending on the specific needs of their institution, such as data volumes, user base, and computational requirements. **Table 6**, p. 26, briefly describes the intended purpose of each node type and which software components it currently includes.

The need for integrating distinct software components into a cohesive, reliable, highly functional system will become increasingly important to ESGF's success as the developer community grows to meet increased demands for expanded functionality and widespread adoption. The 2015 ESGF F2F Conference identified

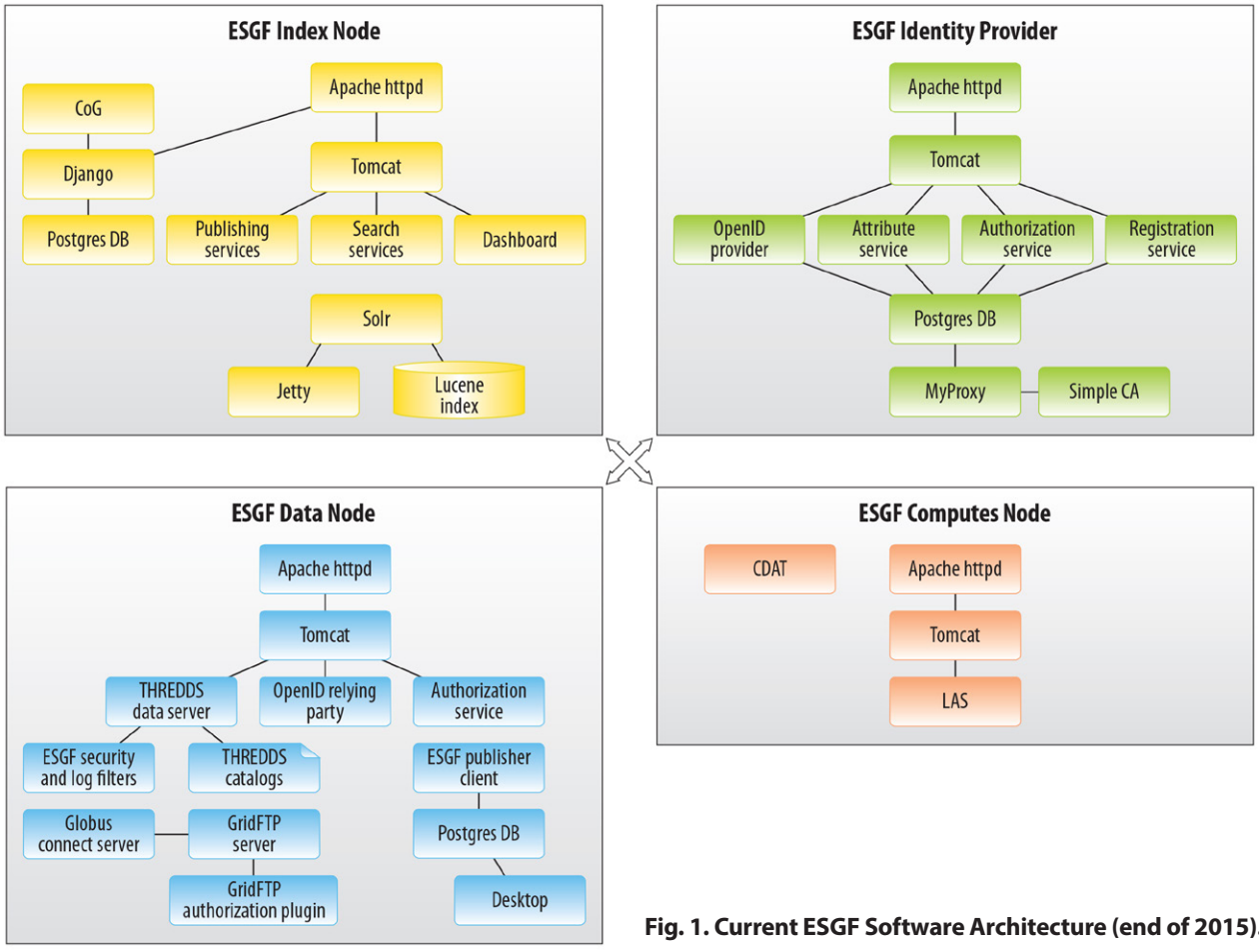


Fig. 1. Current ESGF Software Architecture (end of 2015).

Table 6. Intended Functionality and Implementation of Software Components for ESGF Node Types

Node Type	Node Functionality	Node Components
Index Node	Authorization to publish Metadata publishing Metadata searching System metrics User accounts Web user interface	Authorization Service CoG Dashboard ESGF publishing services ESGF search services
Identity Provider	Access control attributes User authentication X.509 certificates	Identity provider MyProxy and SimpleCA Registration and attribute services
Data Node	Authorization to download Data download Data transfer	THREDDS Data Server Authorization service Globus Toolkit, GridFTP
Compute Node	Data analysis Data visualization	CDAT (i.e., UV-CDAT) and LAS

several critical areas of development that need to be addressed to support the next generation of high-profile international projects, such as CMIP6, the Coordinated Regional Climate Downscaling Experiment (CORDEX), Observations for Model Intercomparisons (Obs4MIPs), and the Accelerated Climate Modeling for Energy (ACME) project. These requirements are described below.

8.1 Enhanced Need for Service APIs

One of the most critical goals of ESGF software development is to expose each service through well-documented, simple, and stable APIs, which enable interoperability between systems and system components. An API is a "contract" that a service exposes to its clients; the service "obliges" clients to always support its functionality and is invoked through a fixed set of signatures, so that clients do not have to worry about changes in service implementation. Services potentially can be replaced with different, more scalable, or better-performing implementations without affecting the clients, as long as the API remains unchanged.

Whenever possible, APIs should be based on international and industry standards, such as OpenID, secure sockets layer (SSL), OAuth, Security Assertion Markup Language (SAML), or web processing service (WPS). Adherence to these standards allows ESGF to interoperate with other external systems—for example, to enable external users with an already existing OpenID to pass ESGF authentication. When standards are not available or widely accepted, ESGF has tried to formulate and document its own APIs. For example, the ESGF search services expose a well-defined query syntax that can be used by clients to investigate the federated metadata archive.

The 2015 ESGF F2F Conference highlighted the importance of API-driven development to support an ever-growing portfolio of interacting services and applications. The following APIs have been identified as critical to ESGF's continued evolution and will have to be further developed and documented in the upcoming months.

- **Security services:** This application enables users and clients to authenticate into the system and receive authorization to download data and execute operations on the data. The security infrastructure needs to support both human users interacting with the system through a browser and, increasingly, machine-to-machine interactions for programmatic downloads and calculations. Some of the most important goals identified by the ESGF Security Working Team for the near future include:
  - Transitioning human authentication from OpenID 2.0 to OpenID Connect and OAuth.
  - Improving documentation and supporting the acquisition of X509 certificates by clients in a variety of programming languages.
  - Revisiting and simplifying the registration process for ESGF access control attributes.
- **Search services:** ESGF's global, federated metadata archive is arguably one of its most important

features, and the ability to efficiently and reliably query the distributed data holdings is critical to many other services and clients, such as data download, computing, and replication. The ESGF search services API already exists and has been used reliably for several years, in support of CMIP5 and other projects. The major challenge for 2016 will be scaling the ESGF search services to the metadata volumes expected from the upcoming CMIP6 model runs and integrate this metadata with metadata holdings from observational and reanalysis data.

- **Publishing services:** Within ESGF, clients publish metadata to a local or remote index node through two possible mechanisms: (1) Requesting the service to harvest an existing metadata catalog ("pull") or (2) sending already-generated metadata documents to the service ("push"). Both mechanisms are exposed to clients through the simple RESTful API, which relies on X.509 certificates for proper authentication and authorization. In the upcoming year, this API will have to be expanded to support metadata validation through controlled vocabularies, atomic metadata updates (that track the quality control evolution of data sets), and registration of PIDs and DOIs at the data set and file levels.
- **Computing services:** One major thrust within the ESGF collaboration is the development of a powerful infrastructure for remote computation, which will allow users and clients to request execution of climate science algorithms onto distributed nodes and access data throughout the federation. Moving the computation to the data is an essential paradigm of any technology infrastructure that needs to enable scientific research over exabyte-size distributed archives. To achieve this goal, the ESGF Computing Working Team is developing an API that leverages the Open Geospatial Consortium OGC/WPS standard to define its own signatures and execute custom operations and diagnostics over data holdings stored throughout the federation.
- **Provenance capturing:** To reproduce complex analysis processes at various levels of details in a shared environment, provenance capturing is necessary. For ESGF, a comprehensive provenance infrastructure is needed to maintain detailed history information about the steps followed and data derived in the course of an exploratory task (for reproducibility). ESGF will need to maintain provenance of data products and the workflows used to derive these products and their remote executions.

8.2 More Complex and Flexible Node Deployments

As described above, an ESGF node currently can be configured to support one or more of four different service types: index, data, identity provider, and compute. One outcome of the conference was definition of the need to enhance the configuration options of an ESGF installation in two ways:

1. Further decomposing some of the current node functionality types—for example, by separating the user interface from the index node and making the authorization service an optional component that can be installed as part of an index node or a data node.
2. Supporting more complex deployment architectures that "mix and match" an arbitrary number of different node types.

As an example, the new desired ESGF node configuration includes the following:

- **No identity provider:** Some institutions have a requirement to tie their ESGF infrastructure into an already existing authentication service that must be compliant with the OpenID/X.509 ESGF protocols; others are prevented from supporting authentication services and must delegate user registration and login to another ESGF node.
- **Cluster of data transfer nodes:** To support more efficient transfers of massive data sets among major climate centers without affecting clients' data downloads, the International Climate Network Working Group and the Replication and Versioning Working Team have proposed a data node architecture that separates the data read and write functionality into separate servers. In this architecture, a cluster of data transfer nodes would be set up with Globus and GridFTP to execute scripted data transfers within the federation and to publish metadata to the index node. Another data node with access to the shared disk would be responsible for serving the data to the community.

- **Additional data nodes:** Although the new TDS version 5.+ has greatly improved the memory management of a large number of catalogs, the centers needing to support large data holdings (such as a full replica of the CMIP6 archive) and a large number of clients still need to distribute and serve data sets from multiple data nodes.
- **Additional computing nodes:** For the same reason, ESGF foresees a not-to-distant future in which a center needs to support intensive data processing initiated by multiple clients working on very large data collections. In such cases, client requests should be routed and split across multiple computing nodes.
- **Cloud of index nodes:** ESGF is investigating the use of Solr Cloud as the most promising architecture for scaling the search services to the size of metadata archives expected to be generated in the next 3 to 5 years. A Solr Cloud configuration involves multiple index nodes connected by a high-speed local network, hosting different pieces (i.e., "slices") of the overall metadata index. Solr Cloud technology offers many desired features, such as automatic distributed indexing and querying, shard redundancy and automatic failover, and overall enhanced performance. The biggest barrier to adoption is that the technology was designed to support a cluster of local servers that are managed by the same administrative staff, as opposed to a system of remote distributed servers under independent administrative control.

8.3 Consolidation of Common Services

Another theme that emerged during the conference was the need to reduce and consolidate some critical ESGF services to a limited number of nodes that can commit to a higher level of support. Currently, there are two classes of services under consideration for possible consolidation:

- **Identity providers (IdPs):** The ESGF Single Sign On protocol (currently based on OpenID 2.0 and X.509 certificates) supports an arbitrary number of IdP nodes where users can register and authenticate. In reality, an excessive number of IdPs is actually detrimental to a smoothly operating federation because it increases security vulnerability, administrative support for installation and maintenance, the size of the certificate trust-store bundle, and general configuration issues. For these reasons, the ESGF Security Working Team recommends that the number of IdPs throughout the federation be reduced to no more than two or three on each continent. Other nodes would delegate user registration and authentication to one of the approved IdPs.
- **Index nodes:** ESGF has a requirement for returning consistent query results throughout the system, independently of the index node where a client initially sends a search request. As a consequence, each index node must create a local copy of the metadata index of every other index node and then distribute the query to its local replica shards. The query also can be distributed directly to remote index nodes, but this alternative greatly reduces performance. This challenge is true whether traditional Solr replication or the newest Solr Cloud architecture is used. Clearly, for scalability reasons, the number of federated index nodes cannot be allowed to grow unchecked; rather, the ESGF team will need to select a limited number of nodes responsible for hosting a piece of the federated metadata archive. While other index nodes could still be used to publish local data, their content would not be replicated across the federation.

8.4 Redundancy of Critical Federation Services

The number of some deployed services needs to be reduced, but other services critical to ESGF operations need to provide redundancy and automatic fail safes (i.e., resiliency). The services that most need redundancy include:

- **Search services:** Fortunately, the ESGF architecture already includes multiple redundancies for the federated search services. Metadata indexes are continuously replicated from one index node to all others, so if any node is subject to a temporary outage, a client can target its search to another node and obtain the same results.
- **Attribute services:** ESGF authorization is based on matching resource policies to user attributes. Although the software infrastructure supports multiple policies for the same resource and querying of the same attributes from multiple services, the

current system suffers from a lack of redundancy in managing specific classes of attributes, such as the CMIP5 commercial and research attributes. Specifically, those attributes are maintained at only one node (Program for Climate Model Diagnosis and Intercomparison), and, if that node is temporarily unavailable, no one in the federation can be authorized to download those data. The ability to replicate an attributes database across two or more sites is needed so that when a node is down, the same attributes can be obtained from another location.

- **Data downloads:** Data downloading is arguably one of the most critical functionalities provided by ESGF. Obviously, if a data node is down, no data can be downloaded from that node, but if the data are fully replicated to other sites, clients can discover and download replicas from those sites. One very useful improvement of the ESGF infrastructure targeted for the mid term will enable automatic selection of an available (or the closest) data node for data download.

8.5 Federation-Level Registry of Available Services

The ESGF system encompasses dozens of nodes distributed across the world and multiple services within each node. To interoperate, nodes and services need to trust one another and be aware of each other's endpoints, including the very latest state of a service component. To fulfill this requirement, a new, more scalable, and better-performing node manager component is being developed that will be responsible for continuously exchanging complete node information among all federated nodes. As such, the node manager will be a required installed component on all ESGF nodes, no matter the node type. A prototype version of the new node manager is expected before summer 2016, with ubiquitous deployment of a fully functional version by the end of the year. When operational, the new node manager will greatly reduce the effort needed to set up and maintain the federation configuration at each node and will enable nodes to dynamically join and leave the federation as they are set up, updated, or reconfigured.

8.6 Development of Client Tools and Applications

Finally, a growing trend within the ESGF community is to develop higher-end applications that leverage existing ESGF services to provide a targeted function or serve a specific scientific community. Example tools include:

- **Synda:** Command-line tool to search and download data from the ESGF distributed archive. As such, Synda relies on the ESGF search services infrastructure and includes a pluggable mechanism for using different data transfer protocols: scp, http, and GridFTP. Synda has been selected as the core application for enabling automatic replication of CMIP6 data sets across the federation.
- **Climate4Impact:** Web portal that enables visualization of climate model data sets specifically targeted to the climate change impact assessment and adaptation communities. This portal relies on several critical ESGF services, including searching, authentication, authorization, and data download via OPeNDAP. There also have been discussions about integrating the core Climate4Impact functionality as part of a standard compute node deployment.
- **Ultra-scale Visualization–Climate Data Analysis Tools (UV-CDAT):** Rich desktop client that enables complex analysis and visualization of climate data sets, including those stored on distributed ESGF nodes. UV-CDAT also uses the ESGF security, search, and data download services and combines these with powerful data reduction and subsetting capabilities to enable high-performance data access.

9. Community Developments and Integration

9.1 CMIP6 Requirements from Conference

High-level ESGF requirements from CMIP6 were discussed both during the 2015 ESGF meeting and prior to it during ESGF Executive Committee and WIP telephone conferences. These requirements are summarized in **Table 7**, this page.

Details on these overall requirement items are in **Appendix H**, p. 105, and the corresponding WIP position papers.

9.2 Central Community Developments

During the "community development and integration" session of the conference, participants from THREDDS Data Server, Synda, and Globus presented their accomplishments, roadmaps, and estimated resource needs to fulfill roadmap milestones. See **Appendix I**, p. 109, for details.

Table 7. High-Level ESGF Requirements from CMIP6

Requirement	Working Team*	CMIP6 Priority	CMIP6 Phase
Hard-coded data rejected in ESGF data publication if there is a mismatch in mandatory quality specifications.	IWT, PWT, CoG, NMWT	High	Data production
Versioning consistent with modifications of data files and PID registration, as well as preservation of metadata for all data set versions.	CMOR, PWT, RVWT, QCWT	High	Data production
Integration of (early) data citation into the data publication process, accessibility at user interfaces, and transition into long-term archives.	IWT, QCWT, CoG, MSWT, RVWT	High	Data production, dissemination, and long-term archival
Integration of PIDs in NetCDF files and PID registration during the data publication process.	CMOR, IWT, PWT, CoG, NMWT	High	Data production
Integration of errata and annotation into the data publication process, accessibility at user interfaces, and transparency of data versions.	IWT, PWT, COG, WPWT, NMWT	High	Data dissemination
On-the-fly format transformation from compressed NetCDF4 into NetCDF3.	CWT, CoG	Low	Data dissemination
Interpolation to regular grids.	CWT, CoG	Medium	Data dissemination
Automated, monitored data replication among ESGF replica nodes.	ICNWG, RVWT, DTWT, NMWT	Medium	Data dissemination
Handling of multiple projects in ESGF search API.	CoG, MSWT	High	Data dissemination
Integration of Earth System documentation (ES-DOC) and the Common Information Model (CIM) standard. (Status unclear, 12-02-2015)	Unclear	High	Data dissemination, long-term archival

*Note: All working teams are ESGF, except for CoG, CMOR, and ICNWG.

Acronyms: IWT, Interface Working Team; PWT, Publishing Working Team; NMWT, Node Manager Working Team; CMOR, Climate Model Output Rewriter; RVWT, Replication and Versioning Working Team; QCWT, Quality Control Working Team; MSWT, Metadata and Search Working Team; WPWT, Workflow and Provenance Working Team; CWT, Computing Working Team; ICNWG, International Climate Network Working Group; DTWT, Data Transfer Working Team.

10. Report Summary and Implementation Plan for 2016

10.1 Working Team Priorities

This section describes the process followed by each ESGF working team to implement and administer its roadmap (**Appendix G**, p. 85) actions in accordance to the scientific challenges and motivating use cases discussed in **Section 2**. Guidance states that after considering a wide range of actions, projects, and improvements for each component in the overall software stack, the plan must describe a set (or subset) of work actions. Inclusion in the strategy or "action plan" is how work will be prioritized, implemented, and administered by each working team. For cross-cutting integrated work actions, such as replication, networks, and the movement and replication of large data sets among major data centers, there must be identifiable action items specific to each team requesting work. The ESGF Executive Committee must sanction all work actions before work can begin.

By mid- to late April 2016, an in-depth implementation strategy (action plan) will be developed by the ESGF Executive Committee for funding agencies to approve (e.g., the ESGF Steering Committee). For each working team, town hall sessions at the conference prioritized specific action items related to community needs (see **Table 8**, this page). Working team leaders were asked to develop an implementation roadmap from conference findings described by attendees (see **Sections 2–5**, beginning on p. 3) and prioritize the list for their respective projects or communities. The implementation roadmap developed by each working team was based on each leader's technical, administrative, political, economic, developmental benefits, costs, and qualitative work analysis of each component's selected work action.

Table 8. Prioritized Working Team Actions, Based on Conference Findings and Project Needs

Working Team	Task and 2016 Timeline Summary	Needed Software	Satisfying Use Case	Finding	Needed for Project	Funding Agency(s)
Computing	• [Feb]: Instantiate ESGF-compliant WPS API for server-side computing (e.g., analysis and visualization) • [Mar]: Analyze back-end implementation for projects [i.e., CDAS, Climate Analytics-as-a-Service (CAaaS)] • [Jun]: Develop specific compute capabilities for projects • [Sep]: Parallelize server-side computing (e.g., MPI, Hadoop) and streaming	UV-CDAT, CDAS, CAaaS, Ophedia, clusters, cloud servers	Section 2.3 Reducing Effort in Downloading Data	Server-side analysis, derived data, resource management, operations, hardware, and data streaming	All MIPs (including CMIP, Obs4MIPs), ACME, CORDEX, others	DOE, NASA, IS-ENES

Table 8. Prioritized Working Team Actions, Based on Conference Findings and Project Needs

Working Team	Task and 2016 Timeline Summary	Needed Software	Satisfying Use Case	Finding	Needed for Project	Funding Agency(s)
CoG User Interface	• [Mar]: Integrate with Globus download services • [Apr]: Integrate with controlled vocabulary • [Jun]: Implement PIDs, DOIs, errata, others • [Dec]: Integrate with ES-DOC model metadata • [Dec]: Integrate with dashboard	CoG, ES-DOC, Globus	Section 2.1 CoG Search	Application programming interface, operations, and search or controlled vocabulary	All MIPs (including CMIP, Obs4MIPs), ACME, CORDEX, others	NOAA, DOE, IS-ENES
Dashboard	• [Feb]: Develop new front-end presentation layer • [Mar]: Implement RESTful API • [Mar]: Implement back-end interaction with Solr (for fine-grained statistics) • [Aug]: Integrate with new node manager • [Dec]: Develop new views for other supported projects (Obs4MIPs, ACME, others)	CoG	Section 2.4 License Restrictions, Data Citations, Usage Tracking	End-use requirements, operations, performance metrics, and use metrics	All MIPs (including CMIP, Obs4MIPs), CORDEX, others	IS-ENES, DOE
Data Transfer	• [Apr]: Integrate Globus with replication tools • [Apr]: Integrate DTN with Globus Connect Sever input/output • [Jun]: Integrate Globus download with CoG • [Dec]: Develop better delegation model to get certificates to transfer	Globus, CoG, DTN-hardware	Sections 2.1 CoG Search; 2.3 Reducing Effort in Downloading Data; 2.5 Managing Data Nodes and the CMIP Archive	Data storage, data transfers and movement, hardware, network, operations, and replication	All MIPs (including CMIP, Obs4MIPs), ACME, CORDEX, others	DOE

Table 8. Prioritized Working Team Actions, Based on Conference Findings and Project Needs

Working Team	Task and 2016 Timeline Summary	Needed Software	Satisfying Use Case	Finding	Needed for Project	Funding Agency(s)
Identity, Entitlement, and Access Management	• [Mar]: Pilot integration of Globus and OAuth 2.0 services • [Apr]: Pilot integration of compute services (UV-CDAT, LAS) with OAuth 2.0 services • [May]: Implement discovery mechanism for OAuth 2.0, deprecating OpenID 2.0 • [Jul]: Deploy OAuth 2.0 operationally with IdPs • [Dec]: Implement, integrate OpenID Connect into ESGF	Globus, CoG, UV-CDAT, LAS, OAuth 2.0, OpenID, Connect		Data security, data transfer and movement, and operations	All MIPs (including CMIP, Obs4MIPs), ACME, CORDEX, others	IS-ENES, DOE
Installation	• [Jun]: Integrate new node manager into the ESGF installer • [Nov]: Replace bash installer with modular Python installer • [Sep]: Create build scripts for individual component for greater flexibility	All software modules		Hardware, how to bring ESGF back up efficiently and effectively, and operations	All MIPs (including CMIP, Obs4MIPs), ACME, CORDEX, others	IS-ENES, DOE, NASA
Climate Network	• [Dec]: Integrate network infrastructure into replication process • [Dec]: Replicate the DTN network infrastructure and Tier 1 and Tier 2 ESGF node sites	Globus, GridFTP, TN-hardware	Sections 2.3 Reducing Effort in Downloading Data; 2.5 Managing Data Nodes and the CMIP Archive	Data transfers and movement, hardware, network, operations, performance metrics, and replication	All MIPs (including CMIP, Obs4MIPs), ACME, CORDEX, others	DOE, IS-ENES
Metadata and Search	• [Mar]: Develop tools and services to support atomic metadata updates • [Mar]: Support tagging of data sets for multiple projects • [Jun]: Implement data validation against controlled vocabularies • [Dec]: Support partitioning of search space across multiple virtual organizations	CoG and Solr	Sections 2.1 CoG Search; 2.2 Errors in Data Sets; 2.5 Managing Data Nodes and the CMIP Archive	Application programming interface, operations, search or controlled vocabulary	All MIPs (including CMIP, Obs4MIPs), ACME, CORDEX, others	NOAA, DOE

Table 8. Prioritized Working Team Actions, Based on Conference Findings and Project Needs

Working Team	Task and 2016 Timeline Summary	Needed Software	Satisfying Use Case	Finding	Needed for Project	Funding Agency(s)
Node Manager	• [Feb]: Complete the development of major features (e.g., shard files, security, dynamic super node selection, installer integration) • [Jul]: Integrate with dashboard and other components • [Dec]: Release into production		Sections 2.2 Errors in Data Sets; 2.4 License Restrictions, Data Citations, and Usage Tracking; 2.5 Managing Data Nodes and the CMIP Archive	Operations, performance metrics, use metrics	All MIPs (including CMIP, Obs4MIPs), ACME, CORDEX, others	DOE, IS-ENES
Persistent Identifier Services	• [Apr]: Integrate with publisher, errata service, node manager • [Jun]: Set up RabbitMQ queuing system and Handle service operations • [Aug]: Develop additional tools for offline message publication and end-user information viewing	CoG	Sections 2.2 Errors in Data Sets; 2.4 License Restrictions, Data Citations, and Usage Tracking	Data quality, operations, and persistent identification	All MIPs (including CMIP, Obs4MIPs), CORDEX	IS-ENES
Provenance Capture	• [Feb]: Incorporate time-series system environment metrics store • [Mar]: Develop performance metrics reporting user interface • [May]: Develop language bindings for ProvEn Client API • [Jun]: Design and integrate services supporting harvesting of provenance from native source types	CoG, UV-CDAT	Sections 2.2 Errors in Data Sets; 2.3 Reducing Effort in Downloading Data; 2.6 Miscellaneous	Data quality, operations, provenance capture, replication, sever-side analysis, and derived data sets	ACME	DOE

Table 8. Prioritized Working Team Actions, Based on Conference Findings and Project Needs

Working Team	Task and 2016 Timeline Summary	Needed Software	Satisfying Use Case	Finding	Needed for Project	Funding Agency(s)
Publication	• [Feb]: Develop new esgscan-directory tool for map-file generation • [Mar]: Implement schema changes to support publisher integration with Errata and PID services • [Mar]: Implement changes to support new TDS features, DTNs, high-performance storage systems (HPSS) • [Apr]: Develop new drs_lite tool for versioning, management of Data Reference Syntax (DRS) • [Apr]: Integrate user interface publication within CoG • [Jul]: Implement data set versioning	CoG, DRS, TN-hardware, HPSS-hardware	Sections 2.1 CoG Search; 2.2 Errors in Data Sets; 2.4 License Restrictions, Data Citations, and Usage Tracking; 2.5 Managing Data Nodes and the CMIP Archive	Application programming interface, operations, provenance capture, replication, and search or controlled vocabulary	All MIPs (including CMIP, Obs4MIPs), ACME, CORDEX, others	DOE
Quality Control	• [Apr]: Version support, persistent metadata • [Jun]: Errata service • [Sep]: Data citation service	CoG	Sections 2.1 CoG Search; 2.2 Errors in Data Sets; 2.4 License Restrictions, Data Citations, and Usage Tracking	Data quality, operations, and persistent identification	All MIPs (including CMIP, Obs4MIPs), CORDEX	IS-ENES
Replication and Versioning	• [Apr]: Implement Synda replication testbed between Tier 1 sites (i.e., major CMIP data centers: BADC, DKRZ, LLNL, NCI) • [Jun]: Integrate with publisher, errata service, and CoG • [Sep]: Set up operations for RabbitMQ queuing system, Handle service	CoG, Globus, Synda,DTN-hardware	Sections 2.1 CoG Search; 2.5 Managing Data Nodes and the CMIP Archive	Archive size, timeline, expected requirements, data storage, data transfers, movement, hardware, network, operations, and replication	All MIPs (including CMIP, Obs4MIPs), ACME, CORDEX, others	DOE, IS-ENES

Table 8. Prioritized Working Team Actions, Based on Conference Findings and Project Needs

Working Team	Task and 2016 Timeline Summary	Needed Software	Satisfying Use Case	Finding	Needed for Project	Funding Agency(s)
Software Security	• [Feb]: Identify team lead, team members, mission statement • [Feb]: Write software security plan that defines roles, responsibilities, and processes for security reviews of future software release • [Apr]: Implement software security plan	All software modules		Hardware, how to bring ESGF back up efficiently and effectively, and operations	All MIPs (including CMIP, Obs4MIPs), ACME, CORDEX, others	NASA, IS-ENES
User Support	• [Mar]: Create infrastructure for connecting front-line support with team leads for more rapid response to customers • [May]: Overhaul wiki, websites • [Jul]: Transition, integrate the support working team FAQ site to the CoG front end	CoG	Section 2.6 Miscellaneous	End-use requirements, operations, training, and documentation	All MIPs (including CMIP, Obs4MIPs), ACME, CORDEX, others	DOE, IS-ENES

10.2 Preparing the Infrastructure for Development

Ensuring that the data infrastructure is optimally designed to enable the development of new, validated, and verified capabilities with proven technology (as described in **Appendix B**, p. 47, **and Table 4**, p. 20) is just as important as the tasks described in the previous subsection. The data and computing infrastructure needs to undergo rapid development and assessment of new scientific modules and to provide a testing-to-production environment for simulation and evaluation (i.e., metrics, diagnosis, and intercomparison) of observational and reanalysis data. Development and use of the overall enterprise and the individual components as stand-alone systems are driven by scientific challenges and requirements, along with a diverse set of climate use cases (**Section 2**, p. 3). Though some tools are specific to a particular project, wherever possible the development teams have identified common methods and similar APIs and developed tools that satisfy the requirements of many projects (as shown in **Table 6**, p. 26).

To achieve individual project and community goals, the ESGF team will continue to build upon and enforce standards and promote the sharing of resources, such as in the case of NetCDF, climate forecast conventions, ESGF, UV-CDAT, ES-DOC, DRS, Globus, and many others. Recognition and use of these open-source projects by the research community are growing, and the tools and experience resulting from these sponsored projects will provide the foundation on which the data infrastructure will be based. ESGF is building a unique, secure, complete, and 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.

Appendices

Appendix A. Conference Agenda

2015 Earth System Grid Federation Face-to-Face Conference *Jointly held by DOE, NASA, NOAA, IS-ENES, and ANU/NCI*	
Time	**Topic**
Monday, December 7, 2015	
10:00 a.m. – 12:00 noon	Registration: Mezzanine/San Carlos 3
2:00 p.m. – 4:00 p.m.	Registration: Mezzanine/San Carlos 3
5:00 p.m.	Meet-and-greet at London Bridge Pub (no host)
Tuesday, December 8, 2015	
7:00 a.m. – 8:30 a.m.	Registration: Mezzanine/San Carlos 3
8:00 a.m. – 8:30 a.m.	Meet-and-greet
8:30 a.m. – 8:40 a.m.	Welcome, safety, introduction, conference charge, and agenda overview — Dean N. Williams • How conference attendees contribute to the conference's final report • Framing of the ESGF F2F Annual Meeting
8:40 a.m. – 8:45 a.m.	DOE opening comments — Justin Hnilo, program manager for the Data and Informatics program within DOE BER's Climate and Environmental Sciences Division
Science Drivers: Project Requirements and Feedback	
8:45 a.m. – 10:55 a.m.	**Science Drivers** *Session Discussion Lead — Dean N. Williams* **8:45 a.m. – 9:15 a.m.** Karl Taylor — WCRP CMIP and WGCM Infrastructure Panel (WIP) **9:20 a.m. – 9:40 a.m.** David Bader — DOE Accelerated Climate Modeling for Energy (ACME) **9:45 a.m. – 10:05 a.m.** Peter Gleckler — Observations for Model Intercomparisons (Obs4MIPs) **10:10 a.m. – 10:30 a.m.** Sébastien Denvil — ENES and Coordinated Regional Climate Downscaling (CORDEX) **10:35 a.m. – 10:55 a.m.** Jerry Potter — Collaborative REAnalysis Technical Environment Intercomparison Project (CREATE-IP) **Questions and discussion of example use-case requirements for each major supporting project** • What are the key things that are difficult to do today and are impeding scientific progress or productivity? • What is your timeline for data production and distribution? • What is the estimated size of your distributed archive? • What are the administrative/sponsor requirements that arise from each project (basically, metrics collection and reporting)? **Homework assignment** • Before the conference adjourns, convert all known science drivers to use cases.
11:00 a.m. – 11:10 a.m.	**Break**
11:10 a.m. – 12:00 p.m.	**Science Driver Town Hall Discussion** *Session Discussion Lead — Dean N. Williams* Town Hall Panel — Dave Bader, Sébastien Denvil, Peter Gleckler, Justin Hnilo, Tsengdar Lee, Jerry Potter, and Karl Taylor **Questions** • What is working, and what is not? • What are the key challenges that scientists encounter? • What data services would address the identified challenges? • What exists today? • What do we still need? • What are the key characteristics that these services need to have to be successful (e.g., integrated and easy to customize)? • What are the key impediments (for both data providers and service providers) in delivering these services? • Which services should be developed with the highest priority, and what would be their measurable impact on science?
12:00 noon –1:30 p.m.	**Lunch**

Time	Topic
1:30 p.m. –3:30 p.m.	**Required Data Center and Interoperable Services** *Session Discussion Lead — Michael Lautenschlager* **1:30 p.m. – 1:50 p.m.** Dean N. Williams — DOE/LLNL **1:55 p.m. – 2:15 p.m.** Ben Evans — ANU/NCI **2:20 p.m. – 2:40 p.m.** Stephan Kindermann — ENES/DDC/DKRZ **2:45 p.m. – 3:05 p.m.** Sébastien Denvil — ENES/IPSL **3:10 p.m. – 3:30 p.m.** Philip Kershaw — ENES/CEDA **Questions and discussion of example use case requirements for each major supporting data center** • What are the key things that are difficult to do today and are impeding scientific progress or productivity? • What is your timeline for data production and distribution? • What is the estimated size of your distributed archive? • What (or which) projects do you support? • What are the scaling challenges? For example, can we make our data access services (e.g., TDS elastic) scale out to meet demand? • What about provision of hosted processing, whether cloud services, batch computing, or other deployments alongside data center archives? • What issues surround workload and data mobility? How can new technologies such as containers enable us to port whole workloads and data between infrastructures? • How we can attach persistent identifiers and associated metadata to workloads and data, allowing them to be repeatable, referenced, and cited? **Homework assignment** • Before conference adjourns, convert all known data center drivers and interoperable services to use cases
3:30 p.m. – 3:45 p.m.	**Break**
3:45 p.m. – 4:45 p.m.	**Data Center and Interoperable Services Town Hall Discussion** *Session Discussion Lead — Michael Lautenschlager* Town Hall Panel — Ben Evans, Stephan Kindermann, Dean N. Williams, Sébastien Denvil, and Philip Kershaw • Data integration and advanced metadata capabilities • Data and metadata collection and sharing capabilities • Data quality, uncertainty quantification, and ancillary Information • Use of broader ontology for discovery and use of project data sets • Data discovery, access, and downloading, along with subsetting services and capabilities • Data preparation services and tools • Authentication and security • Local and remote publication services • Local and remote catalog and search services and data transfer services • Human–computer interface (e.g., user interfaces and APIs) • Resource discovery and allocation services • Workflow services (link together scientific or project execution) • Computing services • Exploration services (including analytics and visualization) • Identify and prioritize key gaps and benefitting communities
4:45 p.m.	**Adjourn Day 1**
6:00 p.m.	Awards ceremony and ice breaker at Marriott hotel — San Carlos 4 (Cost, $40)

Time	Topic
Wednesday, December 9, 2015	
8:00 a.m. – 8:30 a.m.	Meet-and-greet
8:30 a.m. – 10:05 a.m.	**Advanced Computational Environments and Data Analytics** *Session Discussion Lead — Robert Ferraro* **8:30 a.m. – 8:45 a.m.** Overview of the CWT and target milestones — Daniel Duffy, NASA/GSFC **8:50 a.m. – 9:05 a.m.** WPS overview and demo — Charles Doutriaux, DOE/LLNL **9:10 a.m. – 9:25 a.m.** Analytics-as-a-service framework — Thomas Maxwell, NASA/GSFC **9:30 a.m. – 9:45 a.m.** Ophedia — Sandro Fiore, ENES/CMCC **9:50 a.m. – 10:05 a.m.** WPS service and back-end — Maarten Plieger, ENES/KNMI **Questions** • What are the key challenges that scientists encounter? • What are the key things that are difficult to do today and are impeding scientific progress or productivity? • What capabilities would address the identified challenges? • What exists today? • What do we still need? • What impediments do resource providers and software developers face in providing these missing capabilities? • Which requirements need to be addressed with the highest priority, and what would be their measurable impact on science? • What are the overall integration issues? **Homework assignment** • Before the conference adjourns, convert all known data center drivers to use cases.
10:10 a.m. – 10:25 a.m.	**Break**
10:25 a.m. – 11:00 a.m.	**Computational Environments and Data Analytics Town Hall Discussion** *Session Discussion Lead — Robert Ferraro* Town Hall Panel — Daniel Duffy, Aashish Chaudhary, Charles Doutriaux, Thomas Maxwell, Sandro Fiore, and Maarten Plieger • Definition of a scalable compute resource (clusters and HPCs) for projects' data analysis • Data analytical and visualization capabilities and services • Analysis services when multiple data sets are not co-located • Performance of model execution • Advanced networks as easy-to-use community resources • Provenance and workflow • Automation of steps for the computational work environment • Resource management, installation, and customer support • Identification and prioritization of key gaps and benefitting communities
11:00 a.m. – 11:55 a.m.	**ESGF Development for Data Centers and Interoperable Services** *Session Discussion Lead — Luca Cinquini* ESGF working teams report on meeting project requirements, work achieved over the past year, prioritized development, roadmap, needed resources for meeting goals, and collaborations with other agencies. **11:00 a.m. – 11:10 a.m.** CoG User Interface Working Team — Sylvia Murphy, NOAA/ESRL **11:15 a.m. – 11:25 a.m.** Dashboard Working Team — Sandro Fiore, ENES/CMCC **11:30 a.m. – 11:40 a.m.** Data Transfer Working Team — Luckasz Lacinski, DOE/ANL **11:45 a.m. – 11:55 a.m.** Identity Entitlement Access Team — Philip Kershaw, ENES/BADC
12:00 noon – 1:30 p.m.	**Lunch**

Time	Topic
1:30 – 4:40 p.m.	**ESGF Development for Data Centers and Interoperable Services** *Session Discussion Lead — Luca Cinquini* ESGF working teams report on meeting project requirements, work achieved over the past year, prioritized development, roadmap, needed resources for meeting goals, and collaborations with other agencies. **1:30 p.m. – 1:40 p.m.** Installation Working Team — Prashanth Dwarakanath, ENES/Liu **1:45 p.m. – 1:55 p.m.** International Climate Network Working Group — Eli Dart, DOE/ESnet **2:00 p.m. – 2:10 p.m.** Metadata and Search Working Team — Luca Cinquini, NASA/JPL **2:15 p.m. – 2:25 p.m.** Node Manager Working Team — Sasha Ames, DOE/LLNL **2:30 p.m. – 2:40 p.m.** Persistent Identifier Services — Tobias Weigel, ENES/DKRZ **2:45 p.m. – 2:55 p.m.** Provenance Capture Working Team — Bibi Raju, DOE/PNNL **3:00 p.m. – 3:10 p.m.** Publication Working Team — Sasha Ames, DOE/LLNL **3:15 p.m. – 3:25 p.m.** Quality Control Working Team — Martina Stockhause, ENES/DKRZ **3:30 p.m. – 3:45 p.m.** Break **3:45 p.m. – 3:55 p.m.** Replication and Versioning Working Team — Stephan Kindermann, ENES/DKRZ **4:00 p.m. – 4:10 p.m.** Software Security Working Team — Prashanth Dwarakanath, ENES/Liu **4:15 p.m. – 4:25 p.m.** Support Working Team — Matthew Harris, DOE/LLNL **4:30 p.m. – 4:40 p.m.** User Working Team — Torsten Rathmann, ENES/DKRZ
4:40 p.m.	**Adjourn Day 2**
6:00 p.m.	Happy hour at Blue Fin Café Billiards (Cost, $20)
Thursday, December 10, 2015	
8:00 a.m. – 8:30 a.m.	Meet-and-greet
8:30 a.m. – 9:30 a.m.	**ESGF Development for Data Centers and Interoperable Services Town Hall Discussion** *Session Discussion Lead — Luca Cinquini* **Questions** • What tools have been identified during the previous discussions that should be made more widely accessible to the community? • Are these working team tools addressing community needs? • What other tools are there that could address key community needs? • How should tools and services be made available in the future for the integrated ESGF infrastructure? • What level of support would be expected from the science community? • How do we want to assess the maturity and capability (e.g., benchmarks or crowdsourcing) of the working team tools and services? • Are there any conventions needed for the working teams with respect to the many projects? • What level of service, monitoring, maintenance, and metrics is needed for each working team's data services and tools? • What do working teams want to see from others? • What do scientists want to have access to with regard to the working teams? • What standards and services need to be adopted within the compute environment that will allow projects to participate in the multiagency data initiatives discussed on the first day? • What is needed for data sharing across multi-international agencies?
9:30 a.m. – 12:00 noon	**Coordinated Efforts with Community Software Projects** *Session Discussion Lead — Sébastien Denvil* **9:30 a.m. – 9:55 a.m.** THREDDS Data Server (TDS) — John Caron, Sunya, Inc, USA **10:00 a.m. – 10:15 a.m.** Science DMZ for ESGF Supernodes — Eli Dart, DOE/ESnet **10:20 a.m. – 10:35 a.m.** Named Data Networking (NDN) — Christos Papadopoulos, Colorado State **10:40 a.m. – 10:55 a.m.** Climate Model Output Rewriter Version 3 (CMOR3) — Denis Nadeau, DOE/LLNL **11:00 a.m. – 11:15 a.m.** Synda (synchro-data) — Sébastien Denvil, ENES/IPSL **11:20 a.m. – 11:35 a.m.** Globus — Rachana Ananthakrishnan, DOE/ANL **11:40 a.m. – 11:55 a.m.** On-demand streaming of massive climate simulation ensembles — Cameron Christensen, University of Utah **Questions** • How will your efforts help the ESGF community of users? • What is the timeline for releasing your efforts? • What standards and services need to be adopted within the environment that will allow ESGF to participate in early adoption? • How are you funded for longevity?

