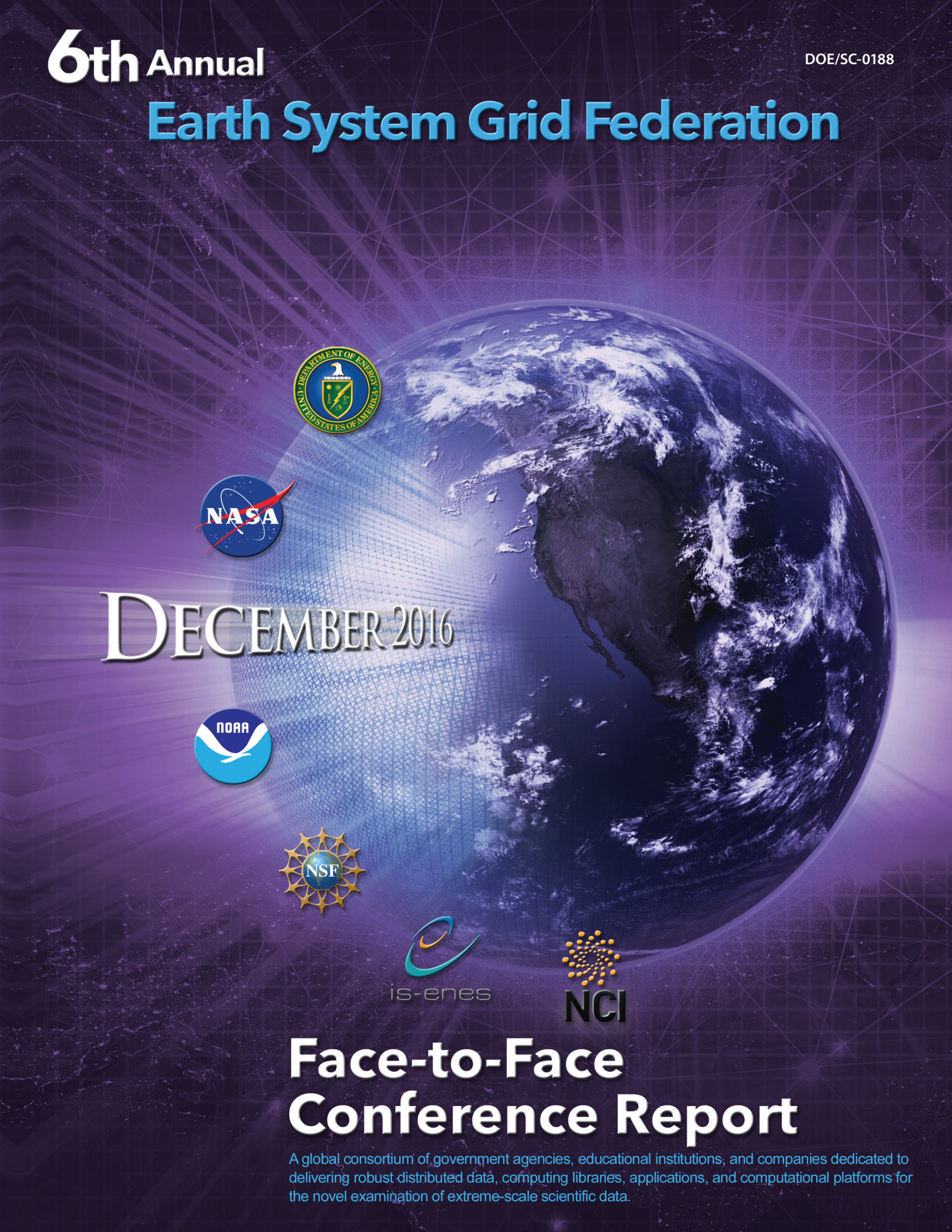
DOE/SC-0188

6th Annual Earth System Grid Federation



DECEMBER 2016

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.