Time	Topic
12:00 noon – 1:30 p.m.	**Lunch**
1:30 p.m. – 2:30 p.m.	**Community Software Projects Town Hall Discussion** *Session Discussion Lead — Sébastien Denvil* Town Hall Panel — John Caron, Eli Dart, Christos Papadopoulos, Denis Nadeau, Sébastien Denvil, Rachana Ananthakrishanan, and Cameron Christensen **Questions** • What standards and services need to be adopted within the environment that will allow projects to participate in multiagency data initiatives? • How should these tools and services be made available in ESGF's future in an integrated way?
2:30 p.m. – 3:00 p.m.	**Poster Session** *Session Discussion Lead — Dean N. Williams* **Posters** **1.** Climate4Impact Portal — Maarten Plieger, KNMI **2.** ACME Workflow — Matthew Harris, DOE/LLNL **3.** HPSS Connections to ESGF — Sam Fries, DOE/LLNL **4.** Distributed Resource for the ESGF Advanced Management (DREAM) — Dean N. Williams, DOE/LLNL **5.** Observation Data Publication into the ESGF — Misha B. Krassovski, DOE/ORNL **6.** Climate Data Management System, version 3 (CDMS3) — Denis Nadeau, DOE/LLNL **7.** Ultrascale Visualization Climate Data Analysis Tools (UV-CDAT) — Aashish Chaudhary, Kitware **8.** CDATWeb — Matthew Harris, DOE/LLNL **9.** NetCDF/HDF5 — Ben Evans, NCI/ANU **10.** PROV — Ben Evans, NCI/ANU **11.** Climate Forecast (CF) Convention — Karl Taylor, DOE/LLNL **12.** ES-DOC — Mark Greenslade, ENES/IPSL **13.** Agreement on Data Management and Publication Workflow — Guillaume Levavasseur, ENES/IPSL **14.** Data Citation Service — Martina Stockhause, ENES/DKRZ **15.** PCMDI's Metrics Package — Paul Durack, DOE/LLNL **16.** DOE UVCMetrics — Jeff Painter, DOE/LLNL; Brian Smith, DOE/ORNL **17.** ESMValTool — Stephan Kindermann, ENES/DKRZ **18.** CMIP6 Errata as a New ESGF Service — Guillaume Levavasseur, ENES/IPSL **19.** Enabling *in situ* Analytics in the Community Earth System Model via a Functional Partitioning Framework — Valentine Anantharaj, DOE/ORNL **20.** The OPTIRAD Project: Cloud-Hosting the IPython Notebook to Provide a Collaborative Data Analysis Environment for the Earth Sciences Community — Philip Kershaw, ENES/CEDA **21.** A NASA Climate Model Data Services (CDS) End-to-End System to Support Reanalysis Intercomparison — Jerry Potter, NASA/GSFC **Questions** • How will your efforts help the ESGF community of users? • What is your timeline for releasing your efforts? • What standards and services need to be adopted within the environment that will allow ESGF to participate in early adoption? • How should these tools and services be made available in ESGF's future in an integrated way? • How are you funded for longevity (i.e., funding source)?
3:00 p.m. – 5:00 p.m.	**Team and Across-Team Discussions**
5:00 p.m.	**Adjourn Day 3**

Time	Topic
Friday, December 11, 2015	
8:00 a.m. – 8:30 a.m.	Meet-and-greet
8:30 a.m. – 10:00 a.m.	**ESGF Development Teams Report Back on Conference Findings** *Session Discussion Lead — Dean N. Williams* • Poster session feedback • Open discussion
10:00 a.m. – 10:15 a.m.	**Break**
10:15 a.m. – 12:00 noon	**ESGF XC and WIP Breakout Meeting** • Discuss development of the annual report • Discuss meeting location and time of the next ESGF F2F meeting **Working Teams Meeting** • All working teams discuss conference findings for their area of the annual report
12:00 noon – 1:30 p.m.	**Lunch**
1:30 p.m. – 5:00 p.m.	**General Data Code Sprint** *Session Discussion Lead — Working Team Leads*
5:00 p.m.	**Adjourn Day 4**
Conclusion of the 5th Annual ESGF F2F Conference	

Appendix B. Presentation and Poster Abstracts

Day 1: Tuesday, December 8, 2015 Science Drivers Project Requirements and Feedback	
Title and Presenter	**Abstract**
WGCM Infrastructure Panel ***Karl Taylor (DOE/LLNL)*** *taylor13@llnl.gov* ***V. Balaji (NOAA/GFDL)*** *balaji@princeton.edu*	The WGCM Infrastructure Panel was formed in response to the WGCM's (2013) expressed need to provide scientific guidance and requirements for the gloal data infrastructure underpinning global climate science and modeling. This infrastructure includes ESGF software and other tools such as ES-DOC, CoG, CMOR, CF Conventions, and others. Chaired by V. Balaji (Princeton/GFDL) and K. Taylor (DOE/LLNL/PCMDI), the panel outlined in 2014 a strategy to develop a series of "position papers" on global data infrastructure and its interaction with the scientific design of experiments. The papers would then be presented to the WGCM annual meeting for endorsement by the WGCM, the CMIP Panel, and the modeling groups. A series of position papers were unveiled at the WGCM-19 meeting (2015) in Dubrovnik, Croatia. The 11 position papers currently in draft, and others in progress, will be available on the WIP website (www.earthsystemcog.org/projects/wip/).
DOE Accelerated Climate Modeling for Energy ***David Bader (DOE/LLNL)*** *bader2@llnl.gov* ***Dean N. Williams (DOE/LLNL)*** *Williams13@llnl.gov* ***Valentine Anantharaj (DOE/ORNL)*** \ *anantharajvg@ornl.gov*	Sponsored by the U.S. DOE Office of Biological and Environmental Research, the Accelerated Climate Modeling for Energy (ACME) project is an ongoing, state-of-the-science Earth system modeling, simulation, and prediction project that optimizes the use of DOE laboratory resources to meet the nation's science needs and DOE missions—"A DOE Model on DOE Machines for the DOE Mission." ACME's initial scientific goals address three areas of importance to both climate research and society: (1) the water cycle: How do the hydrological cycle and water resources interact with the climate system on local to global scales?, (2) biogeochemistry: How do biogeochemical cycles interact with global climate change?, and (3) the cryosphere–ocean system: How do rapid changes in cryosphere-ocean systems interact with the climate system? The high-resolution version of the ACME model simulates the fully coupled climate system at 15 to 25 km, and further development is needed to optimize performance on current and future DOE leadership-class computers. New scalable and extensible solutions for data archival, search, retrieval, and analysis are needed for the size and complexity of the ACME output.
Obs4MIPs ***Peter Gleckler (DOE/LLNL)*** *gleckler1@llnl.gov*	Observations for Model Intercomparisons (Obs4MIPs) is an activity to make observational products more accessible for climate model intercomparisons. It is limited to a collection of well-established and documented data sets that have been organized according to the **CMIP5** (www-pcmdi.llnl.gov/ipcc/data_status_tables.htm) model output requirements and made available on ESGF (HYPERLINK DIDNT WORK) via CoG. Efforts are under way to further align Obs4MIPs with the needs of CMIP6. In brief, satellite products currently available via Obs4MIPs are: (1) directly comparable to a model output field defined as part of CMIP5; (2) open to contributions from all data producers that meet the **Obs4MIPs requirements** (www.earthsystemcog.org); (3) well documented, with traceability to track product version changes; and (4) served through ESGF (and directly available through this CoG. The technical alignment of observational products with model output greatly facilitates model data comparisons. Obs4MIPs was initiated with support from NASA and DOE and is now a WCRP activity overseen by the WCRP Data Advisory Council (WDAC). Additional satellite products are expected from NASA, NOAA, ESA, EUMETSAT, and community efforts such as **CFMIP-OBS** (climserv.ipsl.polytechnique.fr). A WDAC Task Team helps advance Obs4MIPs, and community interest in having the project expanded beyond satellite data is high. This presentation will describe opportunities and challenges for Obs4MIPs, with particular emphasis on the technical needs required to advance the project.
IS-ENES ***Sébastien Denvil (ENES/IPSL)*** *sebastien.denvil@*	European climate modeling groups joined together in 2001 to create ENES with the objectives of helping the development and evaluation of climate models of the Earth system, encouraging the exchange of software and model results, and promoting the development of HPC facilities. Funded by the European Union, the IS-ENES project (Infrastructure for ENES; first phase 2009–13, second phase 2013–17) aims to promote the development of a common distributed climate modeling research infrastructure in Europe in order to facilitate the development and exploitation of climate models and better fulfill societal needs with regards to climate change issues (is.enes.org). IS-ENES supports the integration of the European climate modeling community and recently issued the *Infrastructure Strategy for the European Earth System Modeling Community: 2012–2022*. It promotes the dissemination of European climate model results from the international WCRP CMIP5 and CORDEX experiments developed in preparation for the IPCC 5th Assessment Report by supporting ESGF developments and operations. IS-ENES also aims at enhancing model development and software sharing and supports the preparation of high-end simulations and the use of HPC.

Table continued next page

Day 1: Tuesday, December 8, 2015 Science Drivers Project Requirements and Feedback	
Title and Presenter	**Abstract**
CREATE-IP ***Jerry Potter (NASA/GSFC)*** *jpotter@ucdavis.edu*	The Climate Model Data Services (CDS) group at NASA's Goddard Space Flight Center is collaborating with the world's major reanalysis projects to collect reanalysis data and present it through the Distribution, Visualization, Analytics, and Knowledge Services, resulting in the Collaborative REAnalysis Technical Environment Intercomparison Project (CREATE-IP). CDS has converted monthly mean data from the five major reanalysis projects, including MERRA-2, to the standard ESGF format of one variable per file and published the data in ESGF. The agreement or disagreement among reanalyses enables us to judge the scientific validity of using reanalysis data to evaluate climate models. Differences in the reanalyses may have a variety of causes including, but not limited to, differences in the input observations, changes in observation instrumentation, or differences in the models' physical parameterizations. Data are prepared to the CMIP5 and Obs4MIPs specifications using CMOR and are distributed through the ESGF CoG interface. So far, we have prepared monthly averages of the primary variables produced by the CMIP5 simulations and six-hour frequency of an initial selected set of variables. Additional preparation of six-hour variables is in process. These initial variables were selected because they are useful in evaluating weather events in the past for intercomparison among the different reanalyses. Each reanalysis center produces a different data structure and organization, posing difficulties in preparing the reanalyses for inclusion into ESGF. As a result, each data set requires custom processing. Reanalyses are produced at high horizontal and vertical resolution, and the six-hour data conversion processing has proven to be particularly challenging, requiring several days of dedicated uninterrupted computing to complete one variable. To assist with testing of the processed data, UV-CDAT has proven to be particularly useful for data quality control because of its inherent ease of use and flexibility. In addition to distributing reanalysis data through ESGF, we have implemented a visualization tool, CREATE-V, based on code from the National Center for Atmospheric Research Climate Inspector. This tool utilizes TDS and OpenLayers to support access by interdisciplinary and reanalysis scientists for the exploration and side-by-side comparison of variables by reanalysis, date, and level.

Day 1, December 8, 2015 Required Data Center and Interoperable Services	
Title and Presenter	**Abstract**
ANU/NCI ***Ben Evan (ANU/NCI)*** *Ben.Evans@anu.edu.au*	NCI currently supports more than 10 years of research data collections of Earth systems models and observational products on a high-performance data node with a co-located HPC facility and cloud environment. This integrated large-scale computing infrastructure offers opportunities for enabling scientific users to analyze petabytes of data. NCI's data collections are managed and made available through its National Environmental Research Data Interoperability Platform, which allows access by NCI's Raijin supercomputer, its HPC cloud environment that supports data analysis through a Virtual Desktop Interface with a large catalogue of analysis tools (e.g., UV-CDAT), virtual laboratories such as the Climate and Weather Science laboratory, and for high-speed data transfer such as that used for ICNWG. Provenance of the data is maintained through data product information via the Provenance Management System using the W3C Prov standard. NCI has also been analyzing its data collections in Earth system model, observation, and point-cloud data to provide high-performance data access that addresses the future of HPC computing models and simulations, data analysis tools and engines, and data services. The data formats have been considered in the broader context of data interoperability between science domains. In doing so, we have compared the performance of various I/O approaches (POSIX, MPI-IO, NetCDF3, NetCDF4, GeoTiff, HDF5); the interfaces for accessing data (native libraries, GDAL, Python, OPeNDAP, OGC); and techniques for data subsetting, aggregation, and coordinate transformation.
DDC/DKRZ ***Stephan Kindermann (ENES/DKRZ)*** *kindermann@dkrz.de*	The German Climate Computing Centre (DKRZ) supports the complete data lifecycle of climate data products—model data generation; postprocessing; data ingestion; quality assurance; ESGF publication; long-term archival and assignment of PIDs, DOIs, and early citation information. To support end users, DKRZ acts also as a replication and long-term archival center of external ESGF data products, hosting the world data center for climate (WDCC) and acting as an IPCC data distribution center. Recently, the end-user requirements were strengthened to provide a platform to analyze the huge ESGF data volume hosted at DKRZ. The compute platform is part of the DKRZ high-performance computing solution. Data products derived from the compute platform are stored on the HPC, ~50 PB Luster file system and made available through ESGF. In addition, DKRZ established a data cloud service to support end users in hosting project data collections and to support DKRZ data import and export. A virtual machine environment allows for the flexible deployment of project-specific data servers and services. First investigations to provide docker-based compute services in the future also were performed.
DOE LLNL/PCMDI ***Dean N. Williams (DOE/LLNL)*** *williams13@llnl.gov*	LLNL researchers benefit from an institutional IT infrastructure that provides desktop support and experts in server technologies. The latter includes virtualization expertise that has been applied to provide multiple operating systems on shared resources and to create a wide variety of virtual machines to leverage resources across LLNL. An enterprise team also provides a networking service to implement the Science demilitarized zone (DMZ) and data transfer nodes (DTNs). Connections into LLNL include ESnet, the dynamic Science Data Network, the ALICE grid system, the Open Science Grid, and a wide variety of programmatic networks. Cisco telepresence nodes are also available to facilitate remote collaboration over these networks. In addition to local group resources, LLNL computational scientists deliver a balanced HPC environment with constantly evolving hardware resources and a wealth of HPC expertise in porting; running; and tuning real-world, large-scale applications and data management systems such as ESGF. Currently, LLNL delivers multiple petaflops of compute power, massive shared parallel file systems, powerful data analysis/cluster platforms, and archival storage capable of storing many petabytes of data. This balanced hardware environment supports key collaborations between LLNL application developers and community experts on the creation, debugging, production use, and performance monitoring of large-scale parallel applications, as well as data analysis in a wide variety of scientific climate applications, such as CMIP and ACME. All members of the LLNL climate projects (e.g., PCMDI, ACME) have desktop workstations available to them. In addition, the LLNL Climate Science Program maintains 12 shared computer servers and two internal file servers with more than 250 TB of aggregate storage to support its internal research activities. Data management software (i.e., ESGF) and analysis software (i.e., UV-CDAT) are maintained on these shared systems. The file servers enable seamless integration among the workstation, computer server, and data resources using NFS remote mounting capabilities. The Green Data Oasis and Climate Central Systems host data (e.g., hosting the CMIP3 and CMIP5 archives) served to the external community. Additional archival capacity is required for CMIP6.

Table continued next page

Day 1, December 8, 2015 Required Data Center and Interoperable Services	
Title and Presenter	**Abstract**
IS-ENES/IPSL ***Sébastien Denvil (ENES/IPSL)*** *sebastien.denvil@ipsl.jussieu.fr*	The Institute Pierre Simon Laplace (IPSL) climate modeling group gathers climate modeling teams from the Centre Nationale de la Recherche Scientifique, the Commissariat à l'Énergie Atomique et aux Énergies Alternatives, and from university research disciplines such as meteorology, oceanography, and biogeochemistry. The group's objective is the study of natural and anthropogenic variability in the global climate system. IPSL is also studying climate change impacts and usage of climate projections for adaptation to climate change related to industry. IPSL is one of the climate modeling centers of international repute contributing to the IPCC. IPSL draws upon a team of 50 engineers and informatics experts. Their collaborations within France and internationally, the diversity of the technologies they exploit, and the size and variety of the projects that they handle are a reflection of IPSL's desire to be at the cutting edge of climate modeling. During this talk, we will present our strategies and propositions on required data center and interoperable services.
IS-ENES/CEDA ***Philip Kershaw (ENES/CEDA)*** *philip.kershaw@stfc.ac.uk*	The Centre for Environmental Data Analysis (CEDA) hosts data centers managing climate and Earth observational data on behalf of the UK environmental science community to facilitate access and support its work in collaborations with international partners. CEDA is underpinned by JASMIN, a petascale storage and cloud computing facility. Besides hosting the CEDA data archive, JASMIN provides communities of users with a collaborative environment for analysis of data including group workspaces and hosted processing capability. There are a number of challenges moving forward, driven in part by the success of JASMIN and by the increasing data volumes for both model data and observations. Technically, these can be summarized in terms of the ability to scale computing resources and the effective integration of new and existing technologies to provide services needed by the user community. Currently, the archive holds about 3 PB maintained on spinning disk with a full tape backup. This collection includes more than 250 data sets and in excess of 200 million files. Two key programs in particular, CMIP6 and the data stream from the new generation of European Space Agency Earth observation satellites (the Sentinels), present challenges in terms of both the data volumes (~10 PB reserved for CEDA CMIP6 archive) and velocity (expected 10 TB/day rate for Sentinel data sets). Data from these sources will exceed the disk capacity available to the archive in the near future and will necessitate the development of an integrated solution for disk and near tape storage. Work is under way to pilot new technologies associated with the cloud service, including the use of containers, orchestration tools, and object stores. These will enable the scaling of resources to meet demand and to federate with other cloud providers, be they public or from the research community. For ESGF, there is a need for a robust and stable core service for the projects we host through the infrastructure. These include SPECS, CCMI, CLIPC, and the ESA Climate Change Initiative (CCI). For CCI, CEDA has started a project for ESA over the past year to build an Open Data Portal to serve data products from the program. This effort is reusing and building upon technology from ESGF, including the index and data nodes, and will include innovations such as support for ISO19115 search services and the use of semantic web technology to develop a machine-readable DRS vocabulary and govern its use with client applications.

Day 2: Wednesday, December 9, 2015 Advanced Computational Environments and Data Analytics	
Title and Presenter	**Abstract**
Computing Working Team Overview ***Daniel Duffy (NASA/GSFC)*** *daniel.q.duffy@nasa.gov* ***Charles Doutriaux (DOE/LLNL)*** *doutriaux1@llnl.gov*	A major paradigm shift is occurring as users move from downloading data to perform data analytics to moving analysis routines to the data and performing these computations on distributed platforms. In preparation for this shift, the ESGF Computing Working Team (CWT) is engaged in developing the capability to enable data-proximal analytics throughout ESGF. To guide the discussion, the team created several potential analytical use cases for the data stored in ESGF, using anomalies as the prototypical example of the type of analytics that scientists would want to perform. Next, the team focused on tailoring the interface to the analytic capabilities and reviewed several different potential programming interfaces, before deciding on the web processing service (WPS) as the standard for defining the analytical services, inputs, and outputs. In addition to being rather simple to deploy, the geospatial community makes heavy use of WPS-enabled services. A specification document has been written along with an initial implementation using the pyWPS. This presentation will provide an update on the work performed over the last year and on the CWT's plans for the upcoming year.
WPS Overview and Demo ***Charles Doutriaux (DOE/LLNL)*** *doutriaux1@llnl.gov* ***Daniel Duffy (NASA/GSFC)*** *daniel.q.duffy@nasa.gov*	As the size of remote sensing observations and model output data grows, the volume of the data has become overwhelming, even to many scientific experts. As societies are forced to better understand, mitigate, and adapt to climate changes, the combination of Earth observation data and global climate model projects is crucial not only to scientists but to policymakers, downstream applications, and even the public. Scientific progress on understanding climate change is critically dependent on the availability of a reliable infrastructure that promotes data access, management, and provenance. ESGF has created such an environment for the IPCC. ESGF provides a federated global cyberinfrastructure for data access and management of model outputs generated for IPCC assessment reports. The current generation of the ESGF federated grid allows data consumers to find and download data with limited capabilities for server-side processing. Since the amount of data for future assessment reports is expected to grow dramatically, ESGF is working on integrating server-side analytics throughout the federation. The ESGF CWT has created a WPS API to enable access to scalable computational resources. The API is the exposure point to high-performance computing resources across the federation. Specifically, the API allows users to execute simple operations on ESGF data, such as maximum, minimum, average, and anomalies, without having to download the data. These operations are executed at the ESGF data node site, with access to large amounts of parallel computing capabilities. This presentation will highlight the WPS API and its capabilities, provide implementation details and a demonstration, and discuss future developments.
The Climate Data Analytic Services Framework ***Thomas Maxwell (NASA/GSFC)*** *thomas.maxwell@nasa.gov* ***Mark McInerney (NASA/GSFC)*** *mark.mcinerney@nasa.gov* ***Daniel Duffy (NASA/GSFC)*** *daniel.q.duffy@nasa.gov* ***Jerry Potter (NASA/GSFC)*** *jpotter@ucdavis.edu* ***Charles Doutriaux (DOE/LLNL)*** *doutriaux1@llnl.gov*	Faced with unprecedented growth in the "big data" domain of climate science, NASA has developed the Climate Data Analytic Services (CDAS) framework. This framework enables scientists to execute trusted and tested analysis operations in a high-performance environment close to the massive data stores at NASA. The data are accessed in standard (e.g., NetCDF and HDF) formats in a POSIX file system and are processed using trusted climate data analysis tools (e.g., ESMF, CDAT, and NCO). The framework is structured as a set of interacting modules, allowing maximal flexibility in deployment choices. CDAS services are accessed via a WPS API being developed in collaboration with the ESGF Computing Working Team to support server-side analytics for ESGF. The API can be executed using either direct web service calls, a Python script or application, or a JavaScript-based web application. Client packages in Python or JavaScript contain everything needed to make CDAS requests. The CDAS architecture brings together the tools, data storage, and high-performance computing required for timely analysis of large-scale data sets where the data reside, to ultimately produce societal benefits. It is currently deployed at NASA in support of the CREATE project, which centralizes numerous global reanalysis data sets onto a single advanced data analytics platform. This service permits decision-makers to investigate climate changes around the globe, inspect model trends, and compare multiple reanalysis data sets.

Table continued next page

Day 2: Wednesday, December 9, 2015 Advanced Computational Environments and Data Analytics	
Title and Presenter	**Abstract**
Ophedia ***Sandro Fiore (ENES/CMCC)*** *sandro.fiore@unisalento.it*	The Ophidia project is a research effort on big data analytics facing scientific data analysis challenges in multiple domains (e.g., climate change). Ophidia provides declarative, server-side, and parallel data analysis, jointly with an internal storage model able to efficiently deal with multidimensional data and a hierarchical data organization to manage large data volumes ("data cubes"). The project relies on a strong background in high-performance database management and OLAP systems to manage large scientific data sets. The Ophidia analytics platform provides several data operators to manipulate data cubes and array-based primitives to perform data analysis on large scientific data arrays (e.g., statistical analysis, FFT, DWT, subsetting, and compression). Metadata management support (CRUD-like operators) is also provided. The server front end exposes several interfaces to address interoperability requirements: WS-I^{+}, GSI/VOMS, and OGC-WPS (through PyWPS). From a programmatic point of view, a Python module (PyOphidia) makes straightforward the integration of Ophidia into Python-based environments and applications (e.g., IPython). The system offers a command-line interface (e.g., bash-like) for end users, with a complete set of commands as well as integrated help and manuals. A key point of the talk will be the workflow capabilities offered by Ophidia. In this regard, the framework stack includes an internal workflow management system that coordinates, orchestrates, and optimizes the execution of multiple scientific data analytics and visualization tasks (e.g., statistical analysis, metadata management, virtual file system tasks, maps generation, and import/export of data sets in NetCDF format). Specific macros are also available to implement loops or to parallelize them in case of data independence. Real-time workflow monitoring execution is also supported through a graphical user interface. Some real workflows implemented at CMCC and related to different projects also will be presented, including (1) climate indicators in the FP7 EU Climate Information Platform for Copernicus and EUBRAZIL Cloud Connect projects; (2) fire danger prevention analysis in the INTERREG OFIDIA project; and (3) large-scale climate model intercomparison and data analysis (e.g., analysis of precipitation trends, climate change signals, and anomalies) in the H2020 INDIGO-DataCloud.
WPS Service and Back-End Application of ESGF and WPS Services in climate4impact ***Maarten Plieger (ENES/KNMI)*** *maarten.plieger@knmi.nl*	The aim of climate4impact is to enhance the use of climate research data and the interaction with climate effect and impact communities. The portal is based on impact use cases from different European countries and is evaluated by a user panel consisting of use-case owners. This work has resulted in the ENES portal interface for climate impact communities and can be visited at climate4impact.eu. This presentation discusses how climate4impact uses web processing services (WPS) to perform analysis on climate data stored in ESGF. The WPS calculates climate indices and subsets data using OpenClimateGIS/icclim on data stored in ESGF data nodes. Data are then transmitted from ESGF nodes over secured OpenDAP and become available in a new per-user secured OPenDAP server. Results can then be visualized using ADAGUC WMS. Dedicated wizards for processing of climate indices will be developed in close collaboration with users. The portal is able to generate graphical user interfaces on WPS endpoints and aims to use the WPS being developed by the ESGF CWT. In this presentation, the climate4impact architecture will be detailed, along with the following items: • **Processing:** Transform, subset, and export data into other formats and perform climate indices calculations using WPS implemented by PyWPS, based on NCAR NCPP OpenClimateGIS and IS-ENES2 icclim. • **Visualization:** Visualize data from ESGF data nodes using ADAGUC web map services. • **Security:** Log in using OpenID for access to ESGF data nodes. ESGF works in conjunction with several external websites and systems. The climate4impact portal uses X509-based, short-lived credentials generated on behalf of users with a MyProxy service. Single sign-on is used to make these websites and systems work together.

Day 2: Wednesday, December 9, 2015 ESGF Development for Data Centers and Interoperable Services	
Title and Presenter	**Abstract**
CoG User Interface Working Team ***Sylvia Murphy (NOAA/ERSL)*** *sylvia.murphy@noaa.gov*	Throughout 2015, the ESGF User Interface Working Team (UIWT) has worked on upgrading and expanding the Earth System CoG Collaboration Environment to replace the old ESGF web front end. Major new features include: (1) integration of CoG into the ESGF software stack; (2) federation of distributed CoGs; (3) support for downloading data via Globus; and (4) general site improvements, infrastructure upgrades, and security fixes. CoG is now ready to be deployed as the ESGF front end at each node. For the next 6 months, the priority of the ESGF UIWT will be supporting ESGF administrators and end users, while collecting and prioritizing requirements for additional needed functionality.
Dashboard Working Team ***Sandro Fiore (ENES/CMCC)*** *sandro.fiore@unisalento.it* ***Paola Nassisi (ENES/CMCC)*** *paola.nassisi@cmcc.it* ***Giovanni Aloisio (ENES/CMCC)*** *giovanni.aloisio@unisalento.it*	Monitoring the Earth System Grid Federation is challenging. From an infrastructural standpoint, the dashboard and desktop components provide the proper environment for capturing usage metrics, as well as system status information at local (node) and global (institution and federation) level. The dashboard and the desktop are strongly coupled and integrated into the ESGF stack and represent the back and front end of the ESGF monitoring system. The dashboard acts as an information provider, collecting and storing a high volume of heterogeneous metrics on machine performance, network topology, host/service mapping, registered users and download statistics. The desktop is a web-based environment and provides effective, transparent, robust, and easy access to all the metrics and statistics provided by the dashboard. It is written in Java and JavaScript programming languages and presents enhanced views with several gadgets (enriched with charts, tables, and maps) for a simple and user-friendly visualization of aggregated and geolocalized information. All metrics collected by the ESGF monitoring infrastructure are stored in a system catalog that has been extended to support multiple types of information about the data usage statistics. More specifically, in addition to information such as the number of downloads, downloaded data sets, users who have downloaded some data, and the amount of data downloaded, new metrics are being provided. Some examples are: statistics about data downloads grouped by model, variable, experiment, country or over time; top 10 lists of the most downloaded data sets; and clients' distribution maps. To this end, specific data marts have been created to allow fast access to this information. Finally, to grant programmatic access to the metrics managed by the dashboard, a set of RESTful APIs has been defined (based on a JSON data interchange format), allowing users to design and implement their own client applications.
Data Transfer Working Team ***Luckasz Lacinski (DOE/ANL)*** *lukasz@uchicago.edu*	During the past year, the Data Transfer Working Team's focus has been on updating to the latest software and using the new user interface. The key work completed includes (1) supporting Globus download options with the latest ESGF user interface CoG; (2) updating components on the data node to a recent and supported version of servers; and (3) simplifying the installation process and script. The new release includes all these updates, and this talk will present details on this work and the impact for ESGF users and administrators.
Identity Entitlement Access Team ***Philip Kershaw (ENES/BADC)*** *philip.kershaw@stfc.ac.uk* ***Rachana Ananthakrishnan (DOE/ANL)*** *ranantha@uchicago.edu*	The remit of the Identity Entitlement Access (IdEA) team is to maintain and develop ESGF's system for access control to resources hosted within the federation. Over the past few years, the scope of this work has expanded to other projects beyond the original requirement of securing of access to CMIP5 data. In addition, with the development of compute capability for ESGF, securing access to computing resources is an increasingly important aspect for consideration in the evolution of the system. Activities over the last year have been dominated by the security incident with the federation. Nevertheless, some promising work has been undertaken in piloting a new capability for user delegation using the OAuth 2.0 protocol. User delegation is an important capability to support remote computation of secured resources. Initial integration work has begun between CEDA's OAuth 2.0 service and IS-ENES partners KNMI and DKRZ. This effort will soon be extended to work with Globus transfer. These activities are providing confirmation of the potential for OAuth to simplify access and to provide a common baseline to a number of different access control use cases. We outline the roadmap and resources needed to take this work into a production service for the federation. Besides the technical development of the system, there are considerations with respect to policy and operation of IdPs in the federation. We will set out recommendations for the future to simplify access for users, enhance security, and reduce the operational burden.

Table continued next page

Day 2: Wednesday, December 9, 2015 ESGF Development for Data Centers and Interoperable Services	
Title and Presenter	**Abstract**
Installation Working Team ***Prashanth Dwarakanath (ENES/Liu)*** *pchengi@nsc.liu.se* ***Nicolas Carenton (ENES/IPSL)*** *ncarenton@ipsl.jussieu.fr*	ESGF's Installation Working Team was created in March 2014. Its main responsibilities are ESGF release management, installation tool maintenance, and node administrator support. In 2015, many key deliverables were met, including providing an automated installation mechanism for ESGF, switching to Apache web server as the front end, and providing support for non-Java server-side components. The security incident in June 2015 was the biggest challenge encountered by the IWT. It was successfully handled but meant additional coordination with developers and extra develop-test-deploy cycles. We will present here the work done since the last ESGF Face to Face Conference, highlight features of the major releases to date, and discuss upcoming work on the installer.
International Climate Network Working Group ***Eli Dart (DOE/ESnet)*** *dart@es.net* ***Mary Hester (DOE/ESnet)*** *mchester@es.net*	The ICNWG is a working group of the ESGF focused on end-to-end data transfer performance. Our initial efforts have been focused on enabling the replication of large-scale data sets between the major climate data centers at BADC, DKRZ, LLNL, and NCI. This talk will describe the ICNWG work in 2015 and progress made to date. Next steps for the working group will also be discussed.
Metadata and Search Working Team ***Luca Cinquini (NASA/JPL)*** *Luca.Cinquini@jpl.nasa.gov*	For the ESGF Metadata and Search Working Team (MSWT), 2015 was largely dominated by the general ESGF security incident, which prompted the whole federation to be brought offline. The MSWT took advantage of this unfortunate situation to execute a much-needed upgrade of the ESGF search services infrastructure, which would have been much more difficult as a backward-compatible upgrade. Consequently, the upcoming ESGF 2.0 software stack will utilize Solr 5, deployed as a stand-alone engine embedded within Jetty, which includes many important new features such as automatic updates. The general master/slave/replica architecture has not changed, but the Solr slave shard will be exposed through the standard http port 80 to avoid pesky firewall issues. Additionally, support for publishing data to a new "local shard" has been introduced. From the user interface perspective, many improvements have been added to the search pages, the administrator configuration utilities, and the data cart. In the next year, the main focus of the ESGF-MSWT will be to support the upcoming CMIP6 distributed data archive and related observational data. Major areas of development will include metadata validation, partition of the global search space into virtual organizations, scalability, and performance.
Node Manager Working Team ***Sasha Ames (DOE/LLNL)*** *ames4@llnl.gov* ***Prashanth Dwarakanath (ENES/Liu)*** *pchengi@nsc.liu.se* ***Sandro Fiore (ENES/CMCC)*** *sandro.fiore@unisalento.it*	All ESGF nodes require a node manager component to coordinate automated configuration and federation-wide monitoring activities. To improve scalability over the prior P2P-based node manager, we are implementing a two-tier system that combines aspects of P2P to coordinate the "super-nodes" with the client server to handle the secondary tier of member nodes. We are transitioning to use a Python-based implementation that can run under Apache. Development of this component is approaching readiness to test in a test-federation environment as we shore up more of the functionality. Additionally, this talk will incorporate plans for a "tracking and feedback" effort for ESGF.
Persistent Identifier Services ***Tobias Weigel (ENES/DKRZ)*** *weigel@dkrz.de* ***Stephan Kindermann (ENES/DKRZ)*** *kindermann@dkrz.de* ***Katharina Berger (ENES/DKRZ)*** *berger@dkrz.de*	PID services for ESGF are concerned with the automated assignment and curation of persistent identifiers for CMIP6 data managed in ESGF at several levels of granularity. PIDs will be assigned to all CMIP6 files as well as several higher levels of aggregation covering data sets, simulations, and models. Identifier names are generated by CMOR and registered as part of the overall publishing workflow. An exemplary application based on the PID service is a smart user workspace tool that can pull additional information on given files from the federation, determine whether a new data set version is available, and ultimately provide access to it. The presentation will give an overview on the service design as described in the corresponding WIP paper and provide an update on the current development status. The service architecture is based on a distributed message queue to achieve high availability and throughput. The PID services interact with other ESGF components, including versioning, replication, and citation services.

Table continued next page

Day 2: Wednesday, December 9, 2015 ESGF Development for Data Centers and Interoperable Services	
Title and Presenter	**Abstract**
Provenance Working Team ***Bibi Raju (DOE/PNNL)*** *bibi.raju@pnnl.gov*	The Provenance Working Team aims to focus on developing provenance solutions in support of reproducibility and performance investigations to accomplish ACME's computational goals. This effort includes development of a provenance format that can capture sufficient information to enable scientists to reproduce their previous calculations correctly, as well as capture and link to performance information for specific workflows and model runs to enable in-depth performance analysis. The first step is to investigate methods for the capture, representation and storage, evaluation, access, and use of provenance information. During the last year, we have been developing a comprehensive workflow performance data model called Open Provenance Model-based Workflow Performance Provenance. It enables the structured analysis of workflow performance characteristics and variability. It also links provenance information and performance metrics ontology. The provenance capture ontology and system enable the capture of provenance information from the high-level workflow through all relevant system levels in one integrated environment. A provenance production and collection framework called Provenance Environment (ProvEn) is in place. It provides components supporting the production and collection of provenance information for distributed application environments. Semantic web technologies and ontologies, including the Open Provenance Model-based Workflow Performance Provenance ontology, are used by ProvEn for the representation, storage, and reporting of provenance. We are currently developing a provenance capture mechanism that can handle the high-velocity provenance information.
Publication Working Team ***Sasha Ames (DOE/LLNL)*** *ames4@llnl.gov* ***Rachana Ananthakrishnan (DOE/ANL)*** *ranantha@uchicago.edu*	The Publication Working Team is responsible for the ESGF publisher software, the development of a publications service, and the management of overarching workflows for ESGF publication, including all the required preparation steps. Accomplishments in 2015 included an initial release of a graphical user interface–based publication service running for ACME. ESGF 2.0 contains a handful of changes to the publisher software including support for upgraded components, improved versioning, and optional facets for published data sets. Future work will have a strong focus on workflow and software requirements for CMIP6 and, in addition, the release of a publication service API that incorporates Globus transfer of data.
Quality Control Working Team ***Martina Stockhause (ENES/DKRZ)*** *stockhause@dkrz.de* ***Guillaume Levavasseur (ENES/IPSL)*** *glipsl@ipsl.jussieu.fr* ***Katharina Berger (ENES/DKRZ)*** *berger@dkrz.de*	The ESGF Quality Control Working Team aims to improve the quality of ESGF user services by integration of additional external documentation. The team coordinates the implementation of the errata service (IPSL) and the data citation service (DKRZ). We will present the team's progress over the last 12 months and give a roadmap for the next year with special emphasis on requirements, collaboration, and risk aspects.
Replication and Versioning Working Team ***Stephan Kindermann (ENES/DKRZ)*** *kindermann@dkrz.de* ***Tobias Weigel (ENES/DKRZ)*** *weigel@dkrz.de*	Ensuring ESGF CMIP6 data consistency across sites strongly depends on stable and agreed versioning and replication procedures. On one hand, this requires common software components (versioning support as part of publication procedure and replication software like Synda). On the other hand, operational agreements and the adherence to versioning, replication, and publication best practices are necessary. The presentation will describe the currents status of the software as well as agreement aspects. A short summary of the "replication and versioning" WIP paper will be discussed, along with a roadmap for 2016 highlighting open issues to be resolved. The collaboration aspects with the ICNWG team and the publication team are summarized, as well as future versioning and replication support aspects of the proposed persistent identifier ESGF.
Software Security Working Team ***Prashanth Dwarakanath (ENES/LIU)*** *pchengi@nsc.liu.se*	The security incident discovered in June 2015 not only revealed security vulnerabilities in the ESGF software stack, but also exposed critical lapses in communication between node administrators and developers and highlighted the need for a group to monitor the ongoing status of ESGF from a software security perspective. This led to the formation of the Software Security Working Team. A half-day exercise was also carried out during the ESGF code sprint held in Linköping, Sweden, in September 2015 to discuss security-related issues, develop a best-practices guide for ESGF node administrators, and make recommendations for stipulating mandatory conditions related to security and node operations for sites.

Table continued next page

Day 2: Wednesday, December 9, 2015 ESGF Development for Data Centers and Interoperable Services	
Title and Presenter	**Abstract**
Support Working Team ***Matthew Harris (DOE/LLNL)*** *harris112@llnl.gov*	Last year's presentation covered all the challenges of the technical features of the Support Working Team. This year, with joy, we will cover the new, replaced, and even removed tools for giving our users a better experience. Topics will cover FAQs, wikis, sites, and mail archiving. Although the support process has been significantly improved, continued enhancements are possible with everyone's help.
User Working Team ***Torsten Rathmann (ENES/DKRZ)*** *rathmann@dkrz.de*	In addition to operational support, the User Working Team has been working on the statistics of user questions via esgf-user@lists.llnl.gov and the former Askbot. From December 2013 to September 2015, we received 1,133 requests (excluding spam and ~3 lost Askbot questions). From the statistics, a list of issues shall be distilled, concerning topics such as registration+login, search, Globus Connect, and malfunctioning servers.

Day 3: Thursday, December 10, 2015 Coordinated Efforts with Community Software Projects	
Title and Presenter	**Abstract**
THREDDS Data Server ***John Caron (Independent)*** *jcaron1129@gmail.com*	The THREDDS Data Server (TDS) has developed significantly since first adopted by ESGF. The latest version (5.0) can now scale to the tens of thousands of catalogs and millions of data sets used by ESGF nodes. This presentation will cover these improvements and others of interest to the ESGF community.
Science DMZ for ESGF Super Nodes ***Eli Dart (DOE/ESnet)*** *dart@es.net*	The Science DMZ is a network design pattern that enables data-intensive science by optimizing the network architecture, system design, performance characteristics, security model, and security polices of a specific network enclave for large-scale data transfers. The Science DMZ is now considered best practice for the design, deployment, and operation of cyberinfrastructure for data-intensive science. This talk will describe the Science DMZ model and its application to large (super node) ESGF deployments. In addition, the talk will discuss next-generation portal architectures and their potential applications in the ESGF.
Named Data Networking ***Christos Papadopoulos (Colorado State)*** *christos@colostate.edu*	The Internet currently names hosts, leaving applications to locate the host with the desired data. However, with the emergence of technologies such as Content Delivery Networks and the cloud, along with trends such as mobility and Internet of Things, the need to associate data with an Internet protocol address has become a hindrance. This misalignment requires enormous corrective effort at the expense of application complexity and robust security. This talk will cover named data networking and some of its applications in the climate and high energy physics communities.
Designing Climate Model Output Rewriter, Version 3, for Coupled Model Intercomparison Project, Phase 6 ***Denis Nadeau (DOE/LLNL)*** *nadeau1@llnl.gov*	Many lessons have been learned during CMIP5, and a more flexible version of the Climate Model Output Rewriter (CMOR) has become necessary to handle state-of-the-art model intercomparison projects (MIPs). Flexibility, adaptability, scalability, and robustness are necessary to keep pace with the rapid changes in climate science model development. CMOR is being enhanced to line up with continuously growing CMIP6 requirements. Delineating the structure of new input tables, which empower each model to maintain value delivery into CMIP6, will help enable adoption of those requirements. Additionally, customized global attributes are being designed to accommodate growth in capabilities needed by new MIPs. Finally, the possibility of CMOR parallelization is also being regarded as improvement, since model outputs continuously grow in spatial and temporal resolution.

Table continued next page

Day 3: Thursday, December 10, 2015 Coordinated Efforts with Community Software Projects	
Title and Presenter	**Abstract**
Synda ***Sébastien Denvil (ENES/IPSL)*** *sebastien.denvil@ipsl.jussieu.fr*	Synda is a command-line alternative to the ESGF web front end. Current main features include: • Simple data installation using an apt-get like command • Support for every ESGF project (e.g., CMIP5, CORDEX, and SPECS) • Parallel downloads, incremental process (download only what is new) • Transfer priority, download management and scheduling, and history stored in a database • GridFTP enabled • Hook available for automatic publication upon completion of data sets download • Installation using a docker container and/or Red Hat Package Manager Synda can download files from the ESGF archive in an easy way based on a list of facets (e.g., variables, experiments, and ensemble members). The program evolves together with the ESGF archive back-end functionalities. This talk will walk through Synda's main features from the perspectives of replication and replica publication. Also, ESGF currently only supports an "offline, on-demand" replication procedure, where dedicated replication sites pull replica sets from ESGF sites, reorganize them to fit into their internal ESGF data organization structure, and publish them as "replicas" into the ESGF data federation. No automatic replica synchronization or notification mechanisms are supported. In general, original data can be unpublished or modified without effects on replica sites. We will discuss the current work plan and expose existing possibilities toward an automatic replication workflow.
Globus and ESGF ***Rachana Ananthakrishnan (DOE/ANL)*** *ranantha@uchicago.edu*	Globus provides a hosted research data management service and is widely used for moving and sharing research data on a variety of HPC and campus computing resources. With the recent release of data publication and discovery capabilities, Globus now provides useful tools for managing data across the research lifecycle. This talk will present an overview of Globus capabilities, including recently released features, and provide a quick look at some features that will be released soon. The presentation will discuss how ESGF uses Globus today and opportunities for future work for further ESGF leveraging of Globus.
On-Demand Streaming of Massive Climate Simulation Ensembles ***Cameron Christensen (University of Utah)*** *cam@sci.utah.edu*	The increasing size of climate data sets is a burden that impedes analysis and visualization tasks due to limited storage space, computing power, and network bandwidth available to clients. Our work addresses this issue by providing a framework for interactive visualization and user-directed analysis of massive remote climate simulation ensembles. This framework enables visualization parameters, ensemble members, and analysis scripts to be modified on the fly. It can be used for experimenting with various combinations of analyses for later use in a comprehensive global computation or directly for out-of-core visualization and analysis tasks at any resolution. The framework supports server-side data blending and regridding to minimize client-side storage, computation, and network bandwidth requirements. Python or Java wrappers of the entire framework provide scripting integration, and a JavaScript-based syntax is utilized for in-application dynamic scripting. The framework is built on the IDX multiresolution data format, so we also provide server-side, on-demand data reordering for requested fields of an ensemble to seamlessly use our analysis and visualization tools. Other users cache this data for later access. On-demand data reordering enables streaming multiresolution access and processing of remote data sets that can be too large to download directly. Even devices that could not store a single time step of data could be utilized for visualization and analysis of climate data ensembles. This system has been deployed at Lawrence Livermore National Laboratory and is currently being integrated with the ESGF front end. The client application is available for download from the University of Utah. No modifications to existing data format or infrastructure need to be made for this technology to be utilized by end users. We will present our streaming analysis and visualization framework and demonstrate on-demand data conversion using both disparately located and massive data sets.

Day 3: Thursday, December 10, 2015 Poster Session	
Title and Presenter	**Abstract**
The climate4impact Portal ***Maarten Plieger (ENES/KNMI)*** *maarten.plieger@knmi.nl*	The climate4impact portal aims to enhance both the use of climate research data and the interactions among the climate effect/impact communities. The portal is based on impact use cases from 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 6 years, and its development continues within IS-ENES2. Because the climate impact community is very broad, it currently is the portal's primary focus. This work has resulted in the ENES portal interface for climate impact communities (climate4impact.eu/). Climate4impact is connected to ESGF. A challenge was to describe the available model data and how it can be used. The portal warns users about possible pitfalls when using climate models. All impact use cases are described in the documentation section using highlighted keywords pointing to detailed information in the glossary. The main goal for climate4impact can be summarized by two objectives. The first is to work on a web interface that generates a graphical user interface on WPS endpoints. These endpoints calculate climate indices and subset data using OpenClimateGIS/icclim on data stored in ESGF data nodes. Data are transmitted from ESGF nodes over secured OpenDAP and become available in a new per-user secured OpenDAP server. The results are visualized using ADAGUC. Dedicated wizards for processing of climate indices are developed in close collaboration with users. The second objective is to expose climate4impact services to offer standardized services that can be used by other portals such as the EU FP7 CLIPC portal. This standardization has the advantages of adding interoperability among several portals and enabling the design of specific portals aimed at different impact communities, either thematic or national. In this presentation the climate4impact architecture will be detailed, along with visualization (WMS), processing (WPS), security (OpenID, X509, OAuth2), discovery, and download components.
ACME Dashboard ***Matthew Harris (DOE/LLNL)*** *harris112@llnl.gov*	Supporting the Accelerated Climate Modeling for Energy (ACME) community in model development, testing, and usage requires the utilization of many complex and ever-changing components, from model modules and script versions to computer systems and diagnostics. In particular, it is often difficult in collaborative development efforts to keep track of the latest version of specific models and scripts in which set-up parameters are used by collaborators or runs still need to be completed. The ACME Dashboard is an integrated development environment that aims to support the required "bookkeeping" and coordination effort by integrating into one graphical environment secure resource access (e.g., storage and computing), component registers (e.g., data, models, diagnostics, and workflows), provenance (usage information), and work execution (e.g., run workflow and use diagnostics).
HPSS Connection to ESGF ***Sam Fries (DOE/LLNL)*** *fries2@llnl.gov* ***Alex Sim (DOE/ LBNL)*** *asim@lbl.gov*	Accessing data stored on tape archives is difficult, time consuming, and prone to error. The ACME project plans to create hundreds of terabytes or petabytes of data, not all of which are feasible to store on disk-based archives. To address this challenge, we are bridging high-performance storage systems (HPSS) and ESGF, allowing data sets stored on tape to be accessed through the same methods with which climate scientists are already familiar. The Berkeley Archival Storage Encapsulation (BASE) library at Lawrence Berkeley National Laboratory provides a simple API for retrieving metadata as well as actual data from HPSS and other storage systems. We are creating a Python web application that uses BASE to access and retrieve data and allow that data to be published to ESGF. Our initial platform will test HPSS at the National Energy Research Scientific Computing Center (NERSC) with ESGF nodes at Lawrence Livermore National Laboratory, with plans to deploy at other ACME sites such as the Oak Ridge Leadership Computing Facility and Argonne Leadership Computing Facility.

Table continued next page

Day 3: Thursday, December 10, 2015 Poster Session	
Title and Presenter	**Abstract**
Distributed Resource for the ESGF Advanced Management (DREAM) ***Dean N. Williams (DOE/LLNL)*** *williams13@llnl.gov* ***Luca Cinquini (NASA/JPL)*** *Luca.Cinquini@jpl.nasa.gov*	We envision that the DREAM project will accelerate discovery by enabling climate researchers, among other scientists, to manage, analyze, and visualize data from Earth-scale measurements and simulations. DREAM's success will be built on proven components that leverage existing services and resources. A key building block for DREAM will be ESGF. Expanding on the existing ESGF, the project will ensure that the access, storage, movement, and analysis of the large quantities of data that are processed and produced by diverse science projects can be dynamically distributed with proper resource management. Much of the DOE Office of Science data are currently generated by multiple stand-alone facilities. DREAM can collect data accumulated from these facilities and incorporate it into a fully integrated network accessible from anywhere in the world. The result is a paradigm shift for data management, analysis, and visualization, enabling researchers to: • Manage their calculations, data, tools, and research results. • Ensure that all data are sharable, reproducible, and (re)usable—accompanied by appropriate metadata describing provenance, syntax, and semantics at data creation. • Advance application performance by selectively adapting APIs and services in response to scientific requirements and architectural complexities. • Provide scalable interactive resource management (navigate data and metadata at multiple levels and provide architecture-aware data integration and tools for analysis and visualization). We will engage closely with DOE, NASA, and NOAA science groups working at leading-edge compute facilities. These engagements—in domains such as biology, climate, and hydrology—will allow us to advance disciplinary science goals and inform our development of technologies that can accelerate discovery across DOE and other U.S. agencies more broadly.
Observation Data Publication into the ESGF ***Misha B. Krassovski (DOE/ORNL)*** *krassovskimb@ornl.gov* ***Tom Boden (DOE/ORNL)*** *bodenta@ornl.gov* ***Dali Wang (DOE/ORNL)*** *wangd@ornl.gov*	The ESGF software infrastructure was created to house climate change model output and provide tools to access and analyze it. As Earth system models (ESMs) become more sophisticated, we will need new ways to validate and test models and facilitate collaborations among field scientists, data providers, modelers, and computer scientists. To facilitate ESM validation and testing, the Carbon Dioxide Information Analysis Center (CDIAC) published part of the AmeriFlux data collection to the ESGF system. Although it does not make sense to publish CDIAC's entire data collection to ESGF (more than 1,500 exist), 11 highly relevant and popular data sets were selected and are expected be published to ESGF by December 2015. These data sets have different origins and data formats (e.g., gridded, point source, and vertical profiles) and thus require considerable effort to publish and create appropriate metadata in order to make them harvestable and visible by the ESGF software stack.
Climate Data Management System, Version 3 ***Denis Nadeau (DOE/LLNL)*** *nadeau1@llnl.gov* ***Dean N. Williams (DOE/LLNL)*** *Williams13@llnl.gov* ***Charles Doutriaux (DOE/LLNL)*** *doutriaux1@llnl.gov* ***Jeff Painter (DOE/LLNL)*** *painter1@llnl.gov*	The Climate Data Management System (CDMS) is an object-oriented data management system specialized for organizing multidimensional, gridded data used in climate analyses for data observation and simulation. The basic unit of computation in CDMS3 is the variable, which consists of a multidimensional array that represents climate information in four dimensions corresponding to time, pressure level, latitude, and longitude. As models become more precise in their computations, the volume of data generated becomes bigger and more difficult to handle due to the limit of computational resources. Models today can produce data as frequently as every hour, three hours, or six hours with a spatial footprint close to satellite data. The amount of time for scientists to analyze the data and retrieve useful information is increasingly unmanageable. Parallelizing libraries such as CMDS3 would ease the burden of working with such big data sets. Multiple approaches of parallelizing are possible. The most obvious one is embarrassingly parallel or pleasingly parallel programming where each computer node processes one file at a time. A more challenging approach is to send a piece of the data to each node for computation and have each node save results in a file as a slab of data. This approach is possible with Hierarchical Data Format 5 using the Message Passing Interface. A final approach would be the use of Open Multi-Processing API (OpenMP), where a master thread is split into multiple threads for different sections of the main code. Each method has advantages and disadvantages. This poster brings to light the benefit of each method and seeks to find an optimal solution to compute climate data analyses efficiently using one or a mixture of these parallelized methods.

Table continued next page

Day 3: Thursday, December 10, 2015 Poster Session	
Title and Presenter	**Abstract**
Ultrascale Visualization Climate Data Analysis Tools ***Aashish Chaudhary (Kitware)*** *aashish.chaudhary@kitware.com*	The Ultrascale Visualization Climate Data Analysis Toolkit (UV-CDAT) is a collaborative effort led by Lawrence Livermore National Laboratory. The project's goal is to provide sophisticated data analysis and visualization capabilities at the fingertips of climate scientists. In the past year, we have made tremendous improvements to UV-CDAT, including revamping our plotting capabilities using VTK and MatplotLib, increasing performance and fixing bugs for read and write operations, and enhancing toolkit documentation. In 2016, we are aiming for Sphinx-based documentation, support for system packages, new visualizations, and revamped CDMS for faster performance with parallel capabilities.
CDATWeb ***Matthew Harris (DOE/LLNL)*** *harris112@llnl.gov* ***Jonathan Beezley (Kitware)*** *jonathan.beezley@kitware.com*	CDATWeb is a client/server model for UV-CDAT. With the ever-growing size of data, users' ability to download data and create visualizations on their own hardware is becoming increasingly cumbersome. The CDATWeb visualization server enables users to view and analyze simulation output in place rather than locally, eliminating the need to transfer large data sets over theInternet.
PROV ***Ben Evans (NCI/ANU)*** *Ben.Evans@anu.edu.au*	Capturing provenance information within a computational environment poses new challenges for information infrastructure. NCI has deployed the Provenance Management System (PROMS) to capture workflows for computation, processing, analysis, and publication.
ES-DOC ***Mark Greenslade (ENES/IPSL)*** *momipsl@ipsl.jussieu.fr* ***Sylvia Murphy (NOAA/ERSL)*** *sylvia.murphy@noaa.gov* ***Allyn Treshansky (NOAA/ERSL)*** *allyn.treshansky@noaa.gov* ***Cecilia DeLuca (NOAA/ERSL)*** *cecelia.deluca@noaa.gov* ***Eric Guilyardi (ENES/IPSL)*** *Eric.Guilyardi@locean-ipsl.upmc.fr* ***Sébastien Denvil (ENES/IPSL)*** *sebastien.denvil@ipsl.jussieu.fr* ***Bryan Lawrence (ENES/STFC)*** *bryan.lawrence@ncas.ac.uk*	In 2015, the Earth System Documentation (ES-DOC) project began its preparations for CMIP6 by further extending the ES-DOC tooling ecosystem in support of Earth system modeling documentation creation, search, viewing, and comparison. ES-DOC's online questionnaire, desktop notebook, and Python toolkit will serve as multiple complementary pathways to generating CMIP6 documentation. We envision that institutes will leverage these tools at different points of the CMIP6 lifecycle and will be particularly interested to know that the documentation burden will be either streamlined or completely automated. As all the tools are tightly integrated with the ES-DOC web service, institutes can be confident that the latency between documentation creation and publishing will be reduced to a minimum. Published documents will be viewable with the online ES-DOC viewer (accessible via citable URLs). Model intercomparison scenarios will be supported using the ES-DOC online comparator tool, which is being extended to (1) support comparison of both model descriptions and simulation runs and (2) greatly streamline the effort involved in compiling official tables. The entire ES-DOC ecosystem is open source and built on open standards such as the Metafor Common Information Model (Versions 1 and 2).
Agreement on Data Management and Publication Workflow ***Guillaume Levavasseur (ENES/IPSL)*** *glipsl@ipsl.jussieu.fr* ***Ag Stephens (ENES/BADC)*** *ag.stephens@stfc.ac.uk*	The ESGF publication workflow strongly depends on the data management of each data node. Consequently, the high flexibility of the publication command line allows partner institutes to build their own publication workflows according to their local data policies. Unfortunately, without common use of the publication tools, the ESGF archive became difficult to use and manage, especially for projects containing thousands of data sets (e.g., CMIP5). To ensure high data quality for CMIP6, the IS-ENES Data Task Force is investigating the ESGF publication workflow, taking into account as many use cases of existing data management from ESGF partners as possible. We defined and agreed on several points: • The role and tasks at each data node have to be clearly defined and declared. Who provides, manages, and/or publishes the data? • A modular design could be useful to manage the metadata redundancy among the Postgres database, THREDDS catalogs, and Solr index. • A review of publication tools is required to avoid incorrect use of the publisher and to follow CMIP6 versioning requirements. • A publication test suite is needed. We aim to promote best practices in publishing our CMIP6 data, using enforcement that will be implemented in the publisher code, and improving the end-user experience through new ESGF services (e.g., PID and errata).

Table continued next page

Day 3: Thursday, December 10, 2015 Poster Session	
Title and Presenter	**Abstract**
Data Citation Service ***Martina Stockhause (ENES/DKRZ)*** *stockhause@dkrz.de* ***Katharina Berger (ENES/DKRZ)*** *berger@dkrz.de*	The review of the CMIP6 data citation procedure resulted in the requirement of a citation possibility prior to the long-term archival of the data at IPCC's Data Distribution Centre (DDC) hosted at DKRZ. A concept for a new citation module was developed and described in the WIP paper "CMIP6 Data Citation and Long-Term Archival." This module consists of a repository, a graphical user interface for data ingestion, and an API for data access. This new component has to be integrated in the overall CMIP6 infrastructure. Several connections exist to the long-term archival, the ESGF development (especially the CoG portal, data versioning, and data replication), and the other components providing additional data information (e.g., CIM documents) such as quality information and other annotations. The poster gives a short summary of the citation concept for CMIP6 and the relations between the concept for data citation and long-term archival to other CMIP6 infrastructure components. The focus is the implementation of the citation concept and the technical integration of the citation module into the ESGF infrastructure, which is part of the ESGF Quality Control Working Team's efforts.
PCMDI's Metrics Package ***Paul Durack (DOE/LLNL)*** *durack1@llnl.gov* ***Peter Gleckler (DOE/LLNL)*** *gleckler1@llnl.gov*	Model intercomparison projects (MIPs) provide an effective framework for organizing numerical experimentation and enabling researchers to analyze model behavior and performance. To further our understanding of climate variability and change, the WCRP's CMIP (Taylor et al. 2012; Meehl et al. 2014; Eyring et al. 2016) coordinates a host of scientifically focused MIPs that address specific processes or phenomena (e.g., clouds, paleoclimates, climate sensitivity, and climate responses to natural and anthropogenic forcing). By adopting a common set of conventions and procedures, CMIP provides opportunities for a broad research community to readily examine model results and compare these to observations. Leveraging upon the successes of MIPs, we introduce a new MIP evaluation package—the CMIP PCMDI Metrics Package (PMP; Gleckler et al., 2015). The PMP leverages the vast CMIP data archive and uses common statistical error measures to compare model-simulated climate to observations. The current release includes well-established large- to global-scale mean climatological performance metrics and consists of four components: (1) analysis software, (2) an observationally based collection of observations, (3) a database of performance metrics computed from all the models contributing to CMIP, and (4) usage documentation. The PMP (Doutriaux et al. 2015) is Python-based and utilizes a "lean" version of UV-CDAT (Williams et al. 2015), a powerful analysis package that enables cutting-edge analysis, diagnostic, and visualization capabilities. It is designed to enable users unfamiliar with Python and UV-CDAT to test their own models or observational estimates by leveraging the considerable CMIP infrastructure.
UV-CDAT Metrics ***Jeff Painter (DOE/LLNL)*** *painter1@llnl.gov* ***Brian Smith (DOE/ORNL)*** *smithbe@ornl.gov*	UV-CDAT Metrics is a new framework for climate scientists to analyze, verify, and compare output from multiple model runs (or observational data sets). It implements the functionality of most of the NCAR NCL-based land and atmosphere diagnostics in a more flexible, extensible Python-based framework that uses CDAT and VCS to plot (or summarize in tabular form) typical diagnostic plots. UV-CDAT Metrics has full command-line support, and individual diagnostics can be run from within UV-CDAT. Data produced by the package are in standard formats (e.g., NetCDF, XML, and PNG files) and can be further analyzed or manipulated by scientists in UV-CDAT or other tools. The framework supports recreation of the NCAR plots in their entirety or by an individual plot set. Additional variables, regions, seasons, or variable options can be added easily to expand available diagnostics. Entirely new plot sets can be added as well, and the framework also supports "loose coupling," where existing scripts written in NetCDF Operators (NCO), R, and other languages can be integrated. Current efforts are focusing on multiple levels of parallelization—both computation of individual diagnostics and the running of multiple separate diagnostic computations in parallel.

Table continued next page

Day 3: Thursday, December 10, 2015 Poster Session	
Title and Presenter	**Abstract**
ESMValTool ***Stephan Kindermann (ENES/DKRZ)*** *kindermann@dkrz.de*	The ESM Evaluation Tool (ESMValTool) provides community diagnostic and performance metrics for routine evaluation of Earth system models (ESMs), especially in CMIP6. The effort's priority so far has been to target specific scientific themes focusing on selected essential climate variables and a range of known systematic biases common to ESMs. To support CMIP6, ESMValTool-based processing services must be deployed "near to" ESGF nodes, providing fast access to large amounts of model data (local as well as replicated). An approach for the ESGF integration has been developed and is in a testing phase. This poster will provide an overview of the ESMValTool and the ESGF integration solution, which was implemented. The implementation exploits local ESGF caches as well as a Synda tool–based replication from remote sites. Also, first experiments were done do integrate the ESMValTool in a web processing service framework.
ESGF-QCWT: CMIP6 Errata as a New ESGF Service ***Guillaume Levavasseur (ENES/IPSL)*** *glipsl@ipsl.jussieu.fr* ***Sébastien Denvil (ENES/IPSL)*** *sebastien.denvil@ipsl.jussieu.fr*	Because of the inherent complexity in experimental protocols for projects such as CMIP5 or CMIP6, recording and tracking the reasons for data set version changes is important. During CMIP5, it was impossible for scientists using data sets hosted by ESGF to easily know whether they were using a data set having a known problem and whether this problem was corrected by a newer version. Access to a description of this issue also was very difficult. Movement toward a better errata system is motivated by key requirements: • Provide timely information about newly discovered issues. Because errors cannot be eliminated entirely, we should implement a centralized public interface to data providers, so that they can directly describe problems when discovered. • Provide known issues information prior to download. The user has to be informed of known issues before downloading through the ESGF search interface. • Enable users to interrogate a database to determine whether modifications or corrections have been applied to data they have downloaded. This service could rely on unique file identifiers so end users can discover whether files of interest to them have been (1) affected by known issues, (2) withdrawn, and (3) modified or corrected. • Develop, as part of the errata system, a capability to notify end users of updates to files of interest to them. The ESGF Quality Control Working Team aims to define and establish a stable and coordinated procedure to collect and give access to errata information related to data sets hosted by ESGF.
Enabling *in situ* Analytics in the Community Earth System Model via a Functional Partitioning Framework ***Valentine Anantharaj (DOE/ORNL)*** *anantharajvg@ornl.gov*	Efficient resource utilization is critical for improved end-to-end computing and workflow of scientific applications. Heterogeneous node architectures, such as the graphics processing unit (GPU)–enabled Titan supercomputer at the Oak Ridge Leadership Computing Facility, present us with further challenges. In many HPC applications on Titan, the accelerators are the primary compute engines, while the central processing units (CPUs) orchestrate the offloading of work onto the accelerators and the movement of the output back to the main memory. On the other hand, for applications that do not exploit GPUs, the CPU usage is dominant while the GPUs idle. We utilized a heterogeneous functional partitioning (HFP) runtime framework that can optimize usage of resources on a compute node to expedite an application's end-to-end workflow. This approach is different from existing techniques for *in situ* analyses in that it provides a framework for on-the-fly, on-node analysis by dynamically exploiting underutilized resources therein. We have implemented in the Community Earth System Model (CESM), a new concurrent diagnostic processing capability enabled by the HFP framework. Various single-variate statistics, such as means and distributions, are computed *in situ* by launching HFP tasks on the GPU via the node local HFP daemon. Since our current configuration of CESM does not use GPU resources heavily, we can move these tasks to GPU using the HFP framework. Each rank running the atmospheric model in CESM pushes the variables of interest via HFP function calls to the HFP daemon. This node local daemon is responsible for receiving the data from the main program and launching the designated analytics tasks on the GPU. We have implemented these analytics tasks in C and used OpenACC directives to enable GPU acceleration. This methodology is also advantageous while executing GPU-enabled configurations of CESM when the CPUs will be idle during portions of the runtime. In our implementation results, we demonstrate that it is more efficient to use the HFP framework to offload the tasks to GPUs instead of doing it in the main application. We observe increased resource utilization and overall productivity in this approach by using HFP framework for end-to-end workflow.

Table continued next page

Day 3: Thursday, December 10, 2015 Poster Session	
Title and Presenter	**Abstract**
The OPTIRAD Project: Cloud-Hosting the IPython Notebook to Provide a Collaborative Data Analysis Environment for the Earth Sciences Community ***Philip Kershaw (ENES/CEDA)*** *philip.kershaw@stfc.ac.uk* ***Bryan Lawrence (ENES/STFC)*** *bryan.lawrence@ncas.ac.uk*	We report on experiences deploying the IPython Notebook on the JASMIN science cloud for the OPTIRAD project and its evolution toward a generic collaborative tool for data analysis in the Earth sciences community. The system has been developed in the context of OPTIRAD (OPTImisation environment for joint retrieval of multi-sensor RADiances), a project funded by the European Space Agency and focused on data assimilation of Earth observation products for land surface applications. This domain presents a number of challenges that have provided drivers for the solution developed: the use of computationally expensive processing algorithms, access to large-volume Earth observation data sets, and the need for shared working within the user community. The IPython Notebook has been gaining traction in recent years as a teaching tool for scientific computing and data analysis. It provides an interactive Python shell hosted in an intuitive, user-friendly, and web-based interface that enables the saving and sharing of sessions. The IPython development community is very active, and recent work has led to the creation of a new package JupyterHub. The name Jupyter reflects the fact that the notebook can now support other languages in addition to Python, such as the statistical package R. JupyterHub builds on the baseline functionality of the notebook but incorporates the ability to support multiple user sessions fronted with the required authentication and access control. These developments are significant because they provide the key capabilities needed to enable notebooks to be hosted via a cloud service. Use of cloud technology makes a powerful combination, enabling the notebook to take advantage of cloud computing's key attributes of scalability, elasticity, and resource pooling. In this way, it can address the needs of long-tail science users of "big data," including an intuitive, interactive interface with which to access powerful compute and storage resources. We describe how the notebook has been used in combination with the package ipyparallel to provide a Python-based API to parallel compute capability. Use of Docker containers and the Swarm cluster management system is facilitating scaling of resources to meet demand. We look at how this and other developments are informing the future evolution of the system.
A NASA Climate Model Data Services End-to-End System to Support Reanalysis Intercomparison ***Jerry Potter (NASA/GSFC)*** *jpotter@ucdavis.edu*	Scientists engaged in reanalyses—essentially reforecasts of past weather using the latest forecast models—are interested in reproducing the success of CMIP5. They are studying reanalysis differences and uncertainties to improve reanalysis techniques. Reanalysis data also allow interdisciplinary scientists to compare their data sets (e.g., biodiversity, water planning, and wind power) with 30 or more years of gridded climate data. These research efforts require large sets of monthly and hourly data, formatted identically to facilitate comparisons. NASA's Climate Data Services is collaborating with the world's five major reanalysis projects to collect these data and present the data sets through distribution, visualization, analytics, and knowledge services, resulting in CREATE-IP.

Appendix C. CMIP Requirements Document

Table 9, this page, describes CMIP requirements that were compiled a decade ago but are still largely relevant.