6th Annual Earth System Grid Federation Face-to-Face Conference

December 5–9, 2016
Washington, D.C.

Convened by

U.S. Department of Energy (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)
National Computational Infrastructure (NCI) Australia

Workshop and Report Organizers

Dean N. Williams (Chair; DOE Lawrence Livermore National Laboratory)
Michael Lautenschlager (Co-Chair; German Climate Computing Centre)
Sébastien Denvil (Institut Pierre-Simon Laplace)
Luca Cinquini (NASA Jet Propulsion Laboratory)
Robert Ferraro (NASA Jet Propulsion Laboratory)
Daniel Duffy (NASA Goddard Space Flight Center)
Ben Evans (NCI)
Claire Trenham (NCI)

ESGF Steering Committee

Justin (Jay) Hnilo (DOE, U.S.)
Sylvie Joussaume (IS-ENES2, Europe)
Tsengdar Lee (NASA, U.S.)
Ben Evans (NCI, Australia)
Dean N. Williams (DOE, U.S.; Ex-officio member)

Suggested citation for this report: U.S. DOE. 2017. *6th Annual Earth System Grid Federation Face-to-Face Conference Report*. DOE/SC-0188. U.S. Department of Energy Office of Science.

DOE/SC-0188

6th Annual Earth System Grid Federation Face-to-Face Conference Report

December 5–9, 2016

Washington, D.C.

Publication Date: February 2017

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
National Computational Infrastructure Australia

Contents

Preface by ESGF Executive Committee Chair v

Executive Summary ix

1. Stakeholders' Summary by Steering Committee 1

2. User Survey Results (2016) 3

3. Usage Demographics (2016) 11

3.1 Statistics Overview 11

3.2 Registered Users and Downloads by Project 12

3.3 Downloads by Continent 12

3.4 Registered Users by Continent 12

3.5 Registered Users by Site 12

4. Conference Findings 17

4.1 CMIP6 Data Archive 17

4.2 CMIP6 Tier 1 Node 17

4.3 CMIP6 Tier 2 Node 17

4.4 Software Security 17

4.5 Server-Side Computing 18

4.6 Provenance Capture 19

4.7 Search and Metadata 19

4.8 Metrics 20

4.9 Survey 20

4.10 Modularity 20

4.11 Installation 21

4.12 Container Software 21

4.13 Data Replication and Test Federation 21

4.14 Network 22

4.15 Persistent Identifiers for Data Identification 23

4.16 Digital Object Identifiers 24

4.17 Training and Documentation 24

5. Scientific Challenges and Motivating Use Cases 25

5.1 Open Science Cloud Challenge 25

5.2 Data Challenge 25

5.3 Data Integration Challenge 25

5.4 Computational Environment Challenge 26

6. Computational Environments and Data Analytics 27

6.1 Framework for Collaborative Analysis of Distributed Environmental Data (CAFE) 27

6.2 Climate4Impact (C4I) 27

6.3 Climate Data Analysis Services (CDAS) 28

6.4 Community Data Analysis Tools (CDAT) 28

6.5 Ophidia 29

6.6 Power Analytics and Visualization for Climate Services (PAVICS) 29

6.7 Visualization Streams for Ultimate Scalability (ViSUS) 30

6.8 Growth Areas 30

7. Technology Developments **33**
7.1 Installation 33
7.2 Publishing Services 33
7.3 Search Services 34
7.4 User Interface 34
7.5 Security 34
7.6 Data Transfer, Network, and Replication 34
7.7 Computing Services 35
7.8 Metadata Services 35
7.9 Provenance Capture, Integration, and Usability 36
7.10 Services 36
8. Roadmap **37**
8.1 Short-Term Plans (0 to 2 Years) 37
8.2 Longer-Term Plans (2 to 5 Years) 38
9. Community Developments and Integration **41**
10. Report Summary and Development for 2017 ..... **43**
Appendices **45**
Appendix A. Conference Agenda **47**
Appendix B. Presentation, Demonstration, and Poster Abstracts **55**
Appendix C. ESGF's Current Data Holdings **77**
Appendix D. Conference Participants and Report Contributors **79**
Appendix E. Awards **83**
Appendix F. Acknowledgments **85**
Appendix G. Acronyms **87**

Preface by ESGF Executive Committee Chair

The Sixth Annual Face-to-Face (F2F) 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 5–9, 2016, in Washington, D.C.