Table 9. CMIP Requirements Originally Requested in April 2006

Highest Priority	1. Implement aggregation and subsetting access through OPeNDAP.
High Priority	2. Develop a system by which users may register to receive errata messages via email when errors are found in the database. Upon registration, users should be able to elect to receive only preferred messages (e.g., all messages pertaining to a single variable or a single experiment). 3. Uniquely identify (e.g., in a global attribute) each "release" of model output, so that when files are withdrawn and replaced, a user will be able to tell that the new data supersede the old data.
Medium Priority	4. Enable access to CMIP3 data using the new ESGF system. 5. Implement server-side calculations beyond the capabilities covered in No. 1 above (e.g., climatologies, multimodel means, and standard deviations). 6. Further document the impact of the IPCC AR4 archive for the benefit of PCMDI and other groups that contributed data. A highlights page would link to: a. Summary of archive size (including a page showing "data availability"). b. Amount of data downloaded (including graphs of download rate). c. World map showing dots for each registered user and open (larger) circles showing modeling centers that contributed output. d. Summary that highlights the most popular data (by experiment and model) and identifies whether data from simulation ensembles by a single model have been used. e. Scientific impact tracking through citation index forward search. 7. Include in the search capability a "date" option, which would display data sets and files that have been added, withdrawn, or replaced since a user-specified date. 8. Develop an option by which a user who registered for a certain category of data (e.g., a single variable, experiment, or model) can elect to be automatically notified when new data become available.
Low Priority	9. Provide an unsubscribe option for users to remove themselves from the registered users list. This feature could be implemented together with some of the notification options described above. 10. Develop further quality control monitoring programs that can be distributed with the ESG publication software to catalog output.
Lowest Priority	11. Standardize file names for the CMIP3 archive based on data reference syntax (or create links with standard file names). 12. Automatically add user email addresses to the master email list once a new user account is approved. 13. Cache zonal means, global means, and climatologies.

Table continued next page

Update to CMIP Requirements (July 2014)	
Highest Priority	14. Produce a summary (at least annually) of how often each variable and data set is being downloaded. (This information is needed to help evolve the CMIP standard output list.)
High Priority	15. Enable data publishers to publish to multiple projects (e.g., both CMIP5 and PMIP3). For example, this capability would enable scientists looking for all PMIP data to select the PMIP project and locate not only the data produced specifically for PMIP but also PMIP data that are considered CMIP5 data as well. (See use cases in Sections 2 and 4, p. 3 and p. 15, respectively, and Appendix E, p. 69) 16. Implement a standard directory structure for CMIP across all nodes to facilitate automated replication of data or simplify scripting for sophisticated users. 17. Implement a consistent data set "version" treatment and an automated system whereby all replicas can be deprecated when the original is withdrawn or a newer version becomes available. 18. Create a summary of available output (at the data set level), similar to what is available for CMIP3 (www-pcmdi.llnl.gov/ipcc/data_status_tables.htm) or CMIP5 (iacweb.ethz.ch/staff/beyerleu/cmip5/). 19. Harvest information about the temporal range stored for each variable, and make it available to users. 20. Generate a branch timetable showing time equivalence between parent and child experiments for stitching purposes.
Medium Priority	21. Provide a service whereby users can inquire about the status of their current data holdings (e.g., if any of their files have been withdrawn or replaced). 22. Implement a capability for scripted downloading that does not rely on http protocol (e.g., GridFTP) and allows unsophisticated users to easily and successfully download data. 23. Provide an enhanced search capability enabling users to employ logical directives (e.g., "and" or "or") and return a list of models that have produced output satisfying the criteria. This search feature could be done through inquiries to the catalog outside the ESGF search engine. 24. Improve web-based methods for reporting questions, issues, and errors in model output. 25. Provide web-based methods for reporting errors, questions, and requests related to the specifications for CMIP and the standard output. 26. Generate a table showing the "forcings" active in each modeling run. 27. Provide methods for recording the provenance of data sets used in published work. 28. Provide methods for giving credit to modeling groups when their data are used in published work.

Appendix D. Faceted Search Implementation

This appendix provides guidance on which search categories and options are needed for meeting user needs in searching the CMIP archive. This guidance is motivated by the use cases previously described in Sections 2 and 4, p. 3 and p. 15, respectively, with possible implications for data reference syntax and Earth System Grid Federation (ESGF) publication.

D.1 Faceted Search Categories

User needs can be satisfied by defining the search categories listed in **Table 10**, this page, along with several additional options specified in the notes below the table.

D.2 Additional Search Notes

1. Users should be able to specify a "date" that would limit the search to data sets published after that date.
2. Users should be able to instruct the search engine either to return all versions of data sets or only the latest version.
3. When responding to a search request, the search engine shall consider only published ESGF data sets with the activity_id global attribute set to CMIP-DECK, CMIP5, CMIP6-xxxx, or xxxx, where "xxxx" is one of the labels identifying an MIP endorsed by the CMIP panel (e.g., xxxx

Table 10. Search Categories and Related Examples and Sources

Search Category	Examples	Sources
Model cohort	CMIP-DECK, CMIP6, CMIP5	see note 4 in Section D.2, below
Experiment family	DECK+historical, PMIP, ScenarioMIP	see note 5 in Section D.2, below
Experiment	piControl, decadal-1980, decadal-1980-dwnscl, historical-dwnscl	Global attributes (GAs): experiment_id, sub1_expt_id, sub2_expt_id, sub3_expt_id
Model type	Atmosphere-ocean general circulation model, Earth system model, regional climate model	GA source_type
Model	CanCM4, HadGEM2-AO	GA source
Run variant	r1i1p1f1, r1i2p223f3	GA run_variant_id (and variant)
Frequency	mon, 6hr, da, monClim	GA frequency
Realm	atmosphere, ocean, land	GA realm
Variable table	Amon, Omon	GA table_id
Variable name	tas, pr, psl	as in CMIP5
Variable standard name	(see CF conventions)	as in CMIP5
Variable long name	(see CMIP5 tables)	as in CMIP5
Grid	gn, gs-2p5x5, gr-lo, gr-1	GA grid_id
Data node	BADC, PCMDI-9, PCMDI-22, IPCC DDC, IPSL, GFDL, NCAR	publisher's CMIP6 .ini file

might be "PMIP" or "scenarioMIP"). In fact, ESGF should prevent publication of data sets that are incorrectly labeled with one of these reserved activity_ids. The CMIP panel will provide and update the list of experiment_ids included in each activity. (As part of the publication procedure, ESGF should prevent any inconsistency between the experiment_id and activity_id.)

4. The only choices permitted under the "Model cohort" category are CMIP-DECK, CMIP6, CMIP5, or "no selection." A model cohort limits a search to models that meet certain MIP criteria. For each cohort option, the CMIP panel will provide and update a list of qualifying models. If no selection is made under model cohort, then all MIP models will be considered. The list of models will include some that do not meet the criteria of any of the defined cohorts, but at least one of the endorsed MIPs will have defined them as bona fide MIP participants, which makes them of interest to CMIP scientists. The models not qualifying as CMIP models could be found only if no selection were made under the model cohort category (and only if they met all other search criteria). The CMIP panel will provide and update a list qualifying source_ids (i.e., models) associated with each cohort.

   For a model to qualify for the CMIP-DECK model cohort, all four DECK experiments must be contributed to the archive. Ongoing determination of whether this criterion has been met will be difficult for the CMIP panel. Setting up a chron job would make checking which models have met this criterion easier (by interrogating the ESGF catalogs); whenever a new model is found, the information would then be transmitted to the CMIP panel. This process also would help the panel keep the CMIP-DECK list of models up to date. CoG filtering would have to be kept up to date with the list.

5. The selections under the "Experiment family" category will be DECK+historical, as well as the names of each CMIP6-endorsed MIP (e.g., PMIP and scenarioMIP). Some MIPs also may want to include a phase as part of their acronym (e.g., "PMIP3" rather than "PMIP"). Associated with each choice will be a list of qualifying experiments, so different sets of experiments will be considered depending on which choices are made. Some experiments will be associated with multiple entries in the list (e.g., piControl will probably be associated with many of the experiment families). The CMIP panel will provide and update the list of experiments associated with each family.

Appendix E. ESGF Data Center Challenges and Motivating Use Cases

E.1 Lawrence Livermore National Laboratory/Analytics and Informatics Management Systems Department, USA (LLNL/AIMS)

LLNL/AIMS emphasizes three major challenges for future ESGF development:

1. Network performance optimization, which has two different aspects: data replication between ESGF data nodes and user data access (see Fig. 2, this page).

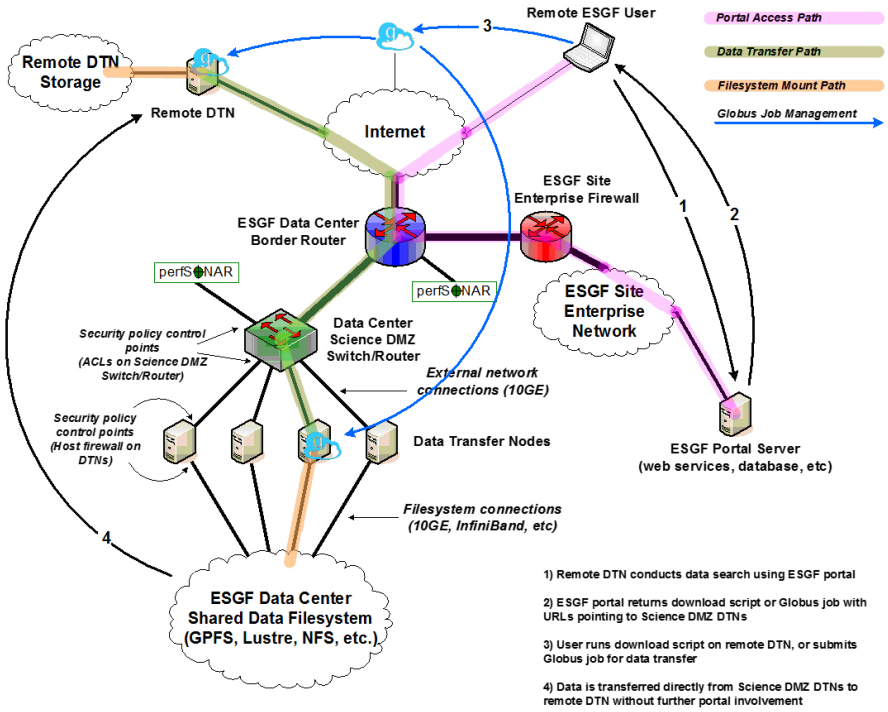


Fig. 2. *Different access paths.* *The replication path (in green) operates directly between ESGF data transfer nodes (DTNs) without passing a firewall. This access path is not valid for all ESGF data nodes, and consideration of firewalls and local data environments will be a challenge as ESGF data replication and user data access are optimized over the coming months.*

2. The ESGF Data Center at LLNL/AIMS (shown in Fig. 2, above) lacks the storage resources required to achieve CMIP6 science goals. If CMIP6 is at a minimum 20 times larger than CMIP5, then an additional 30 PB of spinning disc storage space is needed to meet CMIP6 community requirements.
3. Compute resources for data site data processing and end-user data analysis tools, which already have been requested for CMIP5. LLNL/AIMS' solution, with compute cluster nodes closely located to ESGF data nodes, is shown in Fig. 3, p. 70.

Also presented were motivating use cases for ESGF engagement. In the case of LLNL/AIMS, Accelerated Climate Modeling for Energy (ACME) and CMIP6 were mentioned. The different steps in the ACME workflow, including ESGF data publication of model results, are illustrated in Fig. 4, p. 70.

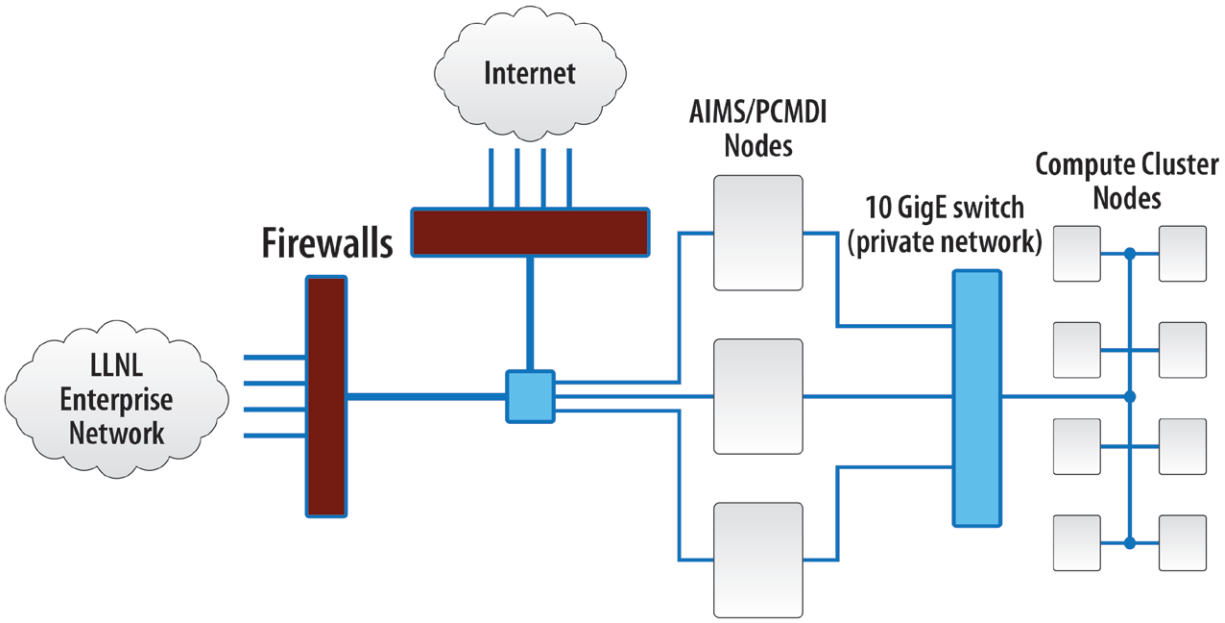


Fig. 3. Such a clear separation between the ESGF infrastructure and an enterprise network is not realized at all sites running ESGF data nodes and not even at all ESGF replication nodes. *A more common picture is the ESGF infrastructure and enterprise network integrated in one common system that shares storage and compute resources. However, this integrated architecture imposes constraints on the accessibility of storage and compute resources via ESGF.*

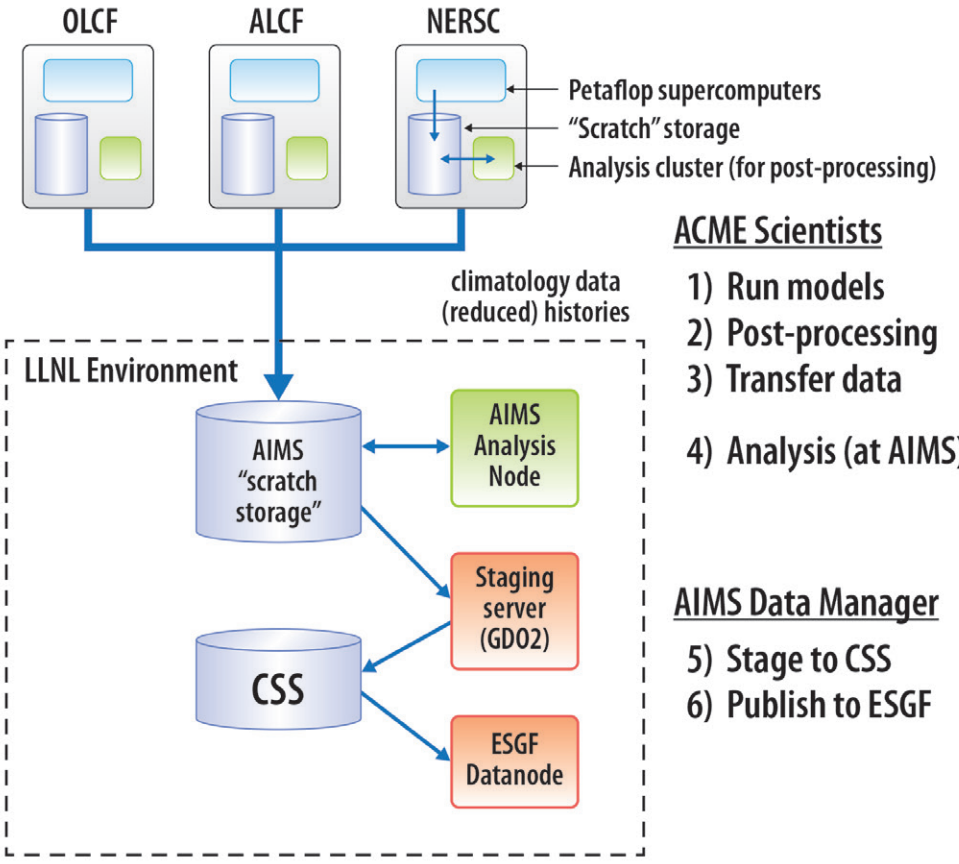


Fig.4 ACME workflow. *The CMIP6 data production workflow is integrated into the ACME workflow and completed by the replication of climate model data from international CMIP6 partners. LLNL/AIMS is one of the replication data nodes that tends to host a considerable amount of data from the entire CMIP3, CMIP5, and CMIP6 projects.*

E.2 National Computational Infrastructure, Australia (NCI)

NCI is a central Australian facility that serves the entire spectrum of Earth sciences and stores more than 10 PB of data from Geoscience Australia, Commonwealth Scientific and Industrial Research Organisation, Australian National University, and other national data archives. This collection includes climate model, Earth observational, geophysical, atmospheric, topographic, hydrological, weather, ocean, marine, and other data sets.

NCI presented a wide variety of challenges for NCI's data infrastructure, including:

1. Migration to transdisciplinary data [i.e., creating unified intellectual frameworks beyond disciplinary perspectives (see Alexander Refsum Jensenius blog at www.arj.no/2012/03/12/disciplinarities-2/) for more details].
2. An integrated scientific computing environment (see Fig. 5, this page, for NCI's solution).

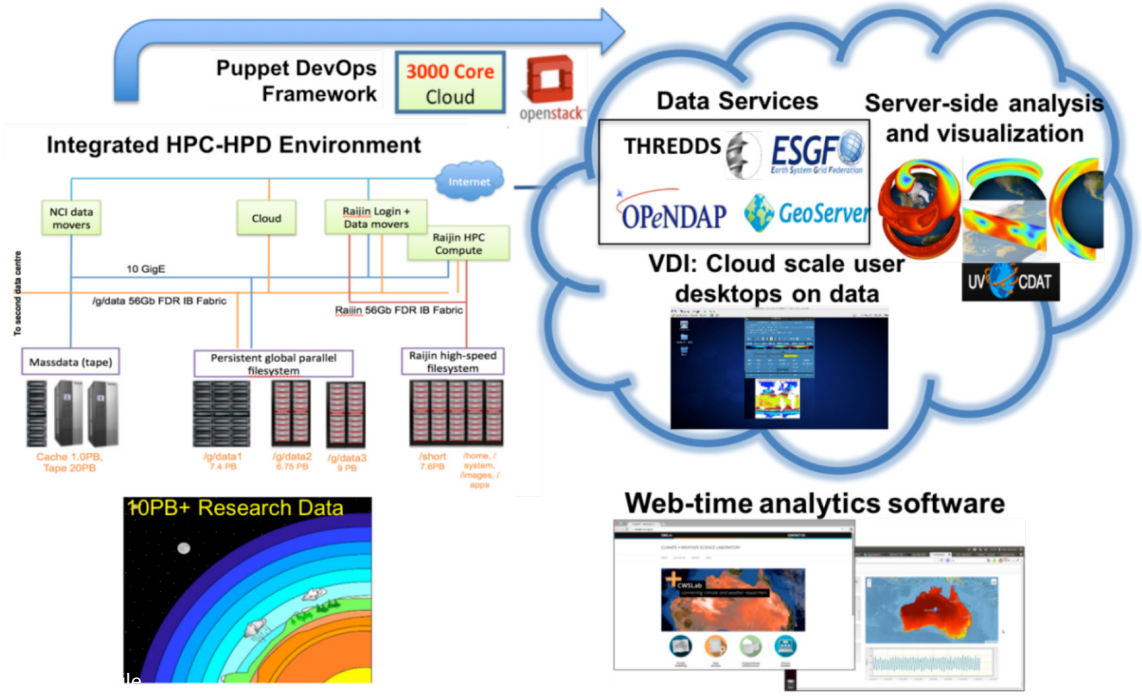


Fig. 5. NCI solution for an integrated scientific computing network. *Integrated scientific computing components include features such as tile map servers, on-the-fly data access and data manipulation, and easy-access analysis environments. Specifically, tools to support NetCDF4/HDF5 with compression are required in the near future.*

3. Data synchronization, which is a specific challenge for global data projects like CMIP6. The emphasis is on ESGF-wide unique versioning to enable automated replication processes.
4. Organization of multipetabyte data archives, which has technical and documentation challenges. Besides the consideration of trade-off performance for storage capacity, implementing a provenance system across NCI (see Fig. 6, p. 72) will increase the scientific usability of the archived data.

NCI specifies two motivating use cases: (1) The National Environmental Research Data Interoperability Platform (NERDIP), which provides a unique interface to a wide variety of geophysical data for diverse research communities (see Fig. 7, p. 72) and (2) CMIP6 and the operation of an Australian ESGF data node emphasizing publication of Australian data and replication from international contributions. Other requirements include quality control and assessment and minting of digital object identifiers (DOIs) for DataCite data publication and deployment of end-user data analysis tools in the NERDIP framework.

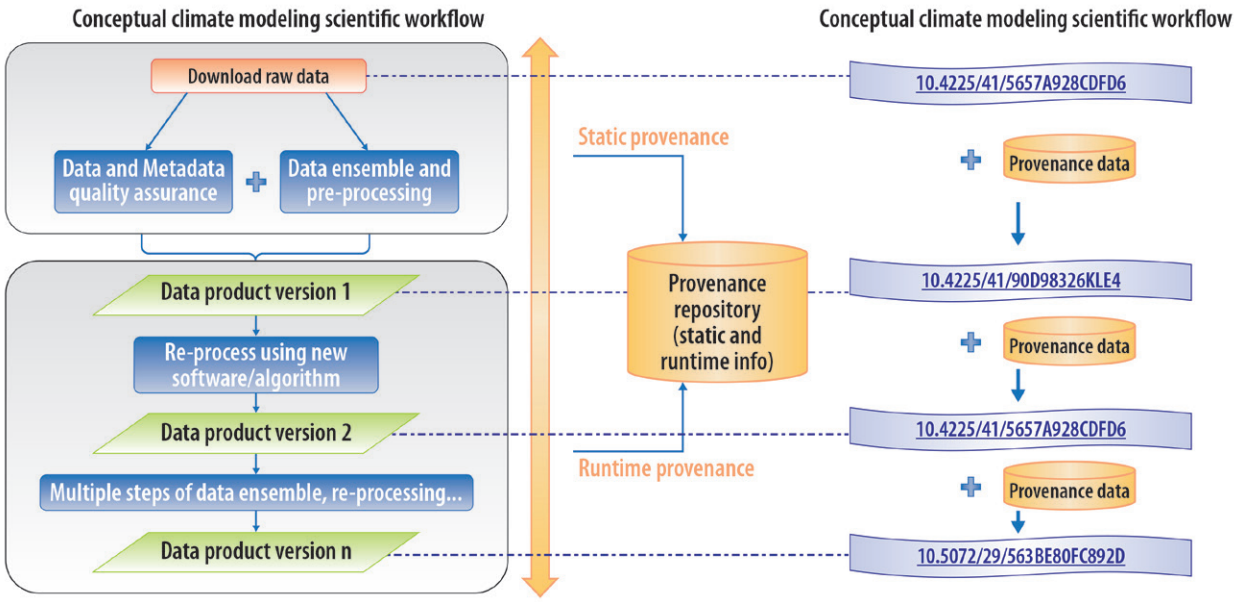


Fig. 6. NCI's climate model data production workflow includes versioning of data products, related data citation management, and provenance information capture. *The Provenance Service captures information at each step within the end-to-end workflow and stores it within the Provenance Repository.*

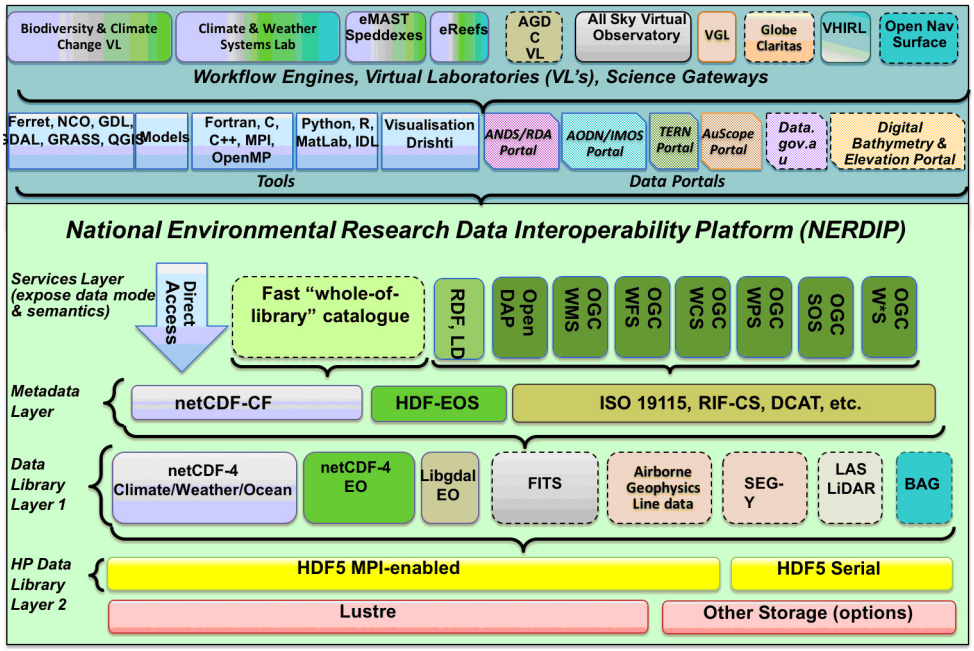


Fig. 7. NERDIP architecture.

E.3 German Climate Computing Centre (DKRZ)

DKRZ-specific challenges related to the center's new high-performance computing (HPC) environment and new ESGF services include:

1. DKRZ's current HPC system merged compute and data service into one system and migrated from a general parallel file system (GPFS) to a Lustre file system. All HPC services are now operating on one system, which

requires modification for DKRZ's integration with ESGF. These modifications include a "national model intercomparison project (MIP) data cache" as part of the ESGF data node and flexible integration of HPC compute resources for ESGF data analysis (see Fig. 8, this page).

2. Long-term archiving and data citation will be continued and further developed within ESGF. Special emphasis is given to the CoG integration of DKRZ's long-term World Data Center Climate (WDCC) data archive (see Fig. 9, this page).

3. An improved version of the CMIP5 quality assurance software will be integrated into ESGF as part of the CMIP6 data management (see Fig. 10, p. 74).

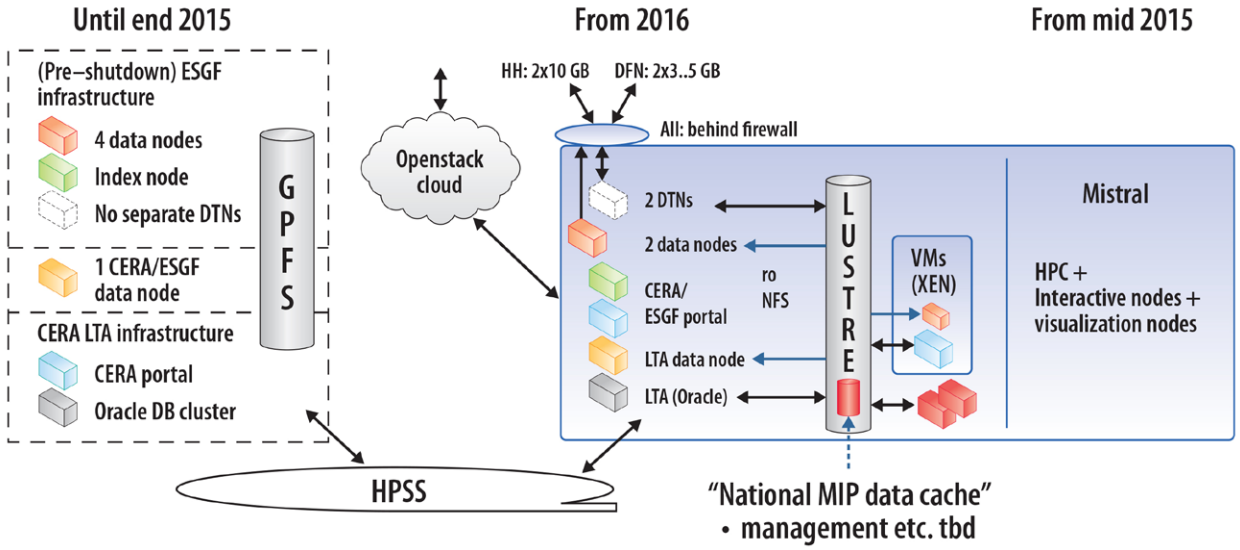


Fig. 8. Integration of compute and data services into Mistral, DKRZ's current HPC system.

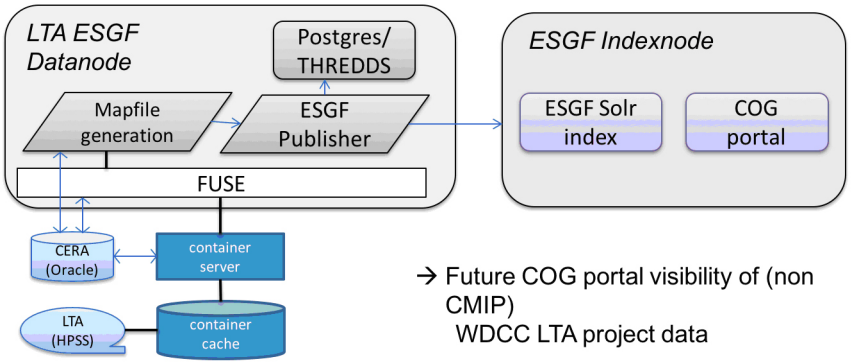


Fig. 9. ESGF Integration of DKRZ's long-term data archive based as a separate long-term archival (LTA) ESGF data node.

4. Also in connection with CMIP6, persistent identifiers (PIDs)—based on the Corporation for National Research Initiative's (CNRI) Handle Server—will replace the standard universally unique identifiers in NetCDF/CF files, moving DKRZ toward PID-based services for versioning and replication.

DKRZ's motivating use cases are the national MIP data analysis platform (see Fig. 11, p. 74) and the center's integration as one of the core ESGF data nodes in the CMIP6 data federation.

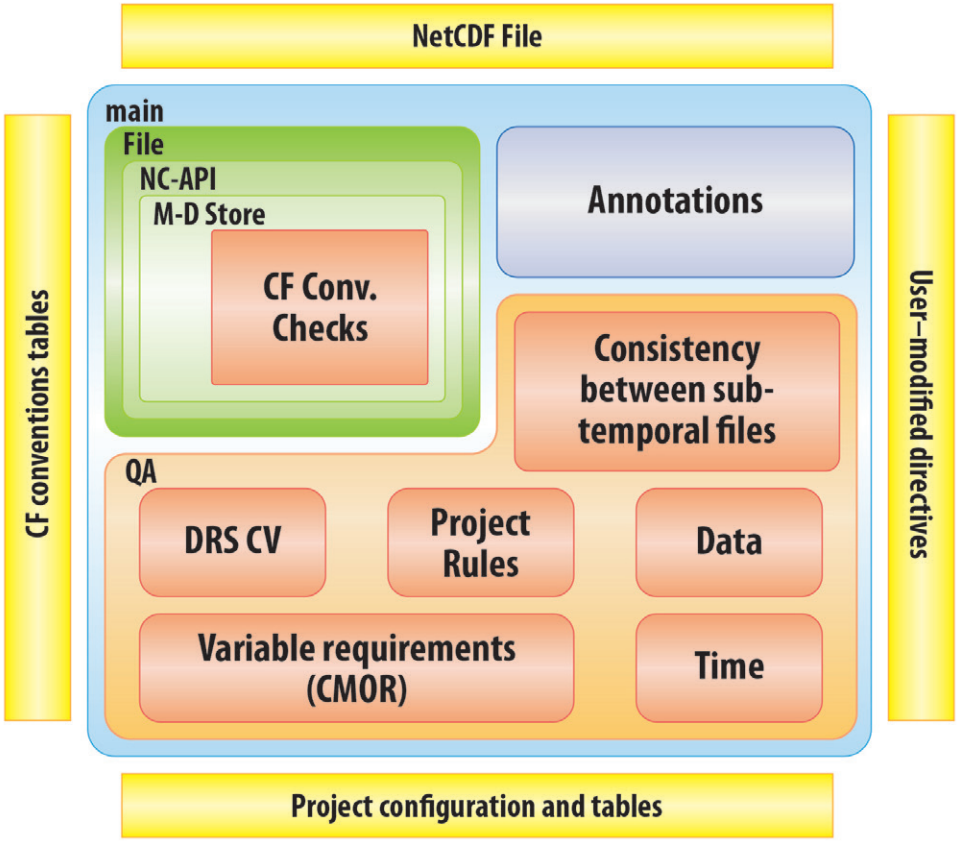


Fig. 10. Architecture of CMIP6 data quality assurance process.

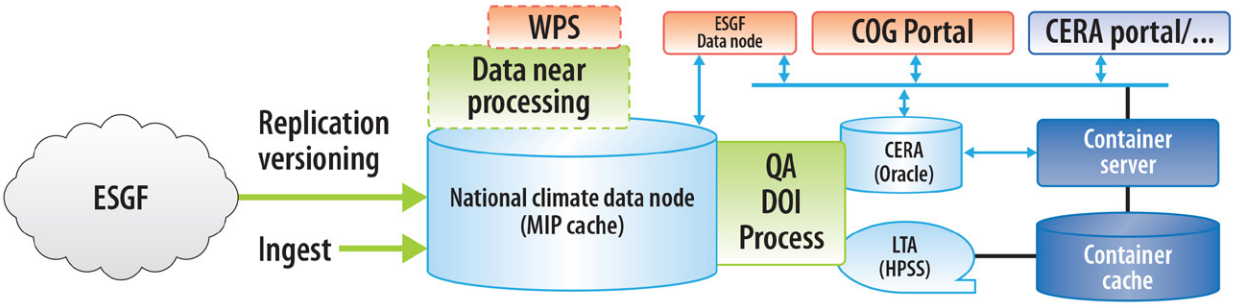


Fig.11. Integration of the German national MIP data analysis platform into DKRZ's ESGF data node.

For CMIP6, emphasis is on the LTA use case including high-performance storage system (HPSS) integration, quality assurance, PID assignment early in the data lifecycle, early data citation reference, and DataCite data publication in the long-term archive for reference data.

E.4 Institut Pierre Simon Laplace, France (IPSL)

IPSL is a federation of nine public research institutes in the field of Earth and planetary environmental sciences. All data gathered by IPSL, whether from field campaigns, observational networks, or numerical simulations, are stored in databases and are available to the scientific community within IPSL and at the national and international levels. Data are transferred to civil society for operational applications.

IPSL's data center challenges for current and future ESGF developments include:

1. Data intercomparison between climate models, ground observations, and satellite observations from different sources.
2. Operation within the European supercomputing ecosystem. Along with other regional centers and universities, IPSL is a "Tier 2" center and provides a national academic platform to analyze CMIP6 outcomes. IPSL is being integrated into the national (Tier 1) and European (Tier 0) HPC landscape.
3. **Adaptation of** ESGF developments to the CMIP6 and IPCC AR6 timelines, which is essential for ESGF acceptance in CMIP6 and the related scientific communities.

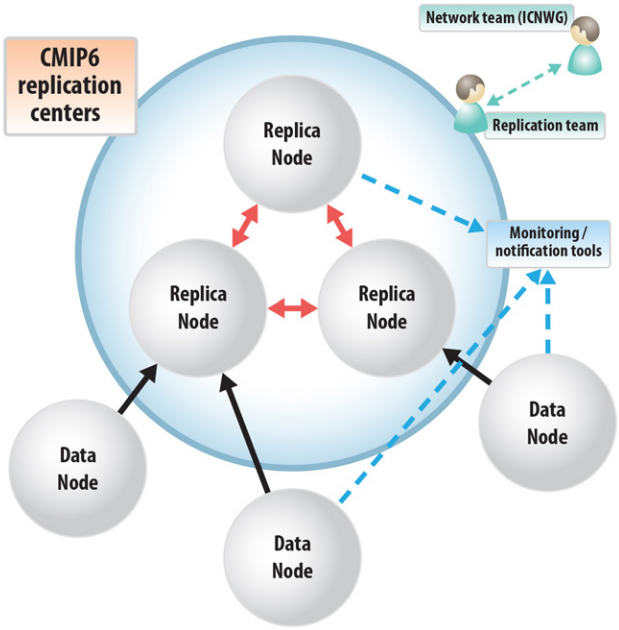


Fig.12. CMIP6 data replication architecture.

Validation of ESGF software release candidates is necessary to improve user confidence in ESGF operational software for scientific work. The suggested ESGF test federation is based on VMware virtual machines. It is completely independent from the production federation and is used to run the ESGF test suite, which performs automated replication and versioning within ESGF. A critical challenge is handling the huge amount of data and number of data entities expected for CMIP6 (see Fig. 12, this page).

IPSL's motivating use cases include a national data analysis environment, which will be operated to serve the national scientific user community, and operation of one of the core CMIP6 ESGF data nodes, which serves the international climate research community and enables data access at the national level.

E.5 Centre for Environmental Data Analysis, United Kingdom (CEDA)

CEDA provides a variety of computational and archival services to the United Kingdom's geophysical scientific community.