The conference brought together more than 100 professionals from 17 countries to share their knowledge and experiences gained during the past year. The goals were to improve the usefulness of interagency software infrastructure development, explore ideas for new spin-off projects for enhancing the federation, and learn from one another in ways that can only happen face-to-face. Conference presentations (esgf.llnl.gov/2016-F2F.html) covered the state of ESGF, discussed development and implementation plans, focused on synergistic community activities, and outlined project deadlines (see **Appendix B**, p. 55). Special town hall discussion panels were held to address the specific needs of the community, which was well represented by the diverse backgrounds and expertise of participants, including climate and weather researchers and scientists, modelers, computational and data scientists, network specialists, and interagency program managers and sponsors. Also attending were researchers interested in incorporating interagency federated service approaches into their science domains such as biology and hydrology.

Posters, presentations, and panel discussions provided ample evidence of ESGF's resiliency in response to the 2015 security incident and other challenges. These events also showed the federation's dedication to build on and extend existing capabilities needed for large-scale data management, analysis, and distribution of highly visible community data and resources.

The past year was one of preparation and stabilization for ESGF, as well as for many of the projects that it supports. In 2016, the ESGF Executive Committee created several foundational living documents (see **Table 1**, p. vi) in anticipation of the receipt and processing of tens of petabytes of simulation and observational data in early 2017 as part of phase six of the Coupled Model Intercomparison Project (CMIP6). These documents will help focus and direct planning and execution for multiple national and international geoscience data projects. In the past, short-term ESGF strategic and work planning were guided by ESGF individual sponsor proposals and coordinated through ESGF F2F meetings, workshops, conferences, and reports. To develop a more cohesive, longer-term strategy, the ESGF Executive Committee in 2016 decided to provide more comprehensive strategic and implementation-oriented living documents that transcend individual sponsor requests. Previous documents and the release of the new documents listed in Table 1 have allowed working teams of ESGF developers to make strong and significant progress on all fronts of the software stack, thus helping ensure ESGF meets the growing needs of the climate community in the coming years. The five documents described in Table 1 highlight the significant amount of time that the ESGF Executive Committee devoted to the construction and development of this effort.

As part of its exciting and productive year, the ESGF community is especially proud of the preparatory work performed in 2016 in anticipation of CMIP6's massive simulation and observational data distribution effort. Conference discussions and presentations noted this achievement in particular, and feedback was extremely positive from participants who greatly enjoyed the chance to meet, network, and learn from like-minded people from many countries and organizations and to explore new and exciting ideas.

On behalf of everyone involved in the organization and production of the conference, I would like to express my sincere thanks to all ESGF developers, supporters, and associated project teams for their dedication to the vision of a successful data federation. The ESGF community holds the premier collection of simulations and observational and reanalysis data for climate change research. Moreover, ESGF is recognized as the leading infrastructure for the management and access of large distributed data volumes for climate change research.

Table 1. Living Documents, Policies, and Guidelines Developed by the Earth System Grid Federation (ESGF) Executive Committee