CEDA's primary challenges include:

1. Data quality, organization, and discovery by end users.
2. Suitable platforms, algorithms, and well-trained experts to support end users in their scientific work.
3. Development of a wide variety of products targeted to different data consumers.
4. Handling of "big data" volume, velocity, and variety, which will be essential with CMIP6 and necessitate solutions for performance, multitenancy, and flexibility. CEDA will be expected to meet the needs of long-tail science users, ensure continuous data availability, and maximize compute, network, and storage use (see Fig. 13, p. 76).

CEDA's motivating use cases include the European Space Agency's (ESA) Climate Change Initiative, OPTImisation environment for joint retrieval of multisensor RADiances (OPTIRAD) project, and CMIP6. The Climate

Change Initiative is an open data portal built on ESGF architecture that aims to provide essential climate variable data products (see Fig. 14, p. 77).

The OPTIRAD project allows for initial experiences with containers and container orchestration in ESGF (see Fig. 15, p. 77).

CMIP6 is a challenging use case that all major ESGF data nodes share.

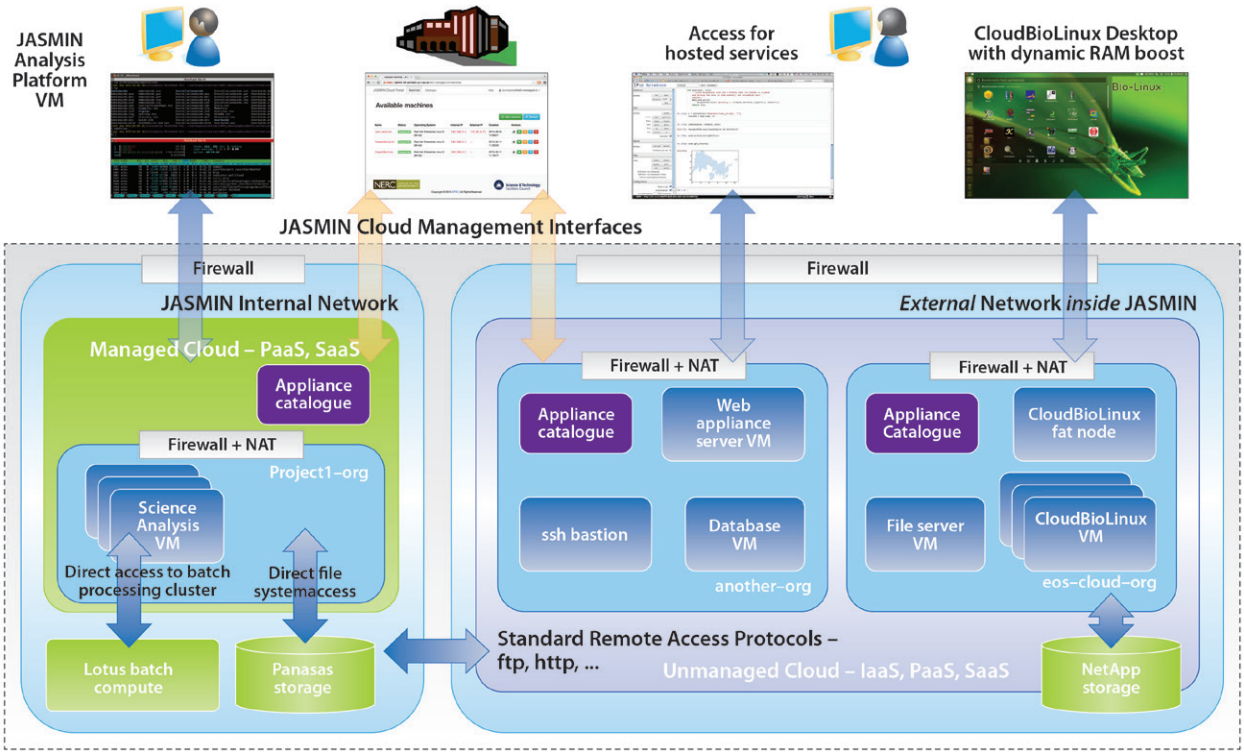


Fig. 13. CEDA's JASMIN analysis platform. *JASMIN integrates cloud architecture, container technologies, and virtual machines to improve flexibility and performance and track maintenance.*

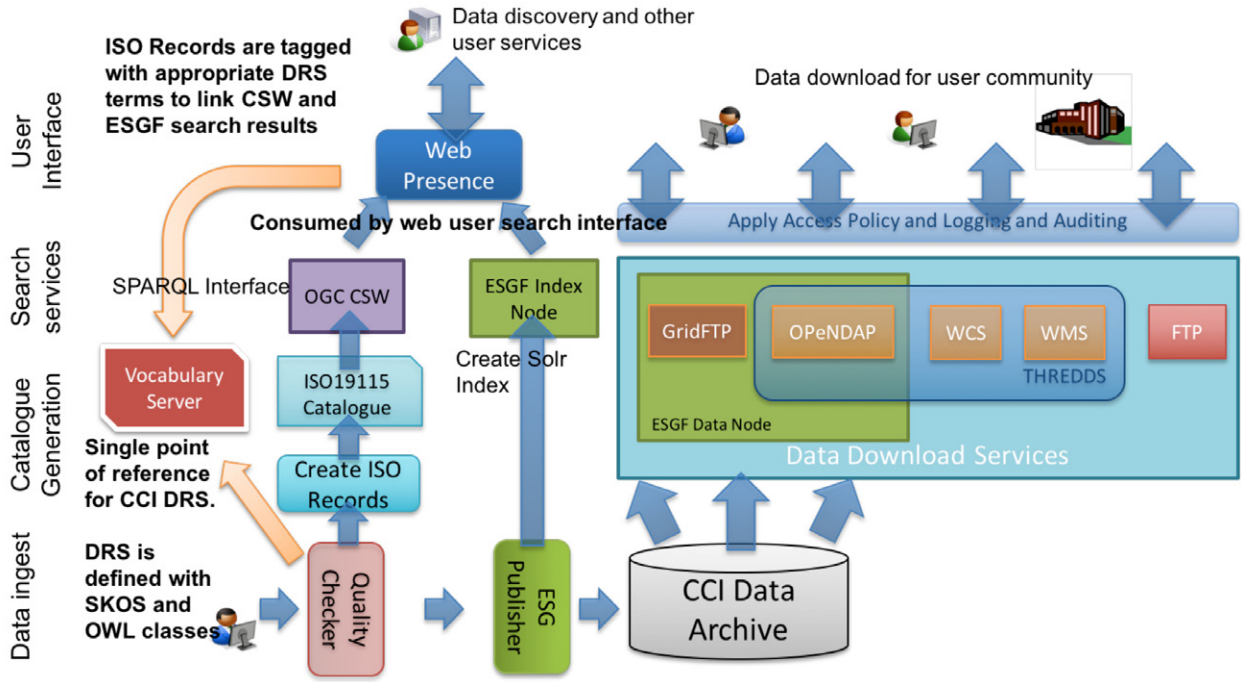


Fig. 14. ESA Climate Change Initiative architecture.

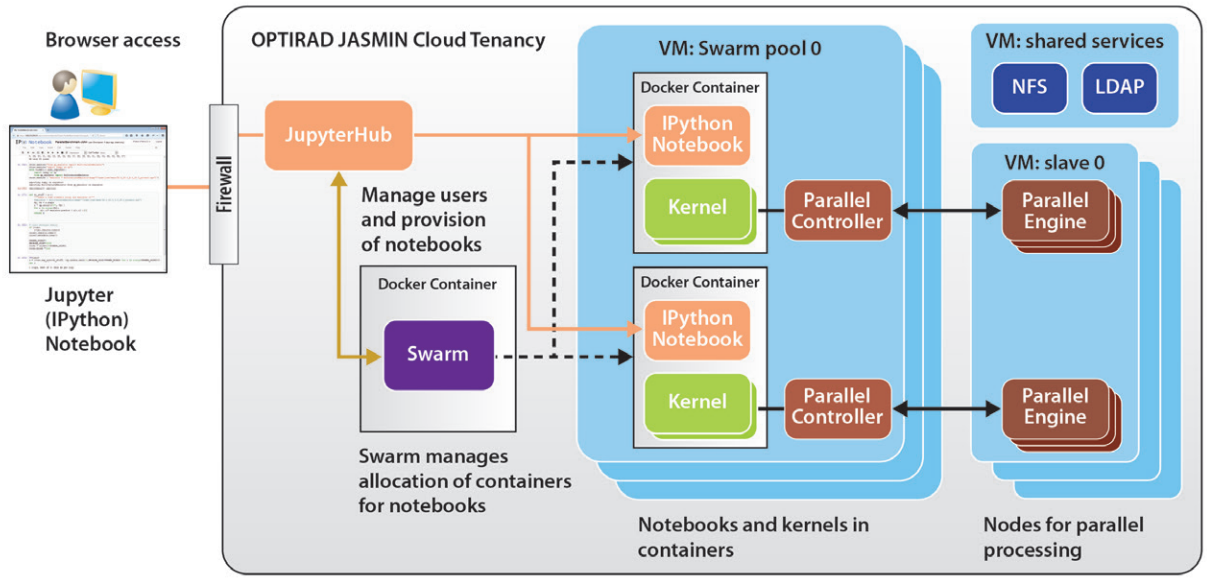


Fig. 15. OPTIRAD deployment architecture.

Appendix F. ESGF Software Security Plan

F.1 Background

The primary purpose of this security plan is to prepare for both major and minor ESGF software releases within the context of the ESGF system development lifecycle (SDLC). This plan's emphasis is on the "release" phase of a typical SDLC and its prerequisites but depends upon development and maintenance aspects of the SDLC as well. Below is a draft of the ESGF Software Security Plan. The completed and signed living document can be found on the ESGF website (esgf.llnl.gov).

SDLC phases prior to a software release

- Requirements definition.
- Design (including secure coding practices and threat modeling).
- Implementation.
- Verification (including security testing).

The release phase of the ESGF SDLC shall include the following:

- Inventory update.
- Change documentation.
- Security review (minor/major).
- Issue resolution.
- Certification of release.

Implement Federal Information Security Management Act (FISMA) Controls SA-3 and SA-8.

F.2 Roles and Responsibilities

F.2.1 ESGF Executive Committee

- Ensure that the ESGF software security plan is agreed upon, signed, and followed by all sites.

F.2.2 ESGF Software Development Team Lead by LLNL

Governance

- Define roles and responsibilities at federation level.
  - Implement FISMA Control SA-2.
- Ensure ESGF points of contact (PoCs) are identified for all sites:
  - Site ESGF Manager PoC.
  - Site ESGF Security PoC.
- Ensure PoC information is maintained and kept current.
  - Implement FISMA Controls CP-3, IR-2, and SA-3.
- Ensure all PoCs have Pretty Good Privacy (PGP) keys and that those keys are managed appropriately.
  - Implement FISMA Controls SC-12, SC-13; en.wikipedia.org/wiki/Pretty_Good_Privacy.

Software Design

- Define goals for future enhancements of the ESGF software design principles that include measures to enhance configuration management.
- Formalize a secure coding standard for ESGF (see **Section F.3**, p. 81).
- Implement FISMA Control SA-11.
- Define ESGF developer roles and responsibilities; at a minimum they shall include:
  - Ensure developers maintain a basic awareness of security principles (e.g., training).
  - Implement an ESGF process for pre-screening third-party libraries and components.
  - Define and formalize an ESGF Change Process that includes enhancing configuration management (e.g., use of Puppet).
  - Implement a formalized peer review process of the ESGF suite and incorporate that into the ESGF Change Process.
  - Conduct threat modeling exercises during roadmapping and long range planning.
  - Include security testing as part of functional testing as an integral aspect of the release process.
  - Implement FISMA Control SA-11.
- Define flaw remediation procedures:
  - Document ESGF software flaws (i.e., bug tracking with Bugzilla, Mantis, Trac, or others).

  - Document ESGF software flaws with security implications.
  - Create a tracking database of flaws and make it accessible to all federation sites (authorization required).
  - Disseminate flaw remediation status to federation.
  - Implement FISMA Controls SI-2 and SA-11.

Software Release

- Define and coordinate minor release procedures (each site is likely to customize this):
  - Ensure independence from the developer team.
- Ensure that the releases are assessed appropriately and that the appropriate Security Review is conducted (see **Section F.4**, p. 81).
- Coordinate all releases and ensure that they include:
  - Current complete software component inventory.
  - Release notes of changes.
  - Current complete source code.
  - Configuration and install scripts.
  - FISMA Controls CM-2, CM-3, CM-6, CM-9, SA-10, and SA-11.
- Certify release for distribution. This is solely the responsibility of the LLNL ESGF team.

Certificates

- Manage and coordinate ESGF certificates:
  - Create ESGF certificates.
  - Distribute ESGF certificates.
  - Maintain ESGF certificates.
  - Implement FISMA Controls IA-5 and SC-12.

Incidents

- Define incident response procedures (to follow local agency and site requirements) including, at minimum, the following:
  - Document incidents.
  - Create a tracking database of ESGF incidents and make it accessible to all federation sites (authorization required).
  - Disseminate incident status to the federation.
  - Define an incident and security call-tree and distribute to all ESGF sites.
  - Implement FISMA Controls IR-4, IR-5, IR-6, and IR-8.

Contingency and Continuity of Operations Plans

- Define contingency plan and Continuity of Operations Plan (COOP) to manage outages, denial of service attacks, and the like:
  - Exercise contingency and COOP plans regularly:
    - » Partial exercise annually at a minimum.
    - » Full exercise every 2 years at a minimum.
  - Implement FISMA Controls CP-2 and CP-4.

Best Practices

- Define, collect, document, and distribute best practices to sites:
  - See **Section F.5**, p. 82.
  - Areas of particular importance include access control, patching, configuration management, account management, incident response, security planning, system development and testing, system and information protection, and monitoring and integrity.
  - Maintain the repository of best practice documentation and distribute updates to all sites.
  - Ensure sites adhere to best practices.
  - Ensure sites contribute to best practices.
- Implement FISMA Controls AC-3, AC-5, AC-6, AC-14, AC-17, AC-22, CM-3, CM-4, CM-6, CM-7, CM-8, IA-8, IR-5, IR-6, PL-2, PL-8, SA-3, SA-8, SA-10, SA-11, SC-2, SC-5, SC-12, SC-13, SC-32, SI-4, and SI-7.

F.2.3 NASA Center for Climate Simulation

NASA Center for Climate Simulation (NCCS) shall conduct the following activities:

Governance

- Define the major release security review procedure (see **Section F.4**, p. 81).
- Maintain the major release procedure.
- Distribute the major release security review procedure to ESGF for review and approval by the Executive Committee.
- Ensure independence from the developer team.

Software Review for Release

- Apply the major release procedure when requested and as per an agreed-upon schedule.
- Communicate results of the Security Review to ESGF and iterate on results as necessary.
- Coordinate the creation of a baseline for recommended configurations of the following:
  - Firewall.
  - Monitoring.
  - Logging.
  - Auditing.
- Ensure baselines are provided back to ESGF for consideration and dissemination as best practices.
- Contribute to minor release security review activities as needed.
- Perform site responsibilities (see below).

F.2.4 ESGF Sites

All ESGF sites shall conduct the following activities:

Governance

- Ensure ESGF PoC contact information is provided to the federation.
- Ensure ESGF PoCs create and maintain PGP keys including the signing of PgP keys by trusted parties (web of trust) and the uploading of those signed keys to a PGP keyserver (e.g., pgp.mit.edu) for distribution and availability.

Incident Response

- Define site-specific incident response procedures and ensure coordination with the federation.

Contingency and Continuity of Operations Plan

- Define a site-specific contingency plan as well as COOP procedures and ensure coordination with the federation, including participation in exercises.

Best Practices

- Follow the best practices as documented by the ESGF community.
- See **Section F.**5, p. 82.

F.3 Secure Software Development Resources

- www.owasp.org/index.php/Secure_SDLC_Cheat_Sheet
- www.securecoding.cert.org/confluence/display/seccode/Top+10+Secure+Coding+Practices
- en.wikipedia.org/wiki/Pretty_Good_Privacy
- pgp.mit.edu

Link to the FISMA assessment case descriptions and downloads

- csrc.nist.gov/groups/SMA/fisma/assessment-cases.html

F.4 Major and Minor Release Security Review Procedures

F.4.1 ESGF Major Release Security Review Procedure

- Responsibility: NCCS Information System Security Officer (ISSO).
- To be performed by NCCS, NASA Jet Propulsion Laboratory (JPL), and others as needed.
- ***Important: Must have independence from developer team.***

Prepare for Audit

- Install release candidate in a test environment (ESGF system administrator).
- Verify inventory and software package (ESGF system administrator and security analyst).
- Review release notes and updated documentation to assess changes (ESGF system administrator and security analyst).
- Develop a plan and schedule for security review, including the use of external resources such as CS Gov, AppSec on Demand, JPL Dynamic scan, and NASA WASP Dynamic scan (NCCS security lead, ESGF system administrator, and security analyst).

Audit Release Candidate

- Schedule a static code scan (e.g., HP Fortify) of the ESGF release (NCCS security lead).

- Perform the common vulnerabilities and exposure (CVE) check (e.g., of "jar" files) of the ESGF release (ESGF system administrator).
- Manually test and scan using local tools; adjust as needed for changes (security analyst).
- Review and code analysis of changed source code (security analyst).
- Review updated configurations (e.g., Apache and Tomcat) (ESGF system administrator).
- Schedule NASA web application security project dynamic scan (e.g., Hailstorm, Nessus scanners) or NASA JPL dynamic scan of installed new version as available (NCCS security lead).
- Activate external resources (e.g., CS Gov Appsec on Demand) as necessary should the assessments and static or dynamic scans be deemed insufficient (NCCS security lead).

Maintenance of Audit Tools

- Perform maintenance of local scanning tools (ESGF system administrator).
- Maintain NCCS local assessment tools (security analyst).
- Maintain access to external scanning resources (e.g., licenses for HP Fortify) (NCCS security lead).
- Update resources and tools in response to ESGF technological changes and requests (NCCS security lead, ESGF system administrator, and security analyst).

Document Audit Results

- Document all high- and moderate-impact issues for tracking and add to ESGF Flaw Tracking Database (NCCS security lead, ESGF system administrator, and security analyst).
- Coordinate the resolution of all high- and moderate-impact issues (NCCS security lead).
- Document for ESGF both the resolved and unresolved issues, recommendations, impacts, and possible risk acceptance for ESGF sites (ESGF system administrator and security analyst).
- ***Important: During this process, feedback to the developer team is crucial for continuous improvement of pre-screening, peer review, threat modeling, and security testing, as well as best practices.***
- Document findings (NCCS security lead, ESGF system administrator, and security analyst).
- Issue final report to ESGF Executive Committee (NCCS ISSO):
- ESGF Executive Committee certifies for release.

F.4.2 ESGF Minor Release Security Review Procedure

- Responsibility: Individual ESGF sites.
- ***Important: Must have independence from developer team.***

Prepare for Audit

- Verify inventory and software package.
- Review release notes.
- Assess changes and define target for CVE check, testing, and code review.

Audit Minor Release Candidate

- Perform targeted CVE check (e.g., of "jar" files).
- Perform targeted manual testing and local scan tools.
- Perform targeted source code review and code analysis.
- Perform targeted configuration review (e.g., Apache and Tomcat).

Document Audit Results

- Document all high- and moderate-impact issues in ESGF Flaw Tracking Database.
- Coordinate the resolution of all high- and moderate-impact issues with ESGF team.
- Document findings and issue report to ESGF as necessary (each site).
- ESGF Executive Committee certifies for release (NOT individual ESGF site).

F.5 ESGF Site Best Practices

All Sites

- Shall adhere, as applicable, to ESGF site best practices.
- Shall contribute recommendations to the ESGF site best practices for consideration and distribution.
- Implement FISMA Controls AC-3, AC-5, AC-6, AC-14, AC-17, AC-22, CM-3, CM-4, CM-6, CM-7,

CM-8, IA-8, IR-5, IR-6, PL-2, PL-8, SA-3, SA-8, SA-10, SA-11, SC-2, SC-5, SC-12, SC-13, SC-32, SI-4, and SI-7.

Installation

- Create and maintain sufficient segregation and isolation of the local ESGF site environment. Where feasible, ESG published datasets should be separate from other data, preferably on storage hardware dedicated to ESG nodes; least privilege shall be used, in regard to permissions on the ESG datasets (i.e., if ESG only requires read access, then only read is granted).
- Implement recommended ESGF site firewall rules, including:
  - Default deny posture, with exceptions allowed for access to the ESG application.
  - Administrative access to the servers hosting the application.
  - Other support functions.
- Ensure that the default password is changed after the ESGF Installer completes.
- Ensure the ESGF software and supporting server operating system environment and supporting elements (e.g., Apache) are maintained and patched.
  - Note/reminder: kernel updates typically require a reboot to take effect.

Monitoring

- Highly recommended: Implement central logging (loghost) of the ESGF environment, including Apache HTTPD and Apache Tomcat logs (ryslog imfile or similar means).
- Highly recommended: Implement the use of the Linux audit capability and auditd daemon to monitor access attempts (Note: a baseline shall be developed for this), with audit logs also forwarded to a central loghost.
- Highly recommended: Implement monitoring (e.g., Nagios) of the ESGF environment and services upon which the ESGF local environment depends.

Specific and actionable guidance shall also be developed in the following areas:

- Access control.
- Patching.
- Configuration management.
- Account management.
- Incident response.
- Security planning.
- System development and testing.
- System and information protection.
- Integrity.

Appendix G. Working Team Accomplishments and Roadmaps

G.1 Compute Working Team

Leads: Daniel Duffy, National Aeronautics and Space Administration (NASA), and Charles Doutriaux, Lawrence Livermore National Laboratory (LLNL)

The amount of data stored across ESGF is expected to grow dramatically with future assessment reports and as future projects utilize ESGF. The availability of more data can be an asset to researchers, but as the data sets grow in size and quantity, users will no longer be able to efficiently download all the data needed for their research. Therefore, a shift is needed from data downloads to remote data analysis within ESGF. In preparation for this shift, the Compute Working Team (CWT) has been developing capabilities to expose data-proximal analytics across ESGF.

CWT's charge is to develop a general application programming interface (API) for exposing ESGF distributed computational resources [e.g., high-performance computing (HPC), clusters, and clouds] to multiple analysis tools. In addition, CWT will codevelop a reference architecture for server-side processing capabilities and use this architecture to test out the API and representative use cases.

G.1.1 2015 Compute Accomplishments

- To drive discussion of analytics, CWT presented some potential scientific use cases. For example, suppose a user wanted to generate a temperature anomaly across Earth's surface for the summer months using data from the Intergovernmental Panel on Climate Change (IPCC). Computing the temperature anomaly requires:
  - Computing the average summer temperature across multiple years for each of the specified collections independently of one another.
  - Computing the average summer temperature for a specific year for each of the specified collections.
  - Regridding all results to the same spatial grid.
  - Computing the ensemble average across all the regridded results.
  - Computing the anomaly by taking the difference between the regridded average and the regridded specific years for each ensemble member.
- CWT analyzed how well different API alternatives would perform under the various use cases before settling on the creation of a web processing service (WPS). WPS APIs are used heavily throughout the open geospatial communities and consist of three basic services (see Fig. 16, this page):
  - GetCapabilities: Provides general information about WPS implementation.
  - DescribeProcess: Provides a full description of an executable WPS process, including required inputs and outputs.
  - Execute: Executes the process and provides the requested output.
- CWT began documenting the WPS API, starting with relatively simple computational processes such as an average function:

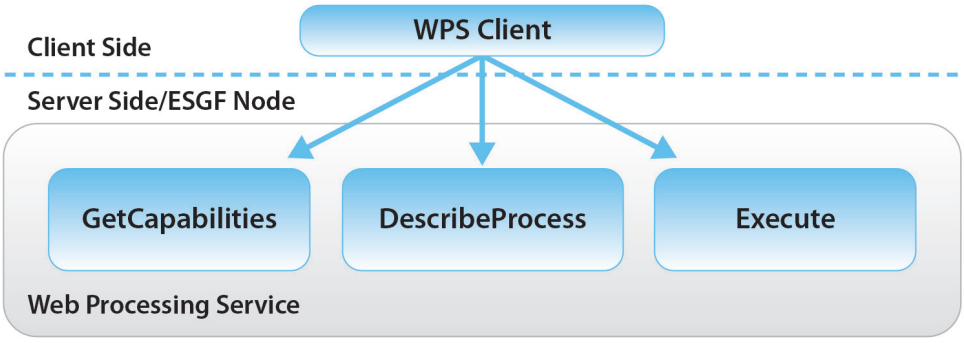


Fig. 16. WPS diagram.

  - Inputs to the average function would require the user to describe the variable, domain, and axes over which to perform the function.
  - Additional inputs could be given to specify any required regridding.
- Charles Doutriaux created a WPS proof of concept at LLNL using PyWPS.
- Tom Maxwell at NASA Goddard Space Flight Center (GSFC) worked closely with CWT to create an ESGF analytics framework.
- Posters based on CWT's efforts were presented at the 2015 American Geophysical Union conference.

G.1.2 Compute 2016 Roadmap

- Expand the point of contact (PoC) to be a production prototype, complete with representative data sets, unit tests, and additional processing capabilities.
- Continue to expand the specification document with additional use cases and update it as necessary.
  - Discussions with scientists are needed to generate additional use cases for CMIP6 data sets.
  - Based on these discussions, CWT will generate a prioritized list of capabilities to expose and begin working on.
- Create unit tests using either standard or synthetic data sets to verify that each ESGF compute node is working as expected.
- Instantiate multiple instances of WPS across the federation.
  - Set up an ESGF-compliant WPS for at least two, and potentially more, sites.
  - LLNL and GSFC plan to be among the first sites.
- Integrate more tightly into ESGF services (e.g., search).
- Analyze multiple back-end implementations to compare the various aspects of implementing and operating different computational platforms. Two such approaches to the storage and compute fabric needed for high-performance data analytics within ESGF are shown in Fig. 17, this page.
  - Climate Data Analysis Service (CDAS): This approach is built on traditional Portable Operating System Interface (POSIX®)-compliant file systems and enables the use of existing tools for data analysis, such as Ultrascale Visualization–Climate Data Analysis Tools (UV-CDAT).
  - Climate Analytics-as-a-Service (CAaaS): This approach is being built on emerging data analysis platforms using object-based (non-POSIX®) storage repositories, such as Hadoop.

G.1.3 Resources Needed to Achieve Compute Goals

- Discussions with the scientific user community to better define and prioritize the capabilities to be exposed (June 2016).

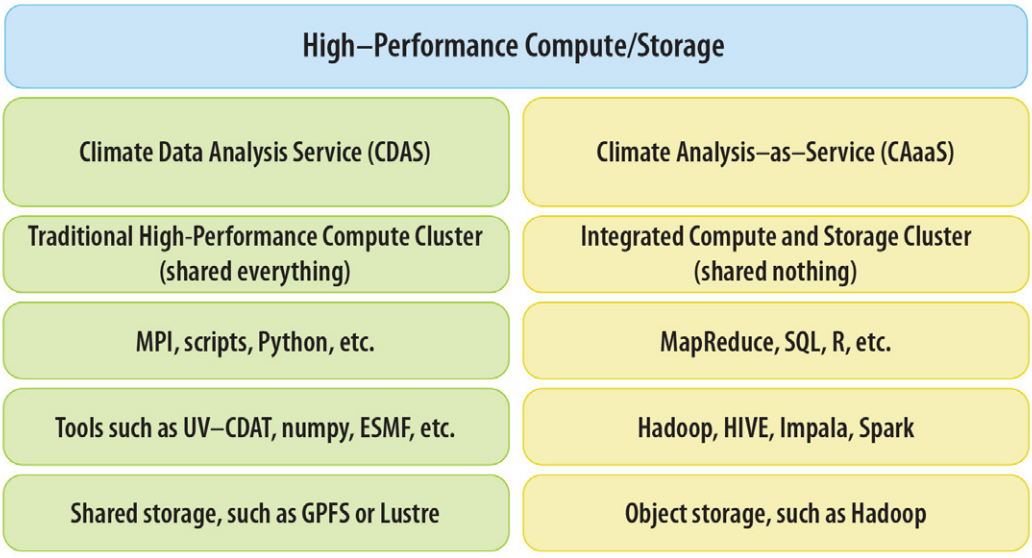


Fig. 17. Comparison of back-end implementations.

- Continued involvement in the development of API and server-side processing routines by the entire CWT (ongoing through December 2016).
- Additional sites (outside LLNL and GSFC) willing to instantiate an ESGF-compliant WPS complete with data and compute capabilities (ongoing through December 2016).

G.2 CoG User Interface Working Team

Leads: Luca Cinquini, NASA and National Oceanic and Atmospheric Administration (NOAA), and Cecelia DeLuca, NOAA

G.2.1 2015 CoG Accomplishments

- Integrated CoG installation with ESGF installer. CoG is now the default node front end together with the rest of the ESGF stack (see Fig. 18, this page).
- Improved and evolved the CoG federation model for exchanging project, user, and tag information across sites.
- Made several improvements and upgrades to the CoG interface for searching ESGF-distributed data (e.g., administrator configuration pages, user search pages, and data cart).
- Integrated with Globus online services for data download (unrestricted data only, for now).
- Improved support for node-specific customization (e.g., header, footer, licenses, and notifications).
- Made several improvements to software infrastructure, including Django upgrade, database migration, security fixes, layouts, styles, and look and feel.

G.2.2 2016 CoG Roadmap

- Support CoG deployment and use throughout the federation (e.g., provide help to administrators and users, February 2016).
- Integrate ESGF ingestion services for publishing smaller data sets from a large community of principal investigators (March 2016).
- Support download of restricted data through Globus services (June 2016).
- Implement Working Group on Coupled Modelling Infrastructure Panel (WIP) requirements in support of CMIP6, including display of quality flags, persistent identifiers (PIDs), (digital object identifiers (DOIs), and errata and integrate these with Earth system documentation (ES-DOC) model metadata (ongoing through December 2016).
- Implement other requirements as they emerge and are prioritized from ESGF administrators and users (ongoing).

G.2.3 Resources Needed to Achieve CoG Goals

- New CoG user group consisting of ESGF administrators (to demonstrate features and provide requirements and feedback).
- Delegation of most support requests to the ESGF support team.
- Model for collaboration with developers of other ESGF services (e.g., ingestion and computing).

Fig. 18. Home pages for currently operational CoG sites at NASA's Jet Propulsion Laboratory (JPL), the University of Colorado, and NOAA's Earth System Research Laboratory (ESRL).

G.3 Dashboard Working Team

Leads: Paola Nassisi and Sandro Fiore, Centro Euro-Mediterraneo sui Cambiamenti Climatici (CMCC)

G.3.1 2015 Dashboard Accomplishments

- Completed implementation of the first data usage statistics system (coarse grained), relying on the existing access logging system.
- Provided the following set of metrics to CMIP5, Coordinated Regional Downscaling Experiment (CORDEX), Observations for Model Intercomparisons (Obs4MIPS), and all projects:
  - **Download statistics** — Data downloaded (in gigabytes and terabytes), number of downloads, number of distinct files, number of distinct users, downloads by user, and downloads by identity provider.
  - **Client statistics** — Geographic map, country and continent distribution, and identity provider (IdP) distribution.
- Designed and implemented preliminary version of the **fine-grained** system, which relies on a new data warehouse for tracking all downloads and a set of data marts for effectively delivering aggregated statistics to the application layer (e.g., ESGF desktop).
  - The fine-grained system will provide statistics starting in 2016, and the coarse-grained system will continue providing aggregated statistics from the past (see Fig. 19, this page).
- Developed a new set of views for coarse-grained statistics including download statistics, download by IdP and by user activity, client statistics, and distribution.
- Developed an initial and new set of cross-project and project-specific views for the fine-grained statistics. A view for CMIP5 has been implemented with project-specific statistics for downloads by variables, model, data set name, and realm (see Fig. 20, p. 89).

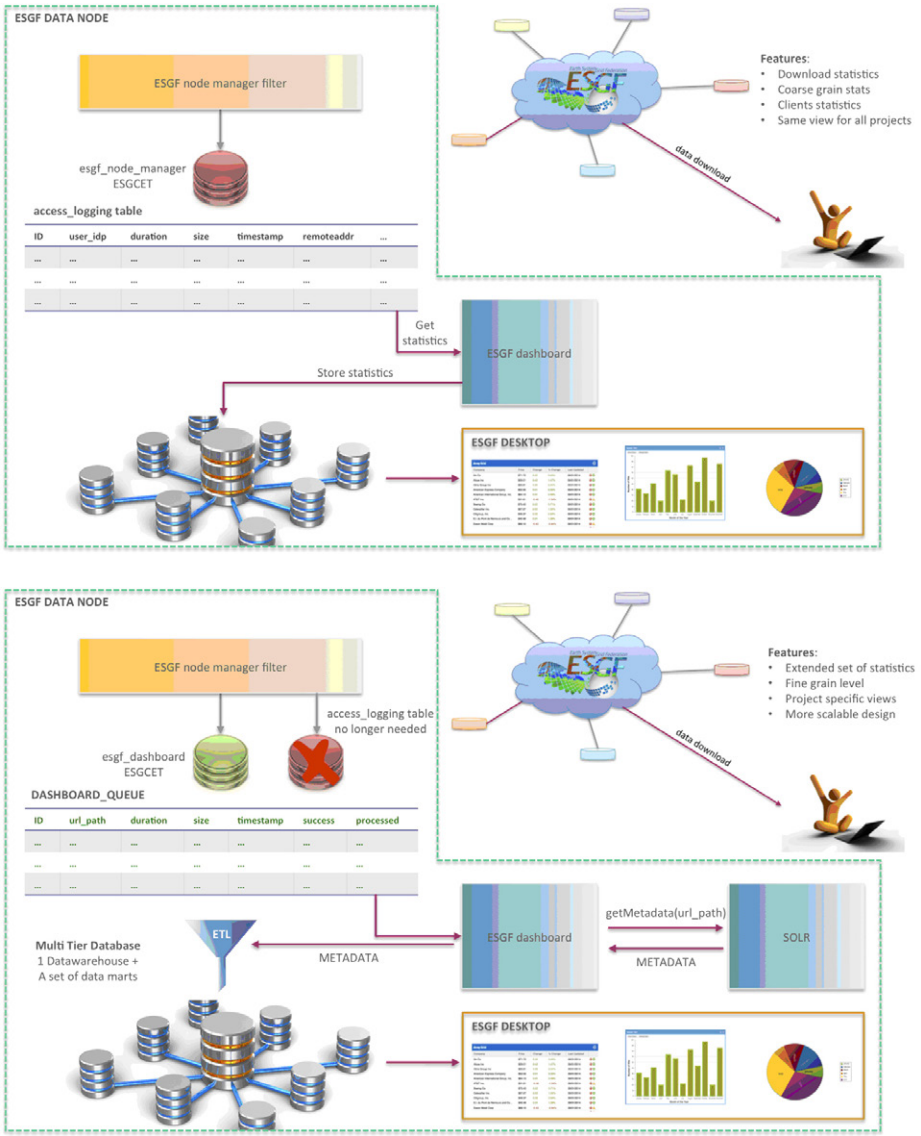


Fig. 19. Architectures for coarse- and fine-grained statistics for the ESGF dashboard component.

- Made several improvements and upgrades to the dashboard's security aspects, look and feel, and layouts.

G.3.2 2016 Dashboard Roadmap

- Implement the back-end interaction with Solr for the fine-grained statistics.
  - Integration of statistics for CMIP5, CORDEX, and Obs4MIPs (February 2016).
- Define new views in the context of the ESGF Dashboard Working Team (2016):
  - In particular, starting from Obs4MIPs, CMIP5, and CORDEX.
  - Single site and federated.
  - Project-specific and cross-project.
- Implement representational state transfer (REST) APIs:
  - Single-node level and federation level (March 2016).
- Extend set of views with geolocation and federation-level statistics (May 2016).
- Ensure close interaction with the Node Manager, Network, and Search Working Teams (2016).
- Document the back-end configuration and architecture (2016).
- Develop the new front-end presentation layer (ongoing through final release in December 2016).

G.3.3 Resources Needed to Achieve Dashboard Goals

- Collaboration with specific projects and related members to define new views and incorporate project-specific requirements.
- Collaboration with developers of other ESGF services to better address statistics needs.

G.4 Data Transfer Working Team

Leads: Rachana Ananthakrishnan and Luckasz Lacinski, University of Chicago

G.4.1 2015 Data Transfer Accomplishments

Fig. 20. Fine-grained statistics. Project-specific view for the CMIP5 project.

- Moved all custom modifications of the Globus mainstream packages to the authorization callout libraries.
- Added support for Globus sharing.
- Supported the CoG User Interface Team in integrating Globus with CoG for data downloads.
- Built the authorization callout libraries as native packages [Red Hat Package Managers (RPMs)] for RHEL 6 and CentOS 6.
- Retired the Bulk Data Movement GridFTP.
- Moved from installing Globus and related software from source code to RPMs. Updated Globus from 5.0.4 to 6, the latest version. Started using Globus Connect Server.