ESGF Living Document	Web Link	Description
Strategic Roadmap	esgf.llnl.gov/media/pdf/2015-ESGF-Strategic-Plan.pdf	The **ESGF Strategic Roadmap** describes the ESGF mission and an international integration strategy for data, database, and computational architecture, and for a stable infrastructure highlighted by the ESGF Executive Committee. These are key developments needed over the next 5 to 7 years in response to large-scale national and international climate community projects that depend on ESGF for success.
Implementation Plan	esgf.llnl.gov/media/pdf/ESGF-Implementation-Plan-V1.0.pdf	The **ESGF Implementation Plan** describes how the ESGF data management system will be deployed, installed, and transitioned into an operational system. It contains an overview of the system, a brief description of the major tasks involved in its implementation, the overall resources needed to support the implementation effort (such as hardware, network, software, facilities, materials, and personnel), and any site-specific implementation requirements.
Software Security Plan	esgf.llnl.gov/media/pdf/ESGF-Software-Security-Plan-V1.0.pdf	The primary purpose of the **ESGF Software Security Plan** is to ensure a systematic approach to releasing to the community a secure ESGF software stack (through both major and minor releases). Developed within the context of the ESGF Software Development Life Cycle (SDLC), this plan's emphasis is on the release phase of a typical SDLC and its prerequisites, which also depends on development and maintenance (design and build) aspects of the SDLC.
Policies and Guidelines	esgf.llnl.gov/media/pdf/ESGF-Policies-and-Guidelines-V1.0.pdf	The ESGF is composed of groups and institutions that have elected to work together to operate a global infrastructure in support of climate science research. Anyone is welcome to download, install, and run a copy of the ESGF software stack as a stand-alone node. However, joining the global federation requires understanding and abiding by **ESGF Policies and Guidelines**, which have been established to provide the best possible experience to the community, ensure security and stability, and facilitate consistent administration of the ESGF nodes.
ESGF Root Certificate Authorities (CA) Policy and Certificate Practices Statement	http://esgf.llnl.gov/media/pdf/ESGF-CA-V1.0.pdf	This document describes the set of rules and procedures established by the ESGF CA Policy Management Authority for the operation of the ESGF Root CA Public Key Infrastructure (PKI) services. The **Certificate Policy** (CP) describes the requirements for PKI operation and granting of PKI credentials as well as for lifetime management of those credentials. The **Certificate Practices Statement** (CPS) describes the actual steps that ESGF takes to implement the CP. These two statements taken together are designed to enable a Relying Party to read them and obtain an understanding of the trustworthiness of credentials issued by the ESGF Root CA.

This work would not be achievable without dedicated developers, ESGF's great user community, and the continued support of interagency sponsors: the Office of Biological and Environmental Research (BER) and Office of Advanced Scientific Computing Research (ASCR), both within the U.S. Department of Energy's (DOE) Office of Science; U.S. National Oceanic and Atmospheric Administration (NOAA), U.S. National Aeronautics and Space Administration (NASA), U.S. National Science Foundation (NSF), Infrastructure for the European Network for Earth System Modelling (IS-ENES), and the Australian National Computational Infrastructure (NCI). Support also comes from other national and international agencies.

I also once again want to thank everyone who attended the conference and gave so freely of themselves and their experiences to make our conference a memorable and successful event. The 2017 Seventh Annual ESGF Face-to-Face Conference will be held in the Washington, D.C., area and is eagerly awaited.

Best wishes to all,

Dean N. Williams

Chair, ESGF Executive Committee

Executive Summary

Since its inception in the late 1990s, the Earth System Grid Federation (ESGF) has evolved significantly as a system. As technology changed and its user base grew larger and more diverse, ESGF gradually developed into the state-of-the-art federated system it is today. Due to the enormity of its scope and the urgency of community needs, many features had to be tested in the real world during the expansion process. The rollout of Version 2.4 in 2016 culminated a series of releases and user feedback initiatives that began in 2013. During this 3-year period, ESGF hosted user group meetings to better understand how the system was being used; this information was then leveraged to begin Version 3.0 planning and preparations. The user meetings were collectively led and organized by the U.S. Department of Energy (DOE), the Infrastructure for the European Network for Earth System Modelling (IS-ENES), the U.S. National Aeronautics and Space Administration (NASA), the Australian National Computational Infrastructure (NCI), the U.S. National Oceanic and Atmospheric Administration (NOAA), and the U.S. National Science Foundation (NSF).