G.4.2 2016 Data Transfer Roadmap

- Integrate Globus with replication tools (April 2016).
- Implement the data transfer node (DTN) and add the new node type to the installer. The DTN hosts only Globus Connect Server input/output (I/O) with a custom authorization callout and supports Science DMZ architecture for replication and downloads (April 2016).
- Implement Globus download in CoG for restricted data (June 2016).
- Perform updates to support data format transformation [e.g., data in Network Common Data Form 4 (NetCDF4) but downloaded as NetCDF3] (October 2016).
- Develop a better delegation model to obtain ESGF transfer certificates (December 2016).

G.4.3 Resources Needed to Achieve Data Transfer Goals

- Collaboration with the International Climate Network Working Group (ICNWG) and replication tool developers.
- Collaboration with CoG developers to integrate Globus download solutions.

G.5 Identity, Entitlement, and Access Management (IdEA) Working Team

Leads: Philip Kershaw, National Centre for Atmospheric Science/British Atmospheric Data Centre (BADC), and Rachana Ananthakrishnan, University of Chicago

G.5.1 2015 IdEA Accomplishments

- Integrated a new user delegation service with ESGF partner services, including the Royal Netherlands Meteorological Institute's (KNMI) Impacts Portal and preliminary work with the German Climate Computing Centre's (DKRZ) Birdhouse WPS suite. The delegation service allows third-party ESGF services to obtain credentials on behalf of a user and access secure data sets. For example, the Impacts Portal is able to load CMIP5 data sets on behalf of the user and visualize them in a web map service (WMS) client. The delegation service also provides a complete replacement to MyProxyCA (see Fig. 21, p. 91).
- Performed upgrades and maintenance to ESGF IdEA software components to patch for security vulnerabilities.

G.5.2 2016 IdEA Roadmap

- **Step 1:** Perform pilot integration of Globus with ESGF OAuth 2.0 service deployed at CEDA. This integration will enable full GridFTP data download with ESGF access control (March 2016).
- **Step 2:** Perform pilot integration of live access service (LAS) with ESGF OAuth 2.0 service deployed at CEDA. This integration will enable LAS to obtain delegated user credentials to access secured data from ESGF services (April 2016).
- **Step 3:** Implement service discovery mechanism for OAuth 2.0. This is an important milestone in deprecating OpenID 2.0 and replacing it with OAuth 2.0 (May 2016).
- **Step 4:** Deploy OAuth 2.0 operationally with ESGF IdPs (July 2016).
- **Step 5:** Retire OpenID 2.0 and replace it with OAuth 2.0 service (September 2016).
- **Step 6:** Retire MyProxyCA; an existing hypertext transfer protocol (HTTP)-based Short-Lived Credential Service (SLCS) replaces it (October 2016).
- **Step 7:** Implement and integrate OpenID Connect into ESGF. OpenID Connect is a direct replacement for OpenID 2.0. It builds on the OAuth 2.0 solution in steps in 2 and 3 (ongoing to December 2016).
- **Step 8:** Implement a security assertion markup language (SAML) bridge to ESGF IdP. This bridge will enable user sign-in using institutional credentials for users whose home organizations are part of Shibboleth federations (e.g., InCommon).

G.5.3 Resources Needed to Achieve IdEA Goals

- Operations and integration effort for OAuth 2.0 integration (steps 1–6).
- Software development effort for OpenID Connect implementation and integration (step 6).
- Software development effort to implement SAML bridge to ESGF IdP (step 7).
- Technical management oversight, which is a critical aspect given the complexity of this work domain area.

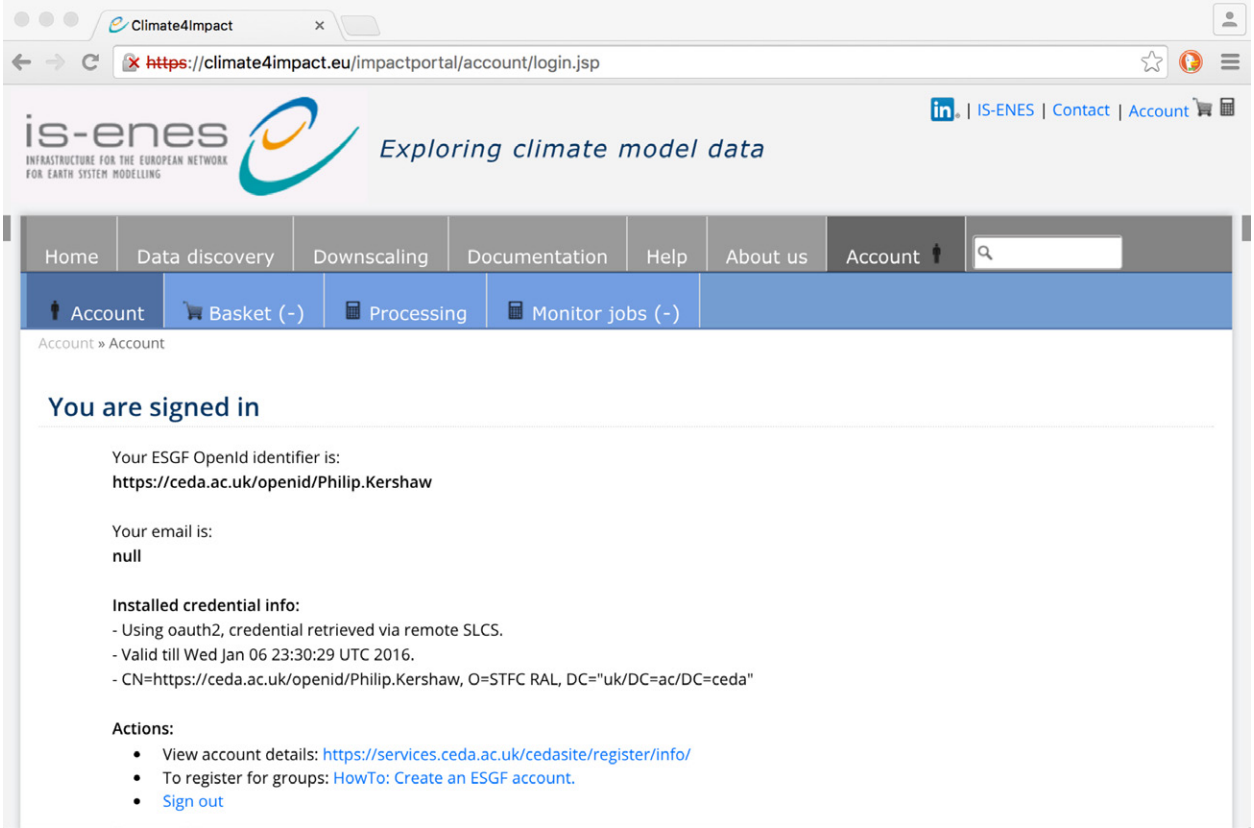


Fig. 21. KNMI Impacts Portal. The screenshot shows the sign-in for Centre for Environmental Data Analysis (CEDA) user Philip Kershaw with CEDA's OAuth 2.0 service. *Here, the sign-in process is complete displaying information about the delegated user certificate the portal has just obtained for Kershaw.*

Achievement of these goals is dependent on operations and development efforts allocated to these tasks. A baseline is achievable for *partial* completion of steps 1–6. With modest additional effort, these steps can all be achieved. Steps 7 and 8 require a dedicated development effort allocated to ensure success.

Steps 7 and 8 are important to ensure that the ESGF system remains current with the latest technologies used by the industry and research sectors. Without ongoing development, IdEA components risk becoming incompatible with comparable systems in the research community or difficult to securely maintain.

G.6 Installation Working Team

Leads: Prashanth Dwarakanath, National Supercomputer Centre (NSC), and Nicolas Carenton, European Network of Earth System Modelling (ENES)/Institut Pierre Simon Laplace (IPSL)

G.6.1 2015 Installation Accomplishments

- Overhauled certificate authority (CA) and trust setup:
  - Replaced source-built packages with distribution-provided packages to improve security and ease maintenance.
  - Set up new CAs with better security features at IPSL and NSC.
  - Eliminated redundant CAs and duplicate hashes from the ESGF trust store.
  - Integrated an all-new temporary CA setup with an installer to allow standalone testing of services upon installation.
- Enhanced multiplatform server-side support:
  - Set up Apache front end, performing URL-based proxying.
  - Supported both Java and Python server-side code.
  - Enabled CoG integration into the installer.

- Implemented automated installation and quicker installations:
  - Set up autoinstall to totally automate installations.
  - Reduces chances of human error due to stray keystrokes during the install process.
  - Supports "all," "data," and data with compute roles when performing clean installations.
  - With good network connectivity and access to a nearby mirror, a complete installation can now be completed in less than 15 minutes.
- Made security and component upgrades:
  - Migrated from Java7 to Java8.
  - Migrated from Tomcat6 to Tomcat8.
  - Included a common vulnerabilities and exposure (CVE) checker utility with the ESGF release to help scan for vulnerabilities.
- Managed multiple releases (see Fig. 22, this page):
  - 1.8.0, February 27, 2015 — Routine upgrade.
  - 1.8.1, March 11, 2015 — Certificate-less Wget downloads, sslv3 fix.
  - 2.0.1, October 2, 2015 — First release subjected to a security audit following the security incident of June 2015.
  - 2.1.0, December 9, 2015 — Stable release of ESGF, incorporating fixes to all known vulnerabilities.

G.6.2 2016 Installation Roadmap

- Integrate the new node manager with the ESGF installer (June 2016).
- Replace the bash installer with a modular Python installer (November 2016).

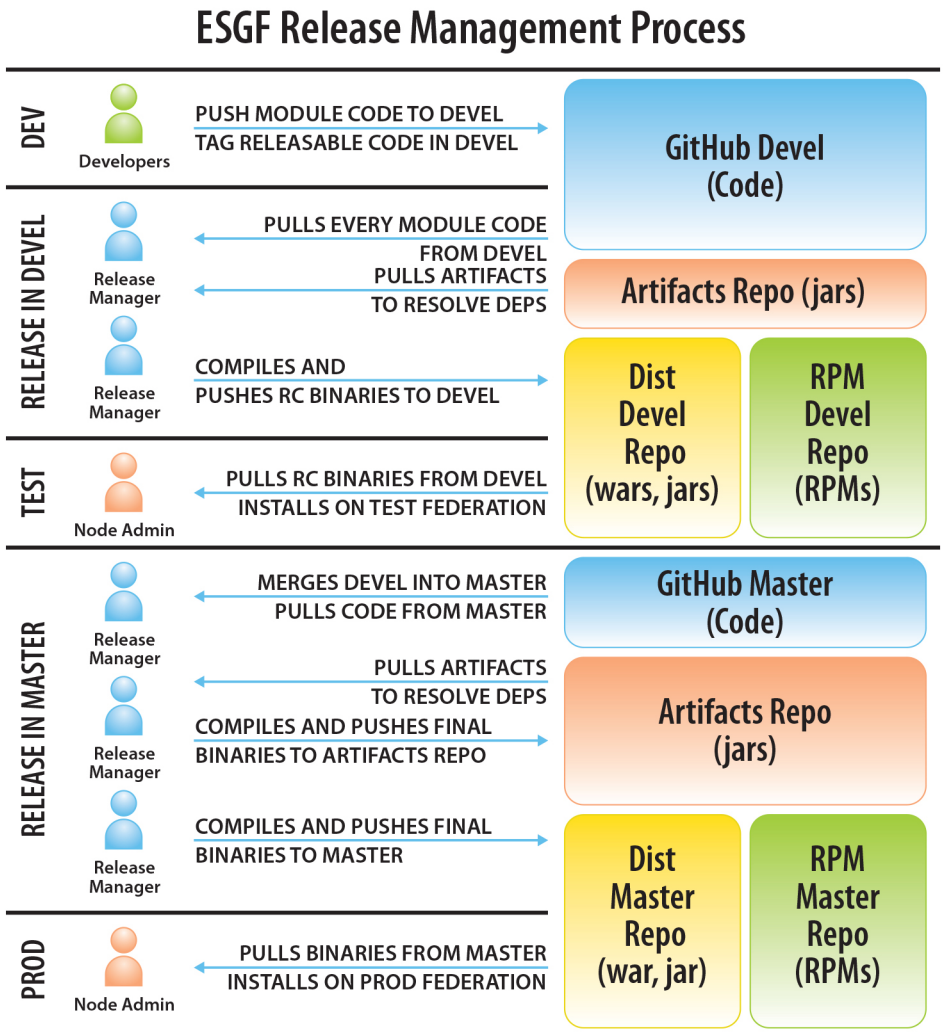


Fig. 22. Release Management Process. *Development of the ESGF software stack adheres to a release management process that ensures the quality of deliverables. Three distinct roles are identified: (1) developers push new features into the system; (2) a release manager is responsible for code freeze, cutting releases, and compilation; and (3) node administrators are requested to test and validate release candidates and the operating ESGF version in production.*

- Create modules and recipes for individual components that could then be maintained by their respective developers (September 2016).

G.6.3 Resources Needed to Achieve Installation Goals

- No additional resources are needed at this time.

G.7 International Climate Network Working Group

Leads: Eli Dart and Mary Hester, DOE Energy Sciences Network (ESnet)

G.7.1 2015 ICNWG Accomplishments

- Created three test data sets composed of real climate data:
  - Each data set is about 240 GB.
  - Made them available on ESnet DTNs for testing.
- Assessed data transfer performance using test data sets:
  - GridFTP and Globus Online performance improved (100 times better than rsync).
  - Characterized host, network, and file system performance.

G.7.2 2016 ICNWG Roadmap

- Collaborate with the Replication and Versioning Working Team on the Synda pilot (ongoing through December 2016).
- Assist with DTN setup at other data centers (ongoing through December 2016; see Fig. 23, this page).

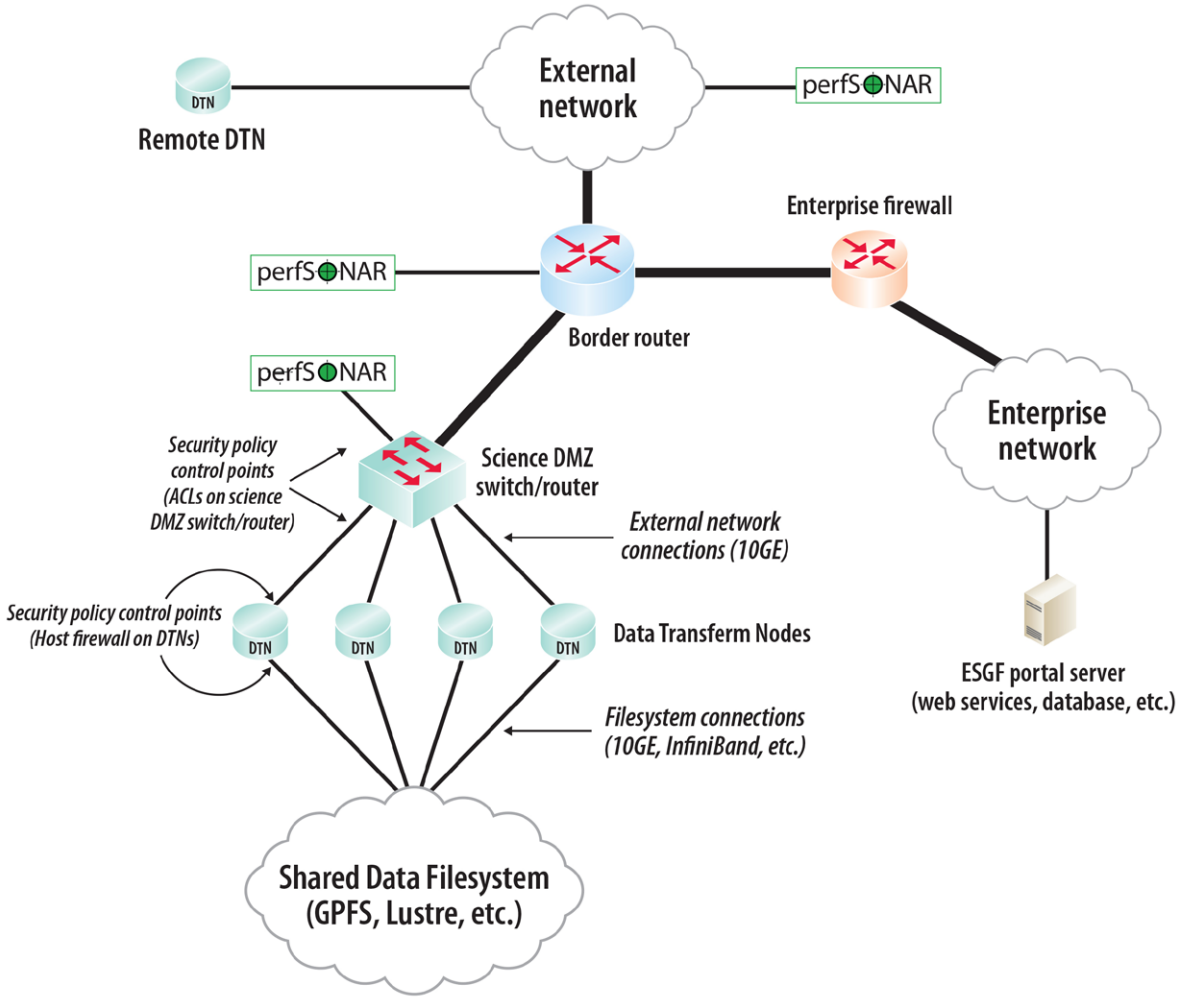


Fig. 23. Separation of ESGF DTNs from the rest of the ESGF portal installation.

- Collaborate with the Data Transfer Working Team to move data more efficiently (ongoing through December 2016).
- Collaborate with the Dashboard Working Team to integrate network performance into the ESGF desktop and dashboard (June 2016).

G.7.3 Resources Needed to Achieve ICNWG Goals

- DTNs established at Tier 1 data centers: CEDA, DKRZ, LLNL, and the National Computational Infrastructure (NCI).
- Republished CMIP5 data with GridFTP URLs.
- Staff time.

G.8 Metadata and Search Working Team

Lead: Luca Cinquini, NASA/NOAA

G.8.1 2015 Metadata and Search Accomplishments

- Upgraded the Solr installation to the latest version (5.2.1), enabling atomic metadata updates, geospatial searches, and better performance and scalability.
- Made several infrastructure improvements, including running "slave" Solr on standard port 80 (instead of port 8983) to avoid firewall issues, use of a Jetty container packaged with Solr distribution, and use of standard Solr scripts for starting and stopping the services.
- Introduced "local shard" for publishing and distributing data that does not need to be shared with the rest of the federation (i.e., data important only to a local user community).
- Made several improvements to the search user interface (UI) as part of CoG development (see Fig. 24, this page).
- Implemented several security fixes to comply with CVE warnings.
- Reviewed, upgraded, and documented ESGF RESTful publishing services.

G.8.2 2016 Metadata and Search Roadmap

- Support deployment of ESGF publishing and search services across the federation (February 2016).
- Develop tools and services to support atomic metadata updates (March 2016).
- Support tagging of data sets for multiple projects (March 2016).

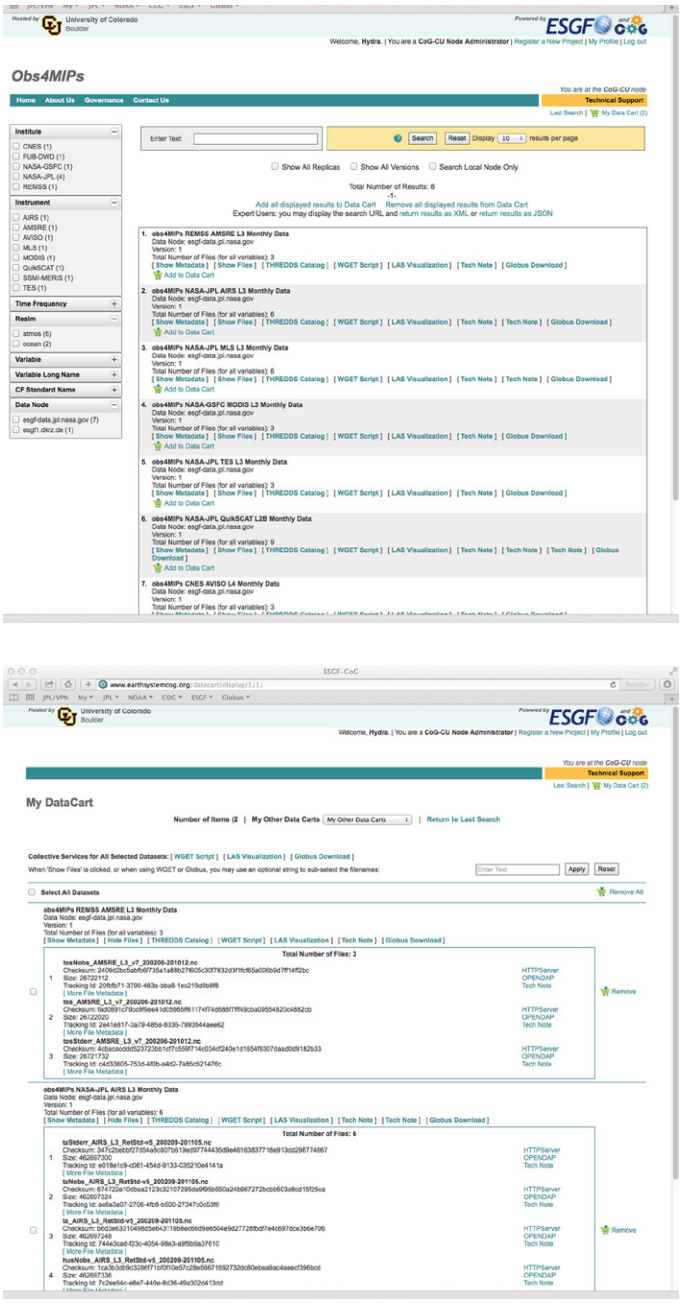


Fig. 24. User interface pages for search results and data cart retrieve information from the ESGF search services.

- Revise and improve documentation (especially the RESTful ESGF query syntax; March 2016).
- Implement data validation against controlled vocabularies (June 2016).
- Make continuous upgrades to the Solr distribution, including defining a process for migrating the metadata indexes (ongoing).
- Implement and prioritize other requirements to search back end and front end (CoG) as they emerge from CMIP6 and other projects (ongoing).
- Support partitioning of search space across multiple virtual organizations [e.g., ESGF and Accelerated Climate Modeling for Energy (ACME); optional, if time permits].
- Research Solr-Cloud architecture (optional, if time permits).

G.8.3 Resources needed to Achieve Metadata and Search Goals

- Extension of working team participation, especially for:
  - Assuming responsibility for metadata standards and validation.
  - Establishing an infrastructure to monitor the consistency of search results across the federation.
- Stabilization of a concrete and usable implementation of controlled vocabularies as soon as possible.
- Establishment of a controlling body that has authority over which data sets can be published into the common search space.

G.9 Node Manager, Tracking, and Feedback Working Team

Leads: Alexander Ames, LLNL, and Prashanth Dwarakanath, NSC

G.9.1 2015 Node Manager, Tracking, and Feedback Accomplishments

- Implemented a software design based on a two-tier architecture.
- Launched the version 0 module featuring health check communications, second-tier node addition and removal, configuration file deployment, and registration.xml with basic metrics.

G.9.2 2016 Node Manager, Tracking, and Feedback Roadmap

- Node manager (see Fig. 25, this page):

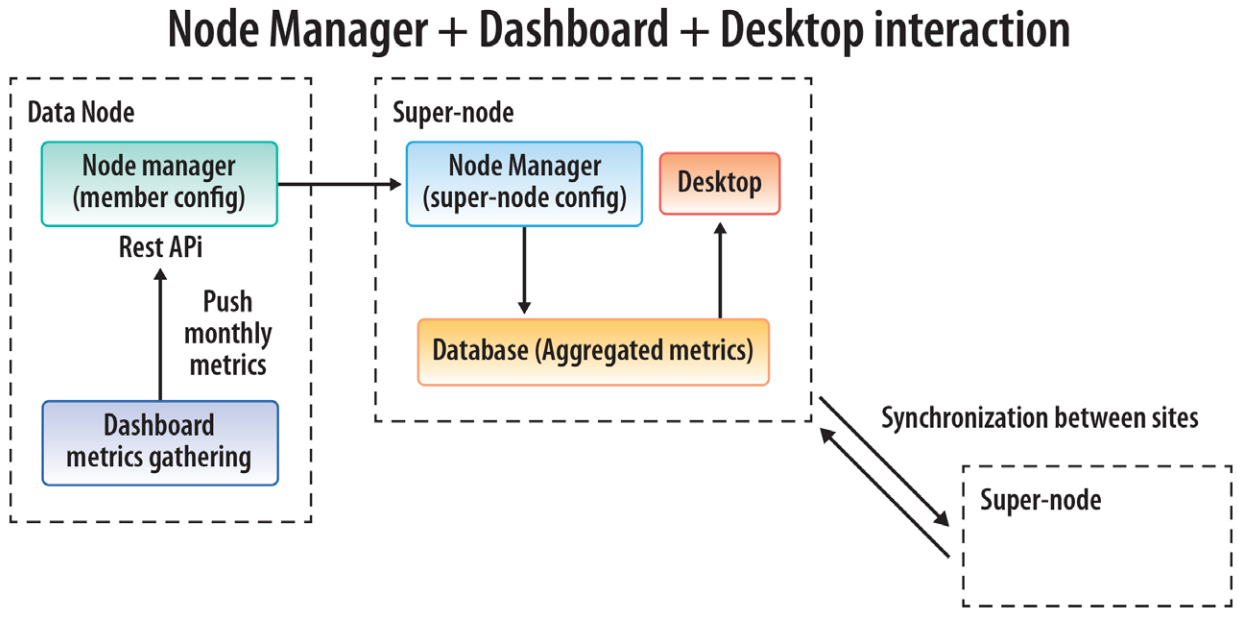


Fig. 25. Software architecture of node manager integration with the ESGF dashboard and desktop.

  - Finish v.0 implementation — Major required features include shards files, security, dynamic super-node selection, and integration with installer (January 2016).
  - Implement feature and requirements for v.1 — Integration with next dashboard, other components, and standby mode (February 2016).
  - Implement and test deployment of v.1 (July 2016).
  - Release production v.0 — Target mid-year major release (to be determined).
  - Release v.1 to production — Target late-year release (to be determined).
- Tracking and feedback:
  - Assemble team (January 2016).
  - Design and review (March 2016).
  - Implement testing v.0 (June 2016).
  - Production release (to be determined)

G.9.3 Resources Needed to Node Manager, Tracking, and Feedback Achieve Goals

- Team members for tracking and feedback.
- Coordination with ESGF node administrators for node manager testing.
- Coordination with Search, CoG, and IdEA Working Teams on how the node manager can support their components.
- Tracking and feedback requirements from science project leads.

G.10 Persistent Identifier Services Working Team

Leads: Tobias Weigel and Stephan Kindermann, DKRZ

G.10.1 2015 PID Accomplishments

- Compiled and finalized a PID services WIP paper with basic agreements:
  - General design of an operational architecture focused on high scalability and availability (see Fig. 26, this page).
  - First high-level agreements on interaction with other ESGF services such as errata and citation.
  - First description (norms) of Handle record contents.
- Developed an initial version of the Python API for publisher integration and Handle service endpoint.

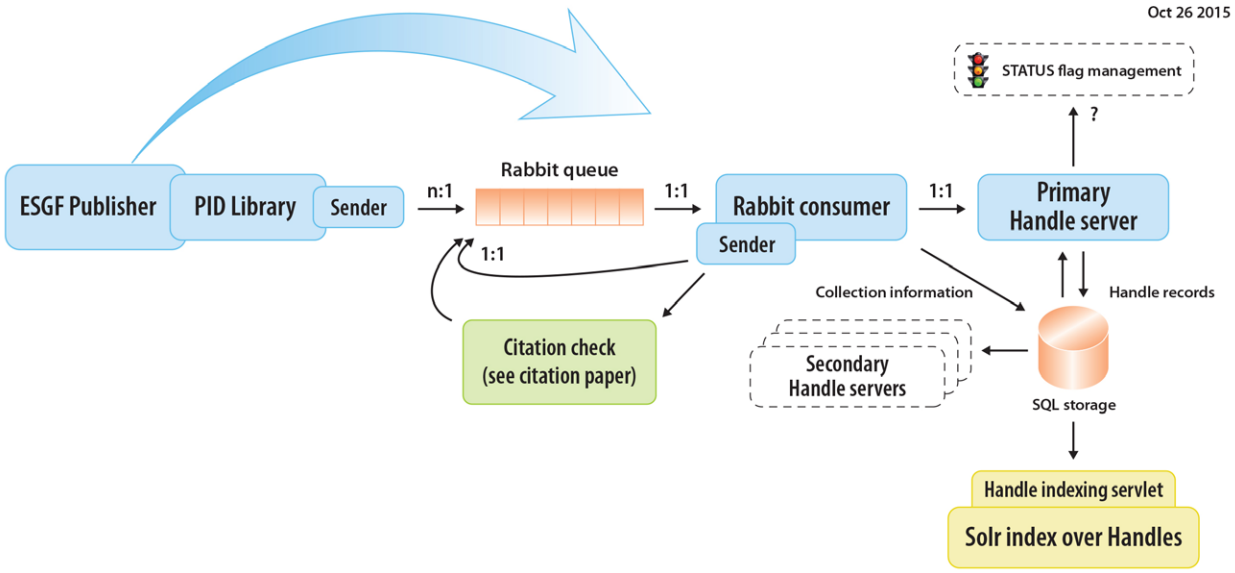


Fig. 26. Basic architecture of ESGF PID services with queueing system and Handle server.

G.10.2 2016 PID Roadmap

- Integrate fully with publisher, errata service, and ESGF end-user services (April 2016).
- Set up operations for the RabbitMQ queuing system and Handle service, including a final decision on the high-availability strategy (June 2016).
- Test software and scalability and perform final roll-out (July 2016).
- Develop additional tools as required, such as an offline message publication tool and a basic end-user information tool (August 2016).

G.10.3 Resources Needed to Achieve PID Goals

- External high-availability hosting of the central RabbitMQ exchange for PID operations, independent from a single ESGF node to reduce operative risks. This may be done through a commercial hosting service, an academic cloud solution, or a cluster across multiple data centers.

G.11 Provenance Capture Working Team

Lead: Bibi Raju, Pacific Northwest National Laboratory

G.11.1 2015 Provenance Accomplishments

- Developed a provenance capture ontology.
- Built a provenance infrastructure (see Fig. 27, this page).
- Developed a scalable provenance capture mechanism.
- Developed a client API that aids in provenance production.

G.11.2 2016 Provenance Roadmap

- Incorporate a time-series system environment metrics store (February 2016).
- Add a provenance capture mechanism that can handle high-velocity provenance information (ongoing through February 2016).

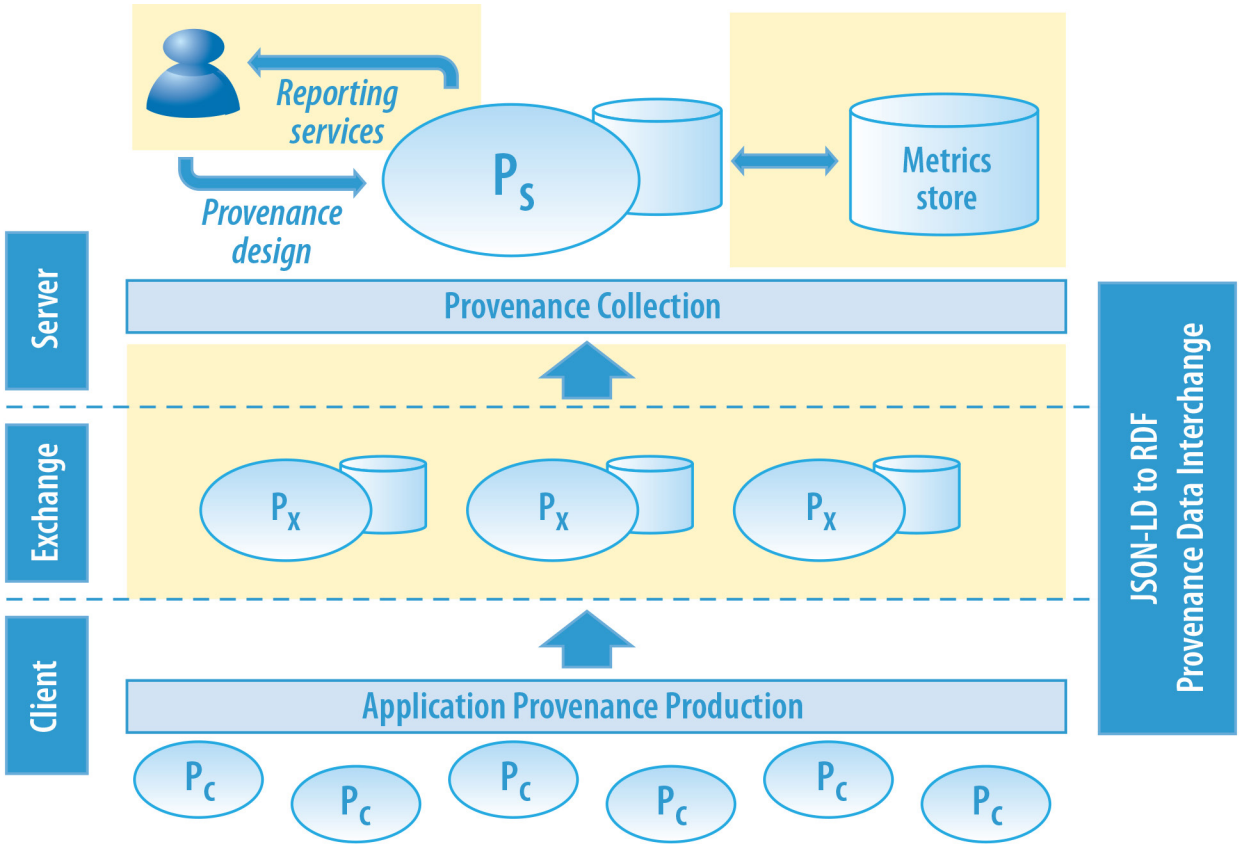


Fig. 27. Provenance environment architecture.

- Develop different language bindings for the ProvEn client API (May 2016).
- Gather requirements for capturing existing sources of provenance information (March 2016).
- Develop a user interface for performance metrics reporting (March 2016).
- Design services supporting provenance harvesting from native source types (May 2016).

G.11.3 Resources Needed to Achieve Provenance Goals

- No additional resources are needed at this time.

G.12 Publication Working Team

Lead: Alexander Ames, LLNL

G.12.1 2015 Publication Accomplishments

- Updated to esg-publisher included in ESGF 2.1.x:
  - Compatible with changes that are part of the 2.x stack [e.g., Thematic Real-time Environmental Distributed Data Services (THREDDS) recheck and open secure sockets layer (SSL) v.1.01].
  - Parallel checksums for map creation.
  - Better support to ingest version tags (these are now part of the mapfile).
  - New script to add optional facets to published data sets.
  - Heterogeneous data support within a project (for ACME; i.e., non-NetCDF files).
  - Cleanup of default settings within the template file for esg.ini.
- Released the Globus/ESGF ingestion graphical user interface (GUI) web service. This service has been used for publishing Atmospheric Model Intercomparison Project (AMIP) data for ACME. It supports both asynchronous ESGF publication and data set transfer via Globus.
- Initial *alpha* test version of the Globus ingestion service API. This API has been tested with *Pegasus Workflow* as a client at Oak Ridge National Laboratory (ORNL).
- Designed the Publication Test Suite at CEDA/BADC.
- Made workflow discussion and documentation recommendations based on the Paris code sprint.

G.12.2 2016 Publication Roadmap

- Publication tool and workflow:
  - Develop a new esgscan_directory tool for mapfile generation (January 2016).
  - Implement 2015 recommended changes to esg-publisher scripts (February 2016).
  - Implement schema changes to support publisher integration with errata and PID services (March 2016).
  - Implement the test suite (March 2016).
  - Implement changes to support new THREDDS Data Server (TDS) features, DTNs, and high-performance storage systems (HPSS; March 2016).
  - Develop a new drs_lite tool for versioning and data reference syntax (DRS) management (April 2016).
  - Write a best practices document (April 2016).
  - Prepare for CMIP6 publication (to be determined; need data availability information).
- Ingestion service (see Fig. 28, p. 99):
  - Integrate UI publication as a CoG feature (March 2016).
  - Facet management and esg.ini files (April 2016).
  - Add replicated data publication as a first-class flow (June 2016).
  - Data set versioning (July 2016).

G.12.3 Resources Needed to Achieve Publication Goals

- Communication with Search Working Team on changes to the index node publishing service.
- CMIP WIP white papers and follow-up coordination with the ESGF executive committee.
- Interface with TDS changes.

G.13 Quality Control Working Team

Leads: Katharina Berger, DKRZ; Guillaume Levavasseur, Infrastructure-ENES/IPSL; and Martina Stockhause, DKRZ

G.13.1 2015 Quality Control Accomplishments

- Version support and persistent metadata:
  - Concept development at the ESGF publication sprint in Paris.
  - Name of version directory used as the default version for ESGF publication.
- Errata service:
  - Concept development at the ESGF publication sprint in Paris.
  - WIP paper review to fix errata service architecture and ESGF dependencies.
  - Exploiting the PID (un)/publication chains and standardized issues information, the errata service records and tracks reasons for data set version changes.
- Data citation service (see Fig. 29, p. 100):
  - Concept development and description in a WIP paper.
  - Repository setup.
  - Start of implementation — GUI development for data creators, export and display of citation information ("landing page"), and integration of "data citation" line template in the ESGF CoG portal.
- Citation link template in the ESGF CoG portal.