The information emerging from the user group meetings influenced requirements, development, and operations. In 2015, representatives from a significant fraction of projects that use ESGF to disseminate and analyze data attended the Fifth Annual ESGF Face-to-Face (F2F) Conference (DOI: 10.2172/1253685). Attendees provided important feedback regarding current and future community data use cases. Discussions focused on maintaining essential operations while developing new and improved software to handle ever-increasing data variety, complexity, velocity, and volume—a task accomplished by the entire consortium (see **Table 2**, p. x).

The most recent conference was held in December 2016 in Washington, D.C. It focused on federation resiliency and reaffirmed the consortium's dedication to extend the existing capabilities needed for large-scale data management, analysis, and distribution of highly visible community data and managed resources (See **Appendices A**, p. 47; **B**, p. 55; **C**, p. 77; and **D**, p. 79.

The federation works across multiple worldwide data centers and spans seven international network organizations to provide users with the ability to access, analyze, and visualize data through a globally federated collection of networks, computers, and software. Its architecture employs a series of geographically distributed peer nodes that are independently administered and united by common federation protocols and application programming interfaces (APIs). The full ESGF infrastructure has been adopted by multiple Earth science projects and allows access to petabytes of geophysical data. These projects include the Coupled Model Intercomparison Project (CMIP), whose output is to be used in the upcoming Intergovernmental Panel on Climate Change's (IPCC) Assessment Reports; multiple model intercomparison projects (MIPs) endorsed by the World Climate Research Programme (WCRP); and the Accelerated Climate Modeling for Energy (ACME) project, which leverages ESGF in its overarching workflow process to store model output. ESGF is a successful example of integrating disparate open-source technologies into a cohesive functional system that serves the needs of the global climate science community.

The success of ESGF can be seen in its associated numbers. As the first and only federated climate archive of its kind, ESGF supports more than 40 projects, including data from 25 worldwide climate research centers spanning 21 countries. Virtually all climate researchers have used ESGF directly or indirectly. The system hosts over 25,000 users, and its petabyte data archive is directly linked to thousands of peer-reviewed climate journal publications. An example of an ESGF gateway can be viewed at the DOE Lawrence Livermore National Laboratory (LLNL) ESGF node site (pcmdi.llnl.gov; see **Fig. 1**, p. xi).

The rollout of the latest ESGF version was accompanied by a transfer of power in the way federation development is governed (i.e., from principal investigators to the multiagency Executive Committee). ESGF was designed to be an open-source software stack that takes advantage of open-source tools, enabling projects, modelers, and researchers to customize and include components of ESGF for their specific

Table 2. ESGF Designated Working Teams		
Team	**Lead Partner Institutions***	**Purpose**
CoG User Interface	NOAA ESRL, Luca Cinquini	Improve ESGF search, data cart management, and interface
Compute	DOE LLNL, Charles Doutriaux NASA GSFC, Daniel Duffy IS-ENES CMCC, Sandro Fiore	Manage server-side computing and develop data analytics capability within ESGF
Data Transfer	DOE ANL, Lukasz Lacinski	Enhance ESGF data transfer and web-based downloads
Dashboard and Stats	IS-ENES CMCC, Sandro Fiore	Monitor ESGF system metrics and data usage statistics
Documentation	DOE LLNL, Matthew Harris	Document the use of the ESGF software stack
Identity Entitlement Access Management	IS-ENES CEDA, Philip Kershaw DOE ANL, Rachana Ananthakrishnan	Implement ESGF X.509 certificate-based authentication and improved interface
Installation	IS-ENES NSC, Prashanth Dwarakanath IS-ENES IPSL, Sébastien Denvil DOE LLNL, William Hill	Install components of the ESGF software stack
International Climate Network Working Group	DOE LBNL, Eli Dart DOE LLNL, Dean N. Williams	Increase data transfer rates between ESGF climate