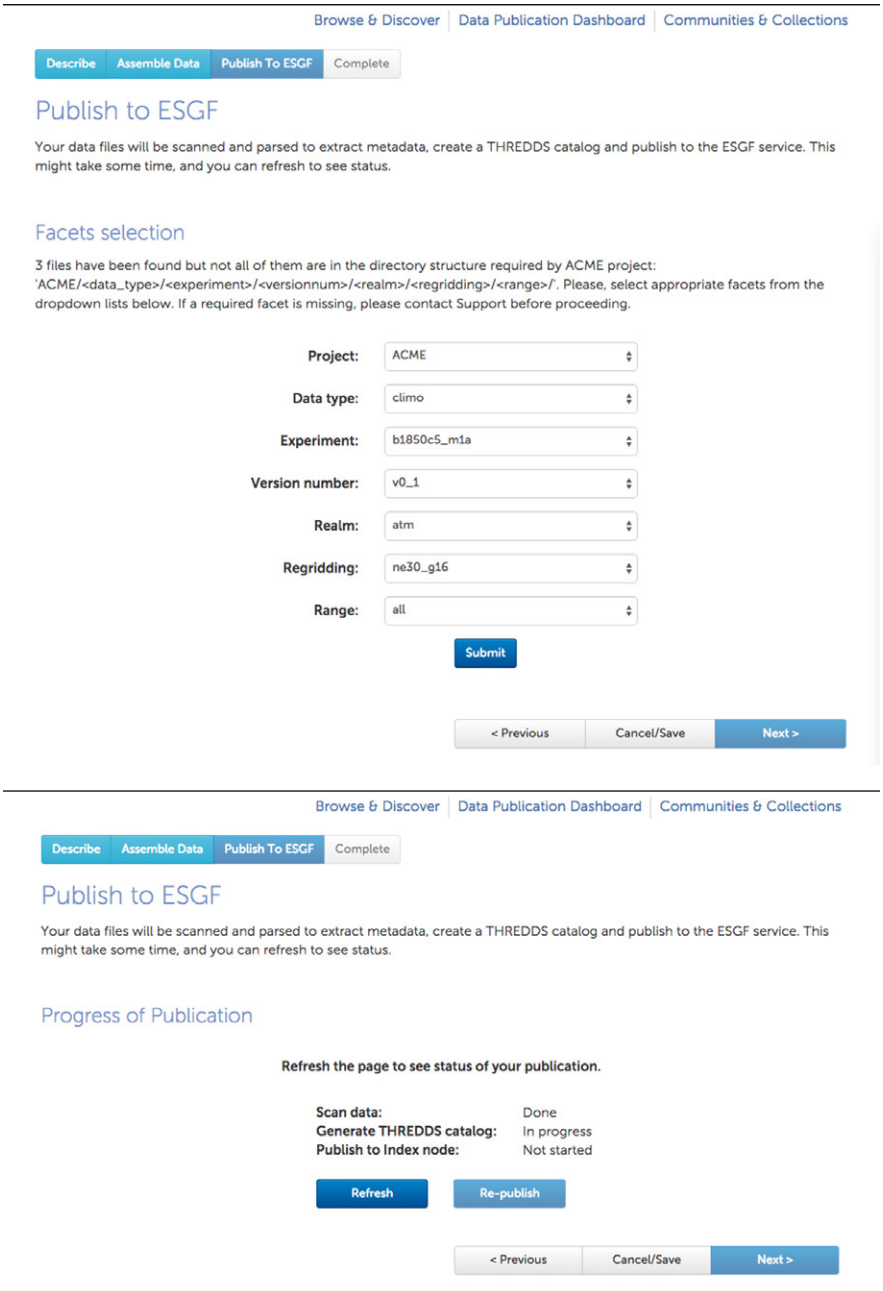


Fig. 28. The ingestion service (top) allows users to select the desired facet values for publication and (bottom) monitors the status of ESGF publication.

G.13.2 2016 Quality Control Roadmap

- Version support and persistent metadata:
  - Continue implementation (with the ESGF Replication and Versioning Working Team).
  - Implement version history service.
- Errata service:
  - Develop and describe the concept in a WIP paper (in review; December 2015).

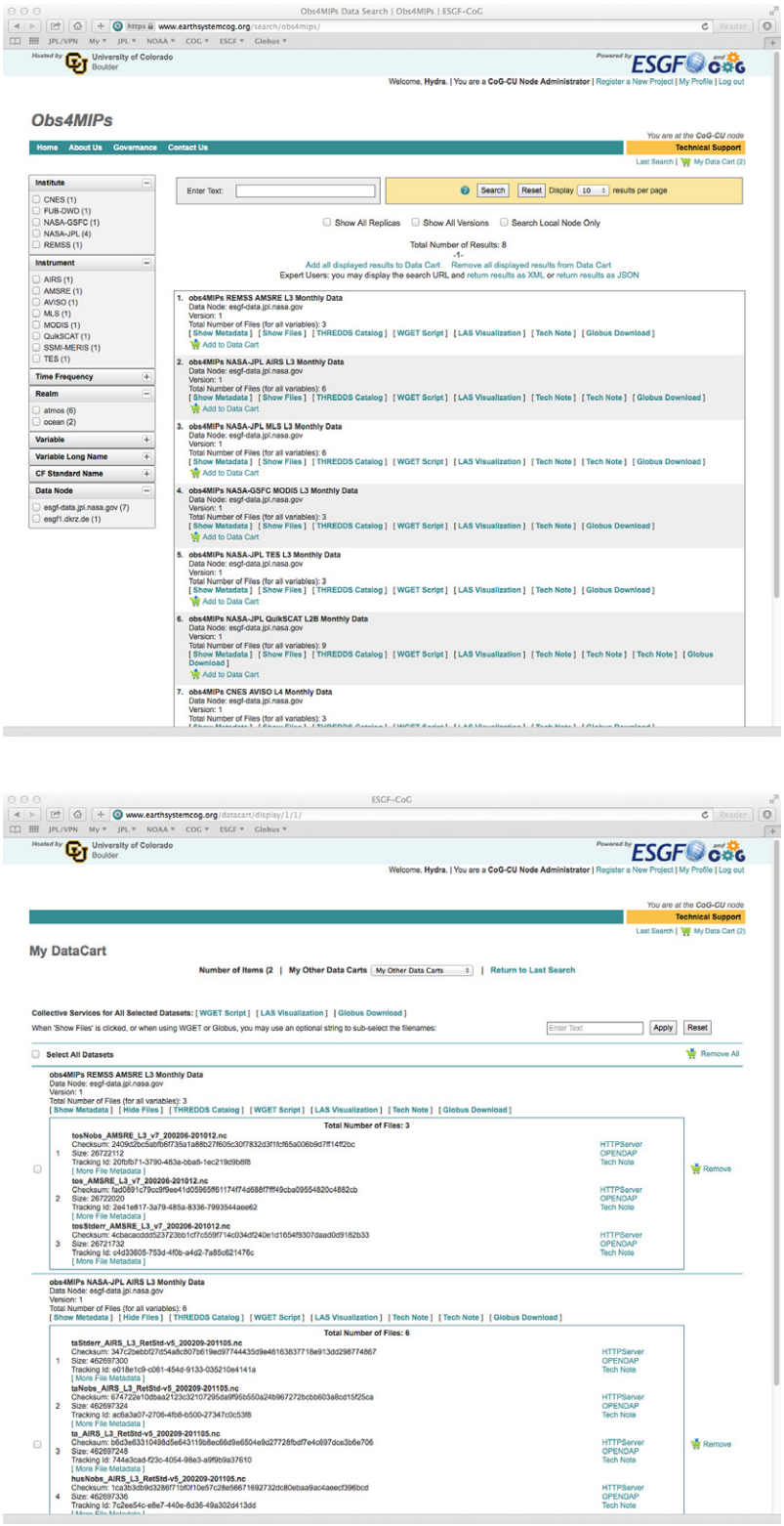


Fig. 29. Citation information in the CoG portal (left) and a landing page with CMIP5 test data (right).

  - Discuss and finalize the issue registration process and form (February 2016).
  - Develop:
    - » The issue manager (June 2016).
    - » The errata module for the ESGF CoG front end (June 2016).
    - » APIs to request Handle service and issue manager (June 2016).
  - Deploy on ESGF index nodes (June 2016).
  - Ensure full operability (September 2016).
- Data citation service:
  - Release the citation GUI for data creators (January 2016).
  - Discuss ESGF version support with the CoG/search API (March 2016).
  - Discuss and define interfaces with ESGF and ES-DOC (March 2016).
  - Discuss and finalize the landing page design and content (March 2016).
  - Provide citation XML for CMIP6 on the open archives initiative (OAI) server (June 2016).
  - Start the integration of DataCite DOI and long-term archival (LTA) services (September 2016).
- General:
  - Draft recommendations for the ESGF integration of external information (June 2016).
  - Prepare the final version of recommendations for ESGF integration of external information (September 2016).

G.13.3 Resources Needed to Achieve Quality Control Goals

Data citation is a critical requirement on version support (evidence use case):
Operable CoG/search for data sets ≤ version/access date and display of metadata for unpublished data sets.

- Collaboration with the ESGF Replication and Versioning Working Team on data versioning and persistence of metadata.
- Collaboration with CoG and search API on version support in the front end.
- Collaboration with the ESGF Publication Working Team and PID service (PIS) for integrating the errata service into the publication workflow and PIS.
- Collaboration with ES-DOC and the CMIP6 quality control (QC) teams on integrating the Common Information Model (CIM), simulation, and QC information into ESGF and drafting recommendations and best practices.

G.14 Replication and Versioning Working Team

Leads: Stephan Kindermann and Tobias Weigel, DKRZ

G.14.1 2015 Replication and Versioning Accomplishments

- Completed the replication and versioning WIP paper (see Fig. 30, this page).
- Completed the CMIP6 versioning requirements and solution approaches living document. (docs.google.com/document/d/1tOaFQEXFyjqAOOlvc-

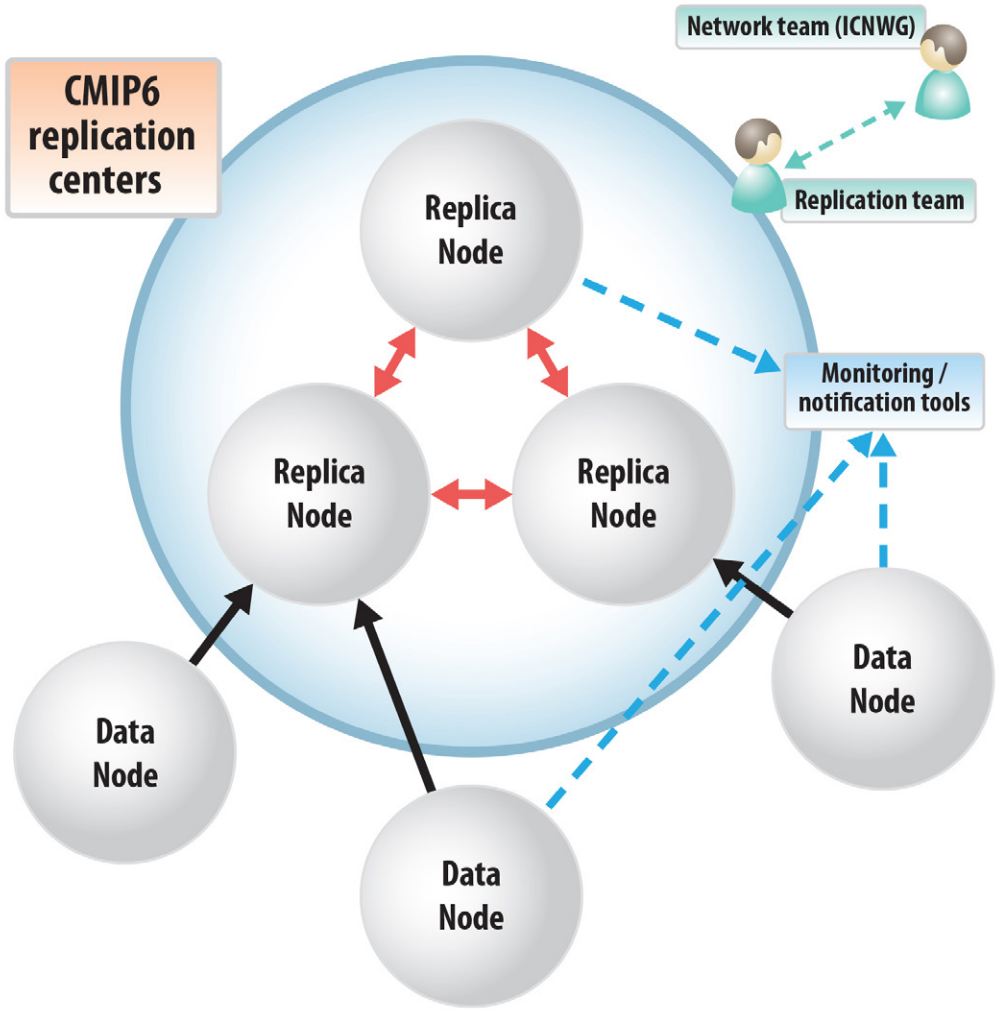


Fig. 30. Overall replication structure after publication data are replicated to core replication nodes, which synchronize their holdings or parts thereof. *A replication team working closely with the network team coordinates the replication process.*

diaX3XrXuxIv_nlE5_FCme1id4/editdocs.google.com/document/d/1tOaFQEXFyjqAOOlvcdiaX3XrXuxIv_nlE5_FCme1id4/edit).

- Improved the Synda replication tool:
  - Installation procedure.
  - Support for GridFTP data transfer (besides http).
- Cooperated with the Publication Working Team on versioning support improvement in the ESGF publisher component:
  - Version information support in mapfile.
- Made PID-related developments:
  - PID services WIP paper with basic agreements.
  - Interaction with errata services.
  - Initial version of Python API for publisher integration and Handle service endpoint.

G.14.2 2016 Replication and Versioning Roadmap

- Cooperate with the Data Transfer Working Group to create the roadmap and perform tests of the Synda replication tool integration with Globus (March 2016).
- Implement the Synda replication test bed between Tier 1 sites (BADC, DKRZ, LLNL, and NCI; April 2016):
  - Set up test bed in cooperation with ICNWG.
  - Install Synda at DTN sites.
  - Run transfer tests between DTNs and between ESGF data node DTN.
  - Develop an initial CMIP5 replication use case.
- Establish the ESGF replication group (core sites initially and others later; April 2016).
- Together with the publication working team, compose a "Replication and Versioning Best Practices for CMIP6" document as a guideline for data centers planning to publish CMIP6 data (June 2016).
- PID related (September 2016):
  - Integrate fully with publisher, errata service, and ESGF end-user services.
  - Set up operations for the RabbitMQ queuing system and Handle service.
  - Test software and scalability and perform the final rollout.
  - Develop additional tools as required such as an offline message publication tool.

G.14.3 Resources Needed to Achieve Replication and Versioning Goals

- Cooperation with system and network administrators at sites to set up, configure, and optimize network resources and data replication software.
- Cooperation with the Coupled Model Intercomparison Project Data Node Operations Team (CDNOT) to establish operational agreements with respect to versioning and replication.
- Initial discussions with sites planning to host "dark archives" supporting the data analysis activities of local projects and user groups with respect to "recommended" replication (and versioning) practices.
- External high-availability hosting of central RabbitMQ exchange for PID operations, independent from a single ESGF node to reduce operations risks; this may be done through a commercial hosting service, an academic cloud solution, or a cluster across multiple data centers.

G.15 Software Security Working Team

Leads: Laura Carriere and Daniel Duffy, NASA/GSFC

G.15.1 2015 Software Security Accomplishments

- Incident response:
  - Identified the attack vector.
  - Improved communications by enabling encryption keys for the development and response team members.
  - Responded to the June 2015 incident by notifying all ESGF organizations of the need to shut down ESGF websites until the software could be secured to prevent additional incidents and possible data corruption.
- Software scans:
  - Established contact at NASA/GSFC to conduct static and dynamic scans and conducted multi-

ple scans to identify security vulnerabilities in the software written by the ESGF development team.
  - Manually audited the CoG GUI software to assess vulnerabilities. Identified three minor issues that were corrected.
  - Wrote code to search the CVE database to identify vulnerabilities in third-party Java components embedded in ESGF software. Ran code against the CVE database (back to 2012) and addressed all issues identified as high or critical.
  - Manually searched the CVE database for vulnerabilities in Python components, as the code involved was relatively small.
- Software development:
  - Supported the rewriting of ESGF and LAS to address security vulnerabilities identified both during the incident and by subsequent scans.
  - Supported the automated generation of a software manifesto to provide a CVE database search list to prevent the future inclusion of vulnerable third-party software packages.

G.15.2 2016 Software Security Roadmap

- Manage a security team:
  - Identify team members, a team lead, and a mission for the security team (February 2016).
- Create a security plan:
  - Write a security plan to define roles, responsibilities, and processes for security reviews of future software releases (February 2016).
  - Obtain approval of security plan actions from the ESGF executive board, including commitment to accept roles and responsibilities (March 2016).
  - Implement the security plan (April 2016).

G.15.3 Resources Needed to Achieve Software Security Goals

- Completion of an agreed-upon ESGF security plan.
- Communication among all ESGF software development team members of the security requirements for future ESGF releases.
- Commitment by all ESGF sites to adhere to the roles and responsibilities outlined in the security plan, including, but not limited to, the identification of PoCs, maintenance of pretty good privacy (PGP) keys, and adherence to identified best practices for site configuration.

G.16 User Support Working Team

Leads: Matthew Harris, LLNL, and Torsten Rathmann, ENES/DKRZ

G.16.1 2015 User Support Accomplishments

- Released the new ESGF website with improved style, layout, and content.
- Transitioned from Askbot to a new frequently asked questions (FAQ) website: esgf.github.io/esgf-swt (see Fig. 31, p. 104).
- Monitored and archived the user mailing list.
- Reviewed and wrote ESGF stack documentation and wikis.

G.16.2 2016 User Support Roadmap

- Create infrastructure for connecting front-line support with team leads (second-level support) for more rapid responses (March 2016).
- Overhaul the wiki, separating user, developer, and administrator information. Remove outdated information and consolidate documentation locations (April 2016).
- Transition and integrate the User Support Working Team FAQ site to the CoG front end (May 2016).

G.16.3 Resources Needed to Achieve User Support Goals

- More front-line support. Power users responding to questions from the esgf-user mailing list may help to reduce response time.
- More second-line support with team leads that can respond to specific questions outside the knowledge of front-line support.
- Working with a CoG developer to complete support site integration into CoG.

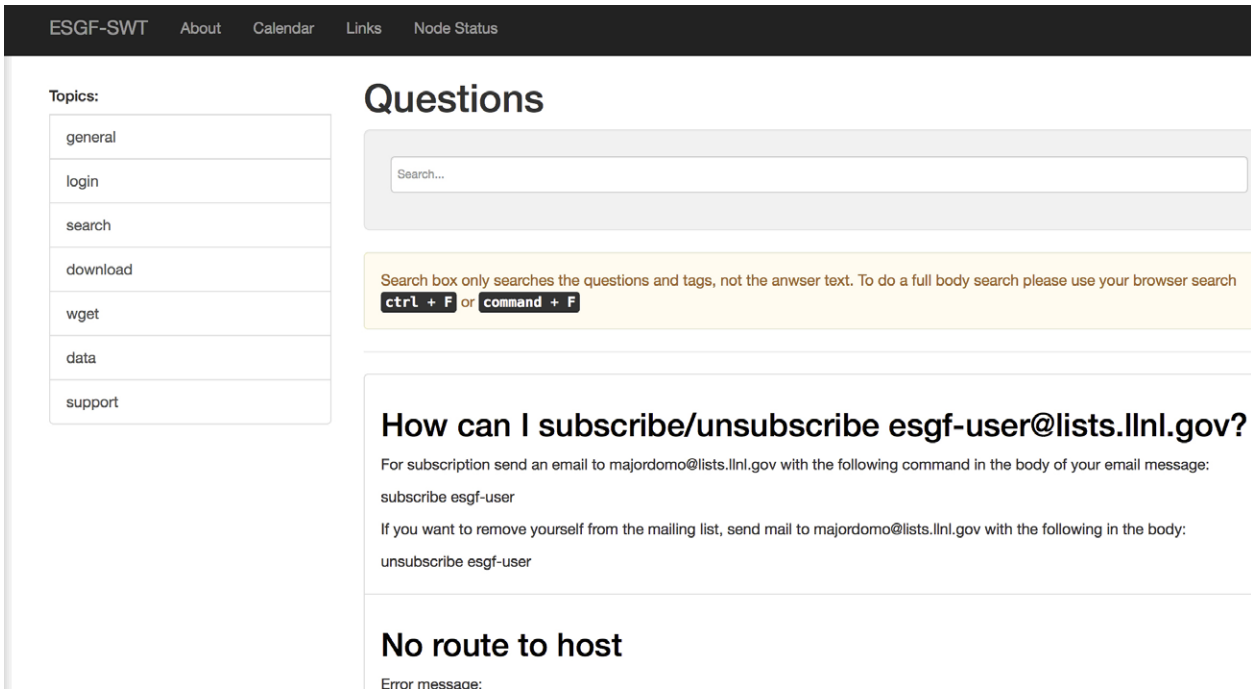


Fig. 31. Home page for the User Support Working Team's current FAQ website.

Appendix H. CMIP6 Requirements from WIP Position Papers

The Working Group on Coupled Modelling's (WGCM) Infrastructure Panel (WIP) outlined in 2014 a strategy to develop a series of position papers on the ESGF global data infrastructure and its impact on scientific experimental design. In October 2015, the WIP—chaired by V. Balaji (Princeton University and NOAA's Geophysical Fluid Dynamics Laboratory) and Karl Taylor (Lawrence Livermore National Laboratory)—presented these papers at the WGCM-19 meeting in Dubrovnik, Croatia, for endorsement by WGCM, the Coupled Model Intercomparison Project (CMIP) panel, and other modeling groups. Highlights of the papers include:

- **Formation of CDNOT:** The CMIP6 Data Node Operations Team (CDNOT) technical consortium is charged with preparing ESGF for CMIP6 operations. With Sébastien Denvil (IPSL) as chair, CDNOT was approved in June 2015. With the formation of CDNOT, ESGF governance is divided among three groups (1) requirements (WIP), (2) implementation (ESGF Executive Committee and other bodies undertaking software development), and (3) operations (CDNOT). The overlapping memberships of WIP, the ESGF Executive Committee, and CDNOT will ensure close cooperation.
- **CMIP6 data request:** Led by Martin Juckes (Science and Technology Facilities Council), the finalized CMIP6 data request is available in machine-readable formats, with associated tools for processing and analysis.
- **Data reference structure:** WIP documents covering data syntax, vocabularies, file names, and global attributes are being finalized. Upon completion, the climate model output rewriter (CMOR) and data reference syntax (DRS) specifications will be considered frozen, and modeling groups can begin constructing workflows on this basis.
- **Data format:** Recommendation – Use NetCDF4 with lossless compression as the data format for CMIP6.
- **Standard grids and calendars:** Initial WIP discussions with modeling groups regarding standard grids and calendars for output data are not yet at a consensus.
- **Model metadata:** Recommendation – Produce Earth System Documentation for model metadata as required elements in quality control and DOI generation.
- **Persistent identifiers (PIDs):** Recommendation – Use PIDs as the basis for tracking data set replication, versioning, error reporting, and usage in peer-reviewed literature.
- **Data citation:** Recommendation – Develop a mechanism for DOI generation with model and simulation granularity because citation is now a terms-of-use requirement.
- **Data licensing:** Recommendation – Create a simplified licensing scheme, wherein licenses are embedded in files. Two alternate licensing schemes are proposed: open access share-alike and noncommercial share-alike.
- **Data volume estimates:** WIP will issue preliminary estimates of aggregate data volume for CMIP6 once some aspects of the data request are finalized, taking into account the number of models, years simulated, and increase in resolution.

The 11 position papers currently in draft and others in progress are available on the WIP website: earthsystemcog.org/projects/wip/resources/. The rough CMIP6 workflow for a data package can be obtained from the quality assurance diagram in the corresponding WIP position paper (see also **Fig. 32**, p. 106).

In addition to classical NetCDF file publication, the CMIP6 ESGF data publication process (see **Fig. 33**, p. 106) includes PID registration (using the Corporation for National Research Initiatives' Handle system PID), citation information confirmation, and errata and annotation registration. Each of these services operates independently; the classic ESGF file publication will work even if one, or all, of the new services fail. The only

consequence of service failure will be that the traffic light icon, which indicates the status of the new services, will not show green.

This service is closely connected to re-publication and versioning of CMIP6 data. A specific challenge lies in ensuring consistency among versions, errata, and annotation information for the entire CMIP6 data lifecycle until long-term archival of CMIP6 reference data, at which point the data are fixed.

Replication and versioning are considered major challenges for CMIP6 and are discussed in a WIP position paper presented during the 2015 ESGF F2F Conference. ESGF, the wide area network bandwidth, CMOR, and DRS all have versioning requirements. The implementation of ESGF replication for CMIP6 critically depends on the CMOR data directory being laid out according to the CMIP6 DRS. This layout corresponds to the point data reference structure presented by the WIP at WGCM-19 (see **Fig. 34**, p. 107). The CMOR and DRS specifications are considered frozen, and modeling groups can begin constructing workflows on this basis. The same is requested for ESGF implementations such as replication, PID integration, or early citation.

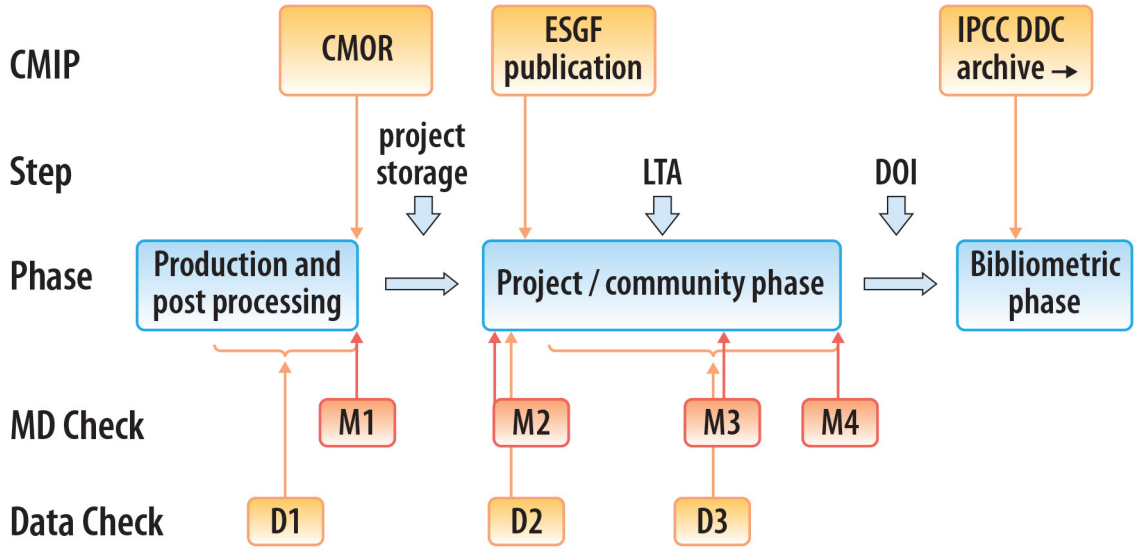


Fig. 32. Diagram from the quality assurance WIP position paper. *(D1, D2 ... Data checks; M1, M2 ... Metadata checks, and quality control of software not represented.)*

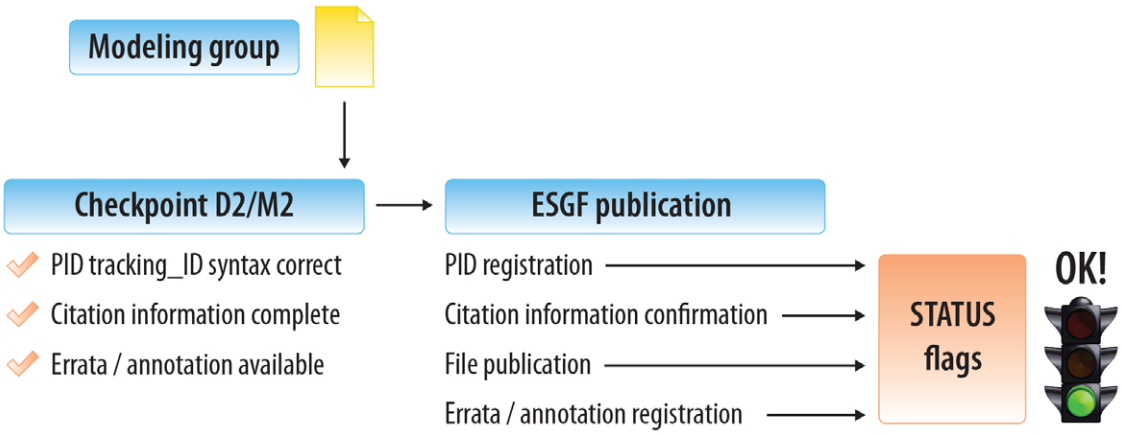


Fig. 33. Schematic CMIP6 Handle PID and early citation data publication service. *Under Checkpoint D2/M2, the new PID and citation information services must be completed before the new errata and annotations service in addition to requirements for classical ESGF file publication. Details on these new ESGF services for CMIP6 can be obtained from the three corresponding WIP position papers.*

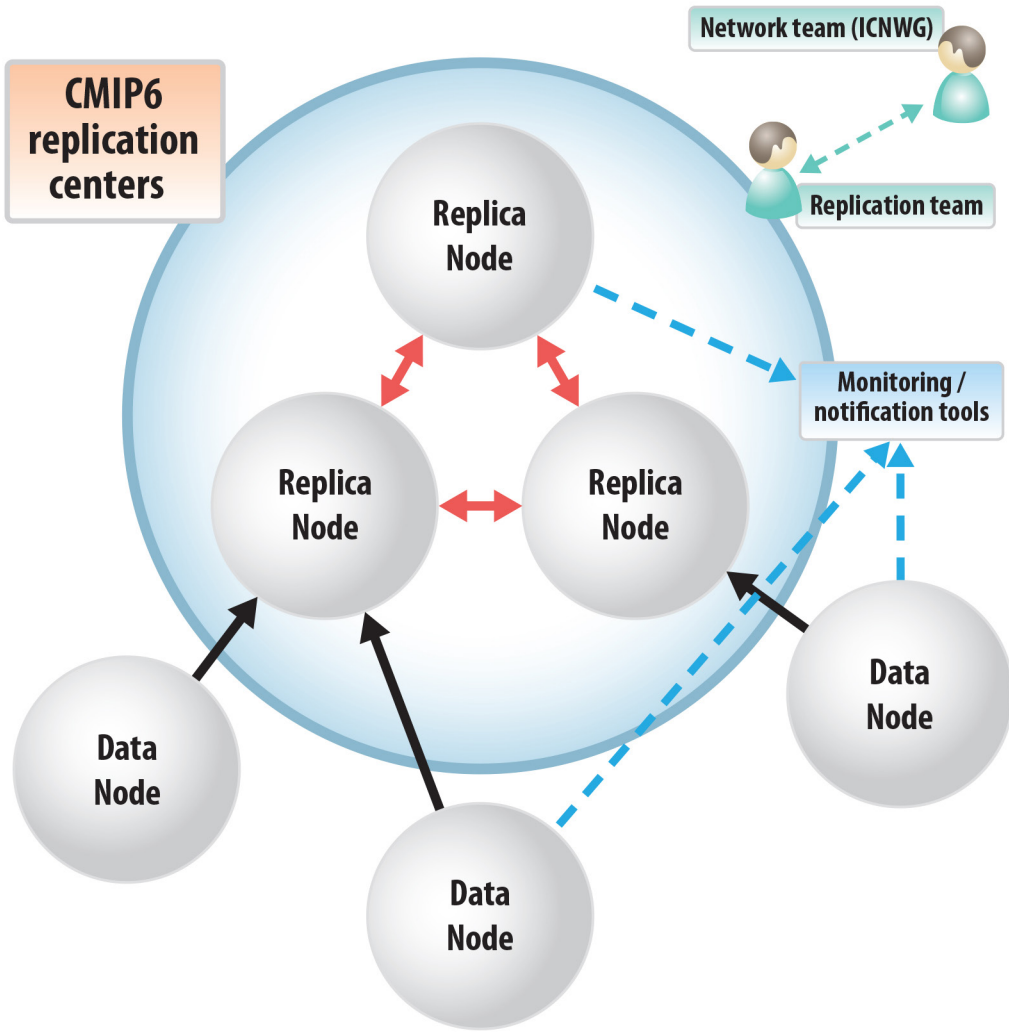


Fig. 34. CMIP6 replication from data nodes to replica centers and among replica centers coordinated by a CMIP6 replication team. *More details can be obtained from the corresponding WIP position paper.*

Appendix I. Community Development Updates

I.1 THREDDS Data Server (TDS)

TDS, a major ESGF component, exposes and catalogs data hosted by ESGF data nodes. In the context of ESGF, TDS notably provides HTTP and OPeNDAP (www.opendap.org) access to NetCDF files. TDS access is controlled by custom ESGF filters coupled with the ESGF attribute service. Consequently, only authorized users can access specific data sets.

I.1.1 2015 Accomplishments

- Completely rewrote server catalog handling and state management.
- Eliminated storing catalogs in memory.
- Refactored state management and persistent key/value store for state information.
- Reinitiated and updated TDS without shutting down ESGF.

I.1.2 2016 TDS Roadmap

- Upgrade Web Map Service (WMS) to ncWMS 2.0.
- Port ncISO tools package to a new third-party pluggable component framework.
- Rewrite GRIdded Binary format (GRIB) and Grid Feature for scalability and performance.
- Prepare TDS 5.0 release candidate by winter 2016.
- Ready TDS 5.x stable release by spring 2016.

I.1.3 2016 Roadmap for TDS in ESGF

- Detect file additions, changes, and deletions in specified data directories and update TDS catalogs and Solr indexes as needed.
- Automatically extract necessary metadata from data files to create catalogs and Solr records.
- Interface with ESGF security infrastructure to enable authorization of publishing to (possibly remote) Solr index.
- Augment the file metadata with metadata from other sources, including:
  - Mapping the directory structure to metadata fields.
  - Ingesting metadata from configuration files (e.g., in XML).
- Implement a pluggable mechanism for parsing metadata from files, enabling third parties to write their own extractors.
- Eliminate the need for ESGF's Publisher component.
- Prepare prototype for testing by May 2016 (4-month goal).
- Have production-ready TDS, incorporating feedback from testers, by June 2016 (5-month goal).

I.1.4 Resources Needed to Achieve TDS Goals

Estimated cost for TDS in ESGF work is $50,000 for 5 months.

I.2 Synda

Synda is a Python program managing discovery, authentication, and download processes from the ESGF archive. Building a local ESGF mirror is facilitating the distribution, access, and analysis of international climate data. This command-line tool is designed to simplify the download of files hosted by ESGF's distributed digital repositories. The download process is achieved by exploring ESGF data repositories using HTTP or GridFTP protocols. The search criteria, called facets, are used to select which files to download and can be set on command line or stored in a file. These attributes are defined by data reference syntax (DRS). The program may be run regularly to find potential new files for download.

I.2.1 2015 Synda Accomplishments

- Developed an easier installation procedure through Docker (www.docker.com)and Red Hat Package Manager packages (RHEL6 and RHEL7).
- Implemented a daemon mode running in the background (i.e., eliminating the need to stop/start during new discovery phase).
- Developed a per-user configuration file.
- Extended support for most ESGF projects [e.g., CMIP5, CORDEX, Obs4MIPs, and

Seasonal-to-decadal climate Prediction for the improvement of European Climate Services (SPECS)].

- Implemented nearest-replica selection.
- Added time coverage and local search filters.
- Added support for SHA256 checksum.
- Implemented support for GridFTP protocol.
- Enabled an option for triggering a postprocessing pipeline applied to downloaded files.

I.2.2 2016 Synda Roadmap

Synda appears to be a good alternative for ESGF data replication for several ESGF partners. The program will evolve together with the ESGF archive back-end functionalities to provide a robust approach to replication. The postprocessing module (sdp : Synda processing) enables users to designate workers to handle various publication steps. A "replication" pipeline can be set, triggering map file generation and publication of downloaded data sets as replicas. A fully automated replication procedure will be developed, in which updates on the source dataset will be pushed to Synda replication instances in the federation.

I.2.3 Resources Needed to Achieve Synda Goals

The Institut Pierre-Simon Laplace (IPSL) has undertaken Synda development since 2011, with Jérôme Raciazek serving as lead developer. Raciazek also provides Synda technical support to the ESGF community. IPSL will continue to support Raciazek but hopes that other partners will be able to contribute to the source code as well.

I.3 Globus

Globus provides a hosted solution for the management, transfer, sharing, publication, and discovery of research data. Globus transfer capabilities provide a reliable, secure mechanism for moving data at any scale among sites using a simple interface. Globus also enables users to share data with collaborators without replicating data to the cloud. Users can make folders with read or read/write access available to collaborators using their identity or email address.

Globus is leveraged extensively for data management and has been used to move more than 100 PB of data. It has over 10,000 active endpoints and 25,000 registered users.

I.3.1 2015 Globus Accomplishments

- Implemented new capabilities for administrators to manage network use, including:
  - Ability to set concurrency and parallelism parameters.
  - Ability to monitor and manage tasks (pause/resume of transfers) via a new administrative management console.
- Added support for various storage systems such as high-performance storage systems (HPSS), Amazon S3, and Spectra Logic BlackPearl.
- Released the publication capability for users to publish data sets by associating metadata and persistent identifiers and making the data available for search and discovery. The publication service keeps the data and a copy of the metadata at the ESGF node for discovery.

I.3.2 2016 Globus Roadmap

An upcoming release will add support for HTTP/S access to files, anonymous sharing, and streamlined provisioning that removes the need for a Globus account. Globus will move to using openID Connect and will support logins with federated identity providers (IdPs), without the need for a Globus username and password.

Globus capabilities, while most commonly accessed using web and command-line interfaces, often are leveraged as a platform by third-party applications. ESGF itself uses Globus for data download via the portal. An effort is under way to integrate download of open and restricted data via CoG. ESGF's upgrade to OAuth will enable a more secure and seamless integration using Globus for data transfer from ESGF archives. ESGF also plans to integrate Globus as an option with the ESGF replication infrastructure and tools, which will provide high-performance and reliable managed data transfer capability to replication tools using ESGF metadata. Collaboration between the Globus and CoG working teams to integrate the

ingestion and publication service with CoG also is planned to provide users access to new capabilities from the familiar CoG interface.

The federation also sees opportunities for ESGF to leverage Globus Auth capabilities to allow access to capabilities using federated logins. Globus Auth uses OAuth and acts as a bridge among several InCommon IdPs and dependent services such as ESGF services. This authentication approach eliminates the need for ESGF sites to run an IdP and for users to create a separate account to access ESGF data.

Appendix J. Conference Participants and Report Contributors

Fig. 35. Group photo of participants in the 2015 international Earth System Grid Federation Face-to-Face Conference.

J.1 Joint International Agency Conference and Report Organizers

Dean N. Williams – Chair of the ESGF Executive Committee, U.S. Department of Energy (DOE) Lawrence Livermore National Laboratory (LLNL)

Michael Lautenschlager – Co-Chair of the ESGF Executive Committee, European Network for Earth System Modelling (ENES)/German Climate Computing Centre (DKRZ)

Luca Cinquini – ESGF Executive Committee, National Aeronautics and Space Administration (NASA)/National Oceanic and Atmospheric Administration (NOAA)

Sébastien Denvil – ESGF Executive Committee, Institut Pierre-Simon Laplace (IPSL)

Robert Ferraro – ESGF Executive Committee, NASA

Daniel Duffy – ESGF Executive Committee, NASA

Claire Trenham – ESGF Executive Committee, National Computational Infrastructure (NCI)

V. Balaji – ESGF Executive Committee, NOAA

Cecilia DeLuca – ESGF Executive Committee, NOAA

J.2 ESGF Program Managers in Attendance

Justin Hnilo – Chair of the ESGF Steering Committee, DOE Office of Biological and Environmental Research (BER)
Tsendgar Lee – ESGF Steering Committee, NASA
Ben Evans – ESGF Steering Committee, NCI

Attendees and Contributors	
Name	**Affiliation**
1. AchutaRao, Krishna	Indian Institute of Technology, Delhi
2. Ames, Alexander	DOE LLNL
3. Ananthakrishnan, Rachana	University of Chicago
4. Anantharaj, Valentine	DOE Oak Ridge National Laboratory (ORNL)
5. Arakawa, Osamu	University of Tsukuba
6. Bader, David	DOE/LLNL
7. Berger, Katharina	ENES/DKRZ
8. Caron, John	Sunya (Independent)
9. Cheng, Huaqiong	Beijing Normal University
10. Christensen, Cameron	Scientific Computing and Imaging Institute/University of Utah
11. Cinquini, Luca	NASA Jet Propulsion Laboratory (JPL) and NOAA Earth System Research Laboratory (ESRL)
12. D'Anca, Alessandro	ENES/Euro-Mediterranean Center on Climate Change (CMCC)
13. Dart, Eli	DOE ESnet
14. Denvil, Sébastien	Centre Nationale de la Recherche Scientifique (CNRS)/IPSL
15. Doutriaux, Charles	DOE LLNL
16. Duffy, Daniel	NASA Goddard Space Flight Center (GSFC) and Center for Climate Simulation (NCCS)
17. Dwarakanath, Prashanth	National Supercomputer Centre
18. *Evans, Ben	NCI/Australian National University
19. Ferraro, Robert	NASA JPL
20. Fiore, Sandro	CMCC
21. Fries, Samuel	DOE LLNL
22. Gleckler, Peter	DOE LLNL/ Program for Climate Model Diagnosis and Intercomparison (PCMDI)
23. Hansen, Rose	DOE LLNL
24. Harney, John	DOE ORNL
25. Harr, Cameron	DOE LLNL
26. Harris, Matthew	DOE LLNL
27. Hester, Mary	DOE ESnet
28. *Hnilo, Jay	DOE BER Headquarters
29. Jhaveri, Sankhesh	Kitware, Inc.
30. Jefferson, Angela	DOE LLNL

31. Kershaw, Philip	National Centre for Atmospheric Science (NCAS)/ British Atmospheric Data Centre (BADC)
32. Kindermann, Stephan	DKRZ
33. Kolax, Michael	Swedish Meteorological and Hydrological Institute
34. Kostov, Georgi	NOAA National Climatic Data Center (NCDC)
35. Koziol, Benjamin	NOAA ESRL and Environmental Software Infrastructure and Interoperability group / Cooperative Institute for Research in Environmental Sciences
36. Krassovski, Misha	CDIAC/ORNL
37. Lacinski, Lukasz	University of Chicago
38. Lautenschlager, Michael	ENES/DKRZ
39. *Lee, Tsengdar	NASA Headquarters
40. Levavasseur, Guillaume	Infrastructure for the European Network for Earth System Modeling (IS-ENES)/IPSL
41. Maxwell, Thomas	NASA GSFC
42. McCoy, Renata	DOE LLNL
43. McEnerney, James	DOE LLNL
44. Nadeau, Denis	DOE LLNL
45. Nikonov, Serguei	NOAA Geophysical Fluid Dynamics Laboratory (GFDL)/Princeton University
46. Ogochi, Koji	Japan Agency for Marine-Earth Science and Technology
47. Papadopoulos, Christos	Colorado State University
48. Painter, Jeffrey	DOE LLNL
49. Peterschmitt, Jean-Yves	Climate and Environment Sciences Laboratory (LSCE)/IPSL
50. Plieger, Maarten	ENES/Royal Netherlands Meteorological Institute (KNMI)
51. Pobre, Alakom-Zed	NASA GSFC and NCCS
52. Potter, Gerald	NASA GSFC
53. Pritchard, Matt	Centre for Environmental Data Analysis's (CEDA) Science and Technology Facilities Council (STFC)
54. Rathmann, Torsten	ENES/DKRZ
55. Schweitzer, Roland	Weathertop Consulting, LLC
56. Smith, Brian	DOE ORNL
57. Stockhause, Martina	DKRZ
58. Story, Matthew	DOE LLNL
59. Taylor, Karl	DOE LLNL/PCMDI
60. Tucker, William	CEDA STFC
61. Vahlenkamp, Hans	NOAA GFDL
62. Wang, Dali	DOE ORNL
63. Weigel, Tobias	DKRZ
64. Williams, Dean	DOE LLNL
65. Wu, Qizhong	Beijing Normal University

* Funding agency program manager

Online Attendees and Contributors	
Last Name	**Affiliation**
1. Aloisio, Giovanni	CMCC
2. Balaji, V.	NOAA GFDL/Princeton University
3. Berkley, Mike	Canadian Centre for Climate Modeling for Analysis (CCCma)
4. Burger, Eugene	NOAA Pacific Marine Environmental Laboratory (PMEL)
5. Carenton-Madiec, Nicolas	ENES/IPSL
6. Carriere, Laura	NASA GSFC
7. Cofiño, Antonio S.	Santander Meteorology Group (University of Cantabria)
8. Chaudhary, Aashish	Kitware, Inc.
9. Cheng, Hua Qiong	Beijing Normal University /Chinese Academy of Meteorological Sciences
10. Chunpir, Hashim	ENES/DKRZ
11. Cram, Thomas	National Science Foundation (NSF)/ National Center for Atmospheric Research (NCAR)
12. DeLuca, Cecelia	NOAA ESRL
13. Durack, Paul	DOE LLNL/PCMDI
14. Greenslade, Mark	IS-ENES2/IPSL
15. Kharin, Slava	CCCma/Environment Canada
16. Mason, Erik	NOAA GFDL
17. Murphy, Sylvia	NOAA ESRL
18. Nassisi, Paola	CMCC
19. O'Brien, Kevin	NOAA PMEL/University of Washington, Joint Institute for the Study of the Atmosphere and Ocean
20. Raju, Bibi	DOE Pacific Northwest National Laboratory
21. Rutledge, Glenn	NOAA NCDC
22. Schuster, Doug	NSF/NCAR
23. Sim, Alex	DOE Lawrence Berkeley National Laboratory
24. Trenham, Claire	NCI/Australian National University
25. Wei, Min	National Meteorological Information Center, China Meteorological Administration

Appendix K. Awards

K.1 Federal Laboratory Consortium Awards

Started in 1974, The Federal Laboratory Consortium (FLC) assists the U.S. public and private sectors in using technologies developed by federal government research laboratories. It is composed of more than 300 federal government laboratories and research centers.

- The ESGF community won the 2013 Far West Region FLC technology transfer award for outstanding partnership.
- ESGF's Ultrascale Visualization–Climate Data Analysis Tools (UV-CDAT) component won the 2014 Far West Region FLC technology transfer award for outstanding partnership and technical achievement.
- On April 29, 2015, ESGF's UV-CDAT component won the National FLC Interagency Partnership award, making this the third consecutive FLC award for the interagency community, This award was one of six team awards presented at the National FLC Meeting, which took place in Denver, Colorado. One of the consortium's highest honors, the FLC award "recognizes the efforts of laboratory employees from at least two different agencies who have collaboratively accomplished outstanding work in the process of science and/or transferring a technology."

  The primary U.S. partnership recognized for the FLC award consists of DOE's Lawrence Livermore (LLNL), Lawrence Berkeley, Los Alamos, and Oak Ridge national laboratories; NASA's Goddard Space Flight Center (GSFC); NOAA's Earth System Research Laboratory (ESRL); New York University; the University of Utah; Kitware Inc.; and Tech-X Corporation.

K.2 Internal Awards

Every year, the climate software engineering community gathers to determine who has performed exceptional or outstanding work in the successful development of community tools for the acceleration of climate science in the ESGF data science domain. This year, the ESGF Executive Committee determined the winners. These awards recognize dedicated members of the ESGF community who are contributing nationally and internationally to federation efforts. Recipients of the awards capture and display the best of the community's spirit and determination to succeed. The Executive Committee's recognition of these members' efforts is but a small token of appreciation and does not exclude others who also are working hard to make ESGF a success.

Award winners contributing to ESGF success:

- **Prashanth Dwarakanath, European Network for Earth System Modelling (ENES) and Linköping University (LiU), and Nicolas Carenton-Madiec, ENES and Institut Pierre-Simon Laplace (IPSL),** won awards in 2014 for their outstanding contributions and leadership in the continued development of the ESGF software stack installer. In 2015, they continued these contributions by leading the difficult ESGF software stack overhaul and helping to coordinate security scans. Through development of the revised installer, they led the release of ESGF 2.0 and 2.1 and helped with the upgrades of many of the underlying system infrastructure components, including Python 2.7.9, Java 1.8, Tomcat 8, Postres 9.4, and OpenSSL 1.0. Their efforts also included switching the ESGF installer to utilize Red Hat Package Managers (RPMs) for faster and easier ESGF installation.
- **Luca Cinquini, NOAA ESRL,** won an award in 2014 for the design and prototype of the new ESGF CoG user interface. In 2015, he won for his technical leadership in providing major upgrades to the search services (Solr5), data downloads through the THREDDS Data Server version 5 (TDS5), and high-performance data transfer work with the Globus-Connect-Server. These efforts are all part of the major ESGF software stack overhaul and will ease users' ability to navigate petabytes of ESGF data sets.
- **Alexander Ames, DOE LLNL,** won an award for his years of ESGF publication leadership and coordination and republishing of projects across the federation, including the Coupled Model Intercomparison Project (CMIP) and Accelerated Climate Modeling for Energy (ACME) project. In addition,

he collaborated with other members of the ESGF Installation Working Team to upgrade, test, and deploy early ESGF v2.x candidate releases. He also was recognized for his contributions to many areas of ESGF development, such as compute clusters, node managers, software security scans, data transfers, and the dashboard, to name a few.

- **Laura Carriere, NASA GSFC**, won an award for her ongoing commitment and support for securing ESGF software security scans. Her work was critical for the release of ESGF v2.x. Members of the ESGF Executive Committee have come to recognize that technologies such as ESGF are useless and would not be deployed at secured sites without the diligent work of those who conduct software security scans to detect vulnerabilities. Laura led the effort to secure NASA funding and resources for this critical effort. Without her involvement, the ESGF software stack would not have been scanned, nor would ESGF v2.x have been released.

Appendix L. Acknowledgments

The 2015 Earth System Grid Federation (ESGF) Face-to-Face Conference organizers wish to thank the national and international funding agencies for providing travel funding for attendees to join the conference in person, Lawrence Livermore National Laboratory (LLNL) for hosting the annual event, and the presenters for their contributions to the conference and this report. The organizers especially acknowledge LLNL's Angela Jefferson for her help in processing endless paperwork, finding the conference location, and arranging many other important logistics. We also acknowledge and appreciate LLNL's video and media services support; Matthew Story for setting up and breaking down presentation equipment; and Technical Information Department technical writer Rose Hansen for taking the detailed conference notes used in this report.

ESGF development and operation continue to be supported by the efforts of principal investigators, software engineers, data managers, projects (e.g., CMIP, ACME, CORDEX, and many others), and system administrators from many agencies and institutions worldwide. Primary contributors to these open-source software products include: Argonne National Laboratory; Australian National University; British Atmospheric Data Centre; Euro-Mediterranean Center on Climate Change (CMCC); German Climate Computing Centre (DKRZ); Earth System Research Laboratory; Geophysical Fluid Dynamics Laboratory; Goddard Space Flight Center; Institut Pierre-Simon Laplace; Jet Propulsion Laboratory; Kitware, Inc.; National Center for Atmospheric Research; New York University; Oak Ridge National Laboratory; Los Alamos National Laboratory; Lawrence Berkeley National Laboratory; LLNL (leading institution); and the University of Utah. Many other organizations and institutions have contributed to the efforts of ESGF, and apologies to any whose names have been unintentionally omitted.

The U.S. Department of Energy, National Aeronautics and Space Administration, National Oceanic and Atmospheric Administration, Infrastructure for the European Network for Earth System Modeling, and the Australian National Computational Infrastructure provide major funding for the community software efforts.

Appendix M. Acronyms and Terms

Acronym	Definition
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 (climatemodeling.science.energy.gov/projects/accelerated-climate-modeling-energy/).
AIRS	Atmospheric Infrared Sounder — One of six instruments onboard Aqua, which is part of NASA's Earth Observing System of satellites. Its goal is to support climate research and improve weather forecasting (disc.gsfc.nasa.gov/AIRS/).
ALCF	Argonne Leadership Computing Facility — DOE Office of Science user facility that provides researchers from national laboratories, academia, and industry with access to high-performance computing capabilities (http://airs.jpl.nasa.gov).
ANL	Argonne National Laboratory — Science and engineering research national laboratory near Lemont, Illinois, operated by the University of Chicago for DOE (www.anl.gov).
API	application programming interface (en.wikipedia.org/wiki/Application_programming_interface/).
AR	Assessment Report
ARMBE	Atmospheric Radiation Measurement Best Estimate — Data products from DOE's Atmospheric Radiation Measurement (ARM) Climate Research Facility that are specifically tailored for use in evaluating global climate models. They contain a best estimate of several cloud, radiation, and atmospheric quantities.
BADC	British Atmospheric Data Centre — The Natural Environment Research Council's (NERC) designated data center for atmospheric sciences (badc.nerc.ac.uk/home/index.html).
BASE	Berkeley Archival Storage Encapsulation
BER	DOE Office of Biological and Environmental Research — Supports world-class biological and environmental research programs and scientific user facilities to facilitate DOE's energy, environment, and basic research missions (science.energy.gov/ber/).
CAaaS	Climate Analytics-as-a-Service
CCCma	Canadian Centre for Climate Modeling for Analysis — Develops and applies climate models to improve the understanding of climate change and make quantitative projections of future climate in Canada and globally (www.ec.gc.ca/ccmac-cccma/).
CDAS	Climate Data Assimilation System — NOAA's CDAS provides access to atmospheric reanalysis data and historical archives including monthly and daily averages of many standard pressure-level data and surface flux quantities (www.cpc.ncep.noaa.gov/products/wesley/cdas_data.html)
CDIAC	Carbon Dioxide Information Analysis Center — DOE's primary climate change data and information analysis center whose data holdings include estimates of CO_2 emissions from fossil fuel consumption and land-use changes, records of atmospheric concentrations of CO_2 and other radiatively active trace gases, carbon cycle and terrestrial carbon management data sets and analyses, and global and regional climate data and time series (cdiac.ornl.gov).
CDNOT	Coupled Model Intercomparison Project Data Node Operations Team
CDS	Climate Model Data Services — Sponsored by NASA, CDS provides a central location for publishing and accessing large, complex climate model data to benefit the climate science community and the public (cds.nccs.nasa.gov).
CEDA	Centre for Environmental Data Analysis — Serves the environmental science community through four data centers, data analysis environments, and participation in numerous research projects that support environmental science, advance environmental data archival practices, and develop and deploy new technologies to enhance data access (www.ceda.ac.uk).
CERES	Clouds and Earth's Radiant Energy System — NOAA's instrument that measures reflected sunlight and thermal radiation emitted by the Earth (www.jpss.noaa.gov/ceres.html).

Acronym	Definition
CMCC	Centro Euro-Mediterraneo sui Cambiamenti Climatici (Euro-Mediterranean Center on Climate Change) — The CMCC scientific organization in Italy enhances collaboration and integration among climate science disciplines (www.cmcc.it/cmccesgf-data-node/).
CMIP	Coupled Model Intercomparison Project — Sponsored by the World Climate Research Programme's Working Group on Coupled Modeling, CMIP is a community-based infrastructure for climate model diagnosis, validation, intercomparison, documentation, and data access (cmip-pcmdi.llnl.gov).
CMIP-DECK	CMIP–Diagnostic, Evaluation, and Characterization of Klima — Ongoing sets of small modeling experiments whose output will be distributed for community use via the ESGF infrastructure.
CMOR	Climate Model Output Rewriter — Comprises a set of C-based functions that can be used to produce NetCDF files that comply with climate forecast conventions and fulfill many requirements of the climate community's standard model experiments (pcmdi.github.io/cmor-site/).
CNRI	Corporation for National Research Initiatives — Nonprofit organization that undertakes, fosters, and promotes research in the public interest, including strategic development of network-based information technologies, providing leadership, and funding for information infrastructure research and development (www.cnri.reston.va.us).
CNRS	Centre Nationale de la Recherche Scientifique (French National Centre for Scientific Research) — Largest fundamental science agency in Europe (www.cnrs.fr/).
CoG	Continuity of Operations Plan
COOP	continuity of operations plan
CORDEX	Coordinated Regional Climate Downscaling Experiment — Provides global coordination of regional climate downscaling for improved regional climate change adaptation and impact assessment (www.cordex.org).
CREATE-IP	Collaborative REAnalysis Technical Environment Intercomparison Project — Data collection, standardization, and ESGF distribution component of CREATE (www.earthsystemcog.org/projects/create-ip/).
CVE	common vulnerability and exposures
CWT	ESGF Compute Working Team
DataCite	Nonprofit organization that develops and support methods to locate, identify, and cite data and other research objects (www.datacite.org).
Data node	Internet location providing data access or processing (en.wikipedia.org/wiki/Node-to-node_data_transfer/).
Data provider	Modeling group (or Obs4MIPs contributor).
Data publisher	Person who publishes CMIP data sets to ESGF.
DECK	Diagnostic, Evaluation, and Characterization of Klima — Ongoing sets of small modeling experiments whose output will be distributed for community use via the ESGF infrastructure.
DKRZ	Deutsches Klimarechenzentrum (German Climate Computing Centre) — Provides high-performance computing platforms and sophisticated, high capacity data management and services for climate science (www.dkrz.de).
DOE	U.S. Department of Energy — Government agency chiefly responsible for implementing energy policy (www.doe.gov).
DOI	digital object identifier — Serial code used to uniquely identify content of various types of electronic networks; particularly used for electronic documents such as journal articles (en.wikipedia.org/wiki/Digital_object_identifier/).

Acronym	Definition
DREAM	Distributed Resources for the ESGF Advanced Management — Provides a new way to access large data sets across multiple DOE, NASA, and NOAA compute facilities, which will improve climate research efforts as well as numerous other data-intensive applications (esgf.llnl.gov/media/2015-F2F/Posters/DREAM-Distributed-Resources-for-the-ESGF-Advanced-Management.pdf).
DRS	Data Reference Syntax — Naming system to be used within files, directories, metadata and URLs to identify data sets wherever they might be located within the distributed ESGF archive.
DTN	data transfer node — Internet location providing data access, processing, or transfer (fasterdata.es.net/science-dmz/DTN/).
DTWT	ESGF Data Transfer Working Team
ENES	European Network for Earth System Modelling — Common infrastructure for distributed climate research and modeling in Europe, integrating community Earth system models and their hardware, software, and data environments (www.eudat.eu/communities/enes-european-network-earth-system-modelling).
ES-DOC	Earth System Documentation
ESG	Earth System Grid
ESG-CET	Earth System Grid Center for Enabling Technologies — DOE project providing climate researchers worldwide with access to data, information, models, analysis tools, and computational resources required to understand climate simulation datasets.
ESGF	Earth System Grid Federation — Led by Lawrence Livermore National Laboratory, a worldwide federation of climate and computer scientists deploying a distributed multipetabyte archive for climate science (esgf.llnl.gov).
ESM	Earth system model — Type of complex, global model that combines physical climate models, global biological processes, and human activities.
ESMValTool	Earth system model evaluation tool
ESnet	DOE Energy Sciences Network — Provides high-bandwidth connections that link scientists at national laboratories, universities, and other research institutions, enabling them to collaborate on scientific challenges including energy, climate science, and the origins of the universe (www.es.net).
ESRL	Earth System Research Laboratory — NOAA ESRL researchers monitor the atmosphere, study the physical and chemical processes that comprise the Earth system, and integrate results into environmental information products that help improve weather and climate tools for the public and private sectors (www.esrl.noaa.gov).
EzCMOR	Front end to an existing free software package, CMOR (Climate Model Output Rewriter), written by Lawrence Livermore National Laboratory. EzCMOR reads many standard data formats and converts them into the CMIP5 data format to allow publication on the ESGF data node (github.com/PCMDI/ezCMOR/).
F2F	Face To Face
FISMA	Federal Information Security Management Act (www.dhs.gov/fisma).
FLC	Federal Laboratory Consortium — Community of more than 300 federal laboratories, facilities, research centers, and their parent agencies (www.federallabs.org).
GA	Global Attribute
GFDL	Geophysical Fluid Dynamics Laboratory — NOAA's GFDL develops and uses mathematical models and computer simulations to improve the understanding and prediction of atmospheric, oceanic, and climatic behaviors (www.gfdl.noaa.gov).
GISS	NASA Goddard Institute for Space Studies — GISS research aims to predict atmospheric and climate changes in the 21st century by combining analysis of comprehensive global datasets with global models of atmospheric, land surface, and oceanic processes (www.giss.nasa.gov).

Acronym	Definition
Globus	Provides high-performance, secure, and reliable data transfer, sharing, synchronization, and publication services for the science community (www.globus.org).
GridFTP	High-performance, secure, reliable data transfer protocol optimized for high-bandwidth wide-area networks (toolkit.globus.org/toolkit/docs/latest-stable/gridftp/).
GSFC	Goddard Space Flight Center — The NASA space research laboratory in Greenbelt, Maryland was established as NASA's first space flight center. GSFC is home to the nation's largest organization of scientists, engineers and technologists who build spacecraft, instruments, and new technology to study Earth, the sun, our solar system, and the universe (www.nasa.gov/centers/goddard/home/index.html).
Hadoop®	Project that develops open-source software for reliable, scalable, distributed computing (hadoop.apache.org).
HDF5	HDF5 Hierarchical Data Format version 5 — Data model, library, and file format for storing and managing a wide variety of high-volume and complex data types (www.hdfgroup.org/HDF5/).
HFP	Heterogeneous Functional Partitioning — Runtime framework that enables on-the-fly, *in situ* analysis on a compute node by dynamically exploiting underutilized node resources.
HPC	high-performance computing
HPSS	high-performance storage systems — Modern, flexible, performance-oriented mass storage system (www.hpss-collaboration.org/).
ICNWG	International Climate Network Working Group — Formed under the Earth System Grid Federation to help set up and optimize network infrastructure for climate data sites around the world (icnwg.llnl.gov).
IdEA	ESGF Identity, Entitlement, and Access Working Team
IdP	identity provider
IPCC	Intergovernmental Panel on Climate Change — Scientific body of the United Nations that periodically issues assessment reports on climate change (www.ipcc.ch).
IPCC DDC	Intergovernmental Panel on Climate Change Data Distribution Centre
IPSL	Institut Pierre-Simon Laplace — Nine-laboratory French research institution whose topics focus on the global environment. Main objectives include understanding (1) the dynamic chemical and biological processes at work in the Earth System, (2) natural climate variability at regional and global scales, and (3) the impacts of human activities on climate (www.ipsl.fr/en/).
IS-ENES2	Infrastructure for the European Network for Earth System Modeling — Second-phase project of the distributed e-infrastructure of models, model data, and metadata of the European Network for Earth System Modelling (is.enes.org).
ISSO	Information System Security Officer
IWT	ESGF Interface Working Team
JPL	Jet Propulsion Laboratory — A federally funded research and development laboratory and NASA field center in Pasadena, California (www.jpl.nasa.gov).
JSON	JavaScript Object Notation — An open- and text-based data exchange format that provides a standardized data exchange format better suited for Ajax-style web applications (www.json.org).
KNMI	Royal Netherlands Meteorological Institute — Dutch national weather service and the national research and information center for meteorology, climate, air quality, and seismology (www.knmi.nl/over-het-knmi/about).
LAS	Live Access Server
LBNL	Lawrence Berkeley National Laboratory — DOE Office of Science laboratory managed by the University of California that conducts fundamental science for transformational solutions to energy and environmental challenges using interdisciplinary teams and advanced new tools for scientific discovery (www.lbl.gov).

Acronym	Definition
LIU	Linköpings Universitet's National Supercomputer Centre in Sweden — Houses an ESGF data node, test node, ESGF code sprint, user support, Bi and Frost clusters (www.nsc.liu.se/).
LLNL	Lawrence Livermore National Laboratory — DOE laboratory that develops and applies world-class science and technology to enhance the nation's defense and address scientific issues of national importance (www.llnl.gov).
LSCE	Climate and Environment Sciences Laboratory — IPSL laboratory whose research focuses on the mechanisms of natural climate variability at different time scales; interactions among human activity, the environment, and climate; the cycling of key compounds such as carbon, greenhouse gases, and aerosols; and the geosphere and its relationship with climate (www.gisclimat.fr/en/laboratory/lsce-climate-and-environment-sciences-laboratory).
LTA	long-term archival
MapReduce	A programming model and associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster (en.wikipedia.org/wiki/MapReduce).
MERRA	Modern Era Retrospective-Analysis for Research and Applications — Data analysis that places observations from NASA's Earth Observing System satellites into a climate context and helps improve on the hydrologic cycle represented in earlier generations of reanalyses (gmao.gsfc.nasa.gov/research/merra/).
Metadata	Data properties, such as origin, spatiotemporal extent, and format (en.wikipedia.org/wiki/Metadata/).
MIP	Model Intercomparison Project
MISR	Multi-angle Imaging SpectroRadiometer — Provides new types of information for scientists studying Earth's climate, such as the partitioning of energy and carbon between the land surface and atmosphere and the regional and global impacts of different types of atmospheric particles and clouds on climate (www-misr.jpl.nasa.gov).
MLS	Microwave Limb Sounder — NASA instrumentation that uses microwave emission to measure stratospheric temperature and upper tropospheric constituents. MLS also measures upper tropospheric water vapor in the presence of tropical cirrus and cirrus ice content (aura.gsfc.nasa.gov/scinst/mls.html).
MSWT	ESGF Metadata and Search Working Team
NASA	National Aeronautics and Space Administration — U.S. government agency responsible for the civilian space program as well as aeronautics and aerospace research (www.nasa.gov).
NCAR	National Center for Atmospheric Research — Federally funded research and development center devoted to service, research, and education in atmospheric and related sciences (ncar.ucar.edu).
NCAS	National Centre for Atmospheric Science — Conducts research on climate change, atmospheric composition, weather, and technologies for observing and modeling the atmosphere. The center also provides scientific facilities for atmospheric research to scientists across the United Kingdom including aircraft, ground-based instrumentation, and computing resources (www.ncas.ac.uk/)
NCCS	NASA Center for Climate Simulation — An integrated set of supercomputing, visualization, and data interaction technologies that enhance capabilities in weather and climate prediction research (www.nccs.nasa.gov).
NCDC	National Climatic Data Center — One of three former NOAA data centers that have been merged into the National Centers for Environmental Information, which is responsible for hosting and providing access to comprehensive oceanic, atmospheric, and geophysical data (www.ncdc.noaa.gov/about).
NCI	National Computational Infrastructure — Australia's high-performance supercomputer, cloud, and data repository (nci.org.au).
NDN	Named Data Networking — Entirely new Internet architecture whose design principles are derived from the successes of today's Internet and can be rolled out through incremental deployment over the current operational Internet (named-data.net/).

Acronym	Definition
NERDIP	National Environmental Research Data Interoperability Platform — NCI's *in situ* petascale computational environment enabling both high-performance computing and data-intensive science across a wide spectrum of environmental and Earth science data collections.
NERSC	National Energy Research Scientific Computing Center — Primary scientific computing facility for the DOE Office of Science, providing computational resources and expertise for basic scientific research (www.nersc.gov).
NetCDF	Network Common Data Form — Machine-independent, self-describing binary data format (www.unidata.ucar.edu/software/netcdf/).
NMWT	Node Manager Working Team
NOAA	National Oceanic and Atmospheric Administration — Federal agency whose missions include understanding and predicting changes in climate, weather, oceans, and coasts and conserving and managing coastal and marine ecosystems and resources (www.noaa.gov).
NSF	National Science Foundation — Federal agency that supports fundamental research and education in all the nonmedical fields of science and engineering (www.nsf.gov).
OAuth	Open standard for authorization (en.wikipedia.org/wiki/OAuth/).
Obs4MIPs	Observations for Model Intercomparisons — Database used by the CMIP modeling community for comparing satellite observations with climate model predictions (www.earthsystemcog.org/projects/obs4mips/).
OLCF	Oak Ridge Leadership Computing Facility — DOE national user facility providing the open scientific community support and access to computing resources including the nation's most powerful supercomputer to address grand challenges in climate, materials, nuclear science, and a wide range of other disciplines (www.olcf.ornl.gov).
OPeNDAP	Open-source Project for a Network Data Access Protocol — A rchitecture for data transport including standards for encapsulating structured data and describing data attributes (www.opendap.org).
OpenID	An open standard and decentralized authentication protocol. (CoG uses an ESGF OpenID as its authentication mechanism.)
OPTIRAD	OPTImisation environment for joint retrieval of multi-sensor RADiances — Collaborative research environment for data assimilation in land applications.
ORNL	Oak Ridge National Laboratory — DOE science and energy laboratory conducting basic and applied research to deliver transformative solutions to compelling problems in energy and security (www.ornl.gov).
PB	petabyte
PCMDI	Program for Climate Model Diagnosis and Intercomparison — Develops improved methods and tools for the diagnosis and intercomparison of general circulation models that simulate the global climate (www-pcmdi.llnl.gov).
PGP	Pretty Good Privacy — data encryption technology.
PID	Persistent identifier — A long-lasting reference to a digital object, a single file or set of files (en.wikipedia.org/wiki/Persistent_identifier/).
PIS	Persistent Identifier Service
PMEL	Pacific Marine Environmental Laboratory — NOAA laboratory that conducts observations and research to advance knowledge of the global ocean and its interactions with the Earth, atmosphere, ecosystems, and climate (www.pmel.noaa.gov).
PMIP	Paleoclimate Modelling Intercomparison Project — Hosted by LSCE, PMIP's purpose is to study the role of climate feedbacks arising for the different climate subsystems (www-pcmdi.llnl.gov/projects/model_intercomparison.php).

Acronym	Definition
PMP	Program for Climate Model Diagnosis and Intercomparison Metrics Package
PNNL	Pacific Northwest National Laboratory — DOE national laboratory in Richland, Wash., where multidisciplinary scientific teams address problems in four areas: science, energy, the Earth, and national security (www.pnnl.gov).
PoC	point of contact
POSIX®	Portable Operating System Interface — Family of standards specified by the IEEE Computer Society for maintaining compatibility between operating systems.
PROMS	Provenance Management System — A collection of tools and methodologies for managing provenance information.
PWT	ESGF Publishing Working Team
Python™	A programming language (www.python.org).
QA	quality assurance
QC	quality control
QCWT	ESGF Quality Control Working Team
Qualifying model	Model that meets certain criteria, including having performed (or having every intent to perform) certain required experiments.
RPM	Red Hat Package Manager
RVWT	ESGF Replication and Versioning Working Team
SAML	Security Assertion Markup Language
SDLC	system development lifecycle
Solr	Open-source enterprise search platform built on Lucene that powers the search and navigation features of many commercial-grade websites and applications (lucene.apache.org/solr/).
SSL	secure sockets layer
STFC	CEDA Science and Technology Facilities Council — Multidisciplinary science organization whose goal is to deliver economic, societal, scientific, and international benefits to the United Kingdom and, more broadly, the world (www.stfc.ac.uk/).
Synda	IPSL-developed software application for downloading data hosted by the distributed digital repositories of the ESGF data infrastructure.
TB	terabyte
TDS	THREDDS Data Server
THREDDS	Thematic Real-time Environmental Distributed Data Services — Web server hat provides metadata and data access for scientific data sets using a variety of remote data access protocols (www.dataone.org/software-tools/thematic-realtime-environmental-distributed-data-services-thredds/).
TRMM	Tropical Rainfall Measuring Mission — Joint mission between NASA and the Japan Aerospace Exploration Agency to study rainfall for weather and climate research (trmm.gsfc.nasa.gov).
UQ	uncertainty quantification — Method determining how likely a particular outcome is, given the inherent uncertainties or unknowns in a system (en.wikipedia.org/wiki/Uncertainty_quantification/).
User	Person accessing data via the ESGF (who is assumed to have an understanding of models and to possess technical skills typical of a Working Group I scientist).
UV-CDAT	Ultrascale Visualization–Climate Data Analysis Tools — Provides access to large-scale data analysis and visualization tools for the climate modelling and observational communities (uvcdat.llnl.gov).

Acronym	Definition
VDI	virtual desktop interface
WCRP	World Climate Research Programme — 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 (www.wcrp-climate.org).
WDAC	WCRP Data Advisory Council — Acts as a single entry point for all WCRP data, information, and observation activities with its sister programs and coordinates their high-level aspects across the WCRP, ensuring cooperation with main WCRP partners and other observing programs (www.wcrp-climate.org/WDAC.shtml).
Web portal	Point of access to information on the World Wide Web (en.wikipedia.org/wiki/Web_portal/).
WGCM	Working Group on Coupled Modelling — Fosters the development and review of coupled climate models. Activities include organizing model intercomparison projects aimed at understanding and predicting natural climate variability on decadal to centennial time scales and the response of the climate system to changes in natural and anthropogenic forcing (www.wcrp-climate.org/index.php/unifying-themes/unifying-themes-modelling/modelling-wgcm).
WIP	WGCM Infrastructure Panel — Serves as a counterpart to the CMIP panel and will enable modeling groups, through WGCM, to maintain some control over the technical requirements imposed by the increasingly burdensome MIPs (www.earthsystemcog.org/projects/wip/).
WPS	web processing service — Provides rules for standardizing inputs and outputs (requests and responses) for geospatial processing services (www.opengeospatial.org/standards/wps/).
WPWT	ESGF Workflow and Provenance Working Team
XML	Extensible Markup Language — A markup language that defines a set of rules for encoding documents in a format that is both human- and machine-readable (en.wikipedia.org/wiki/XML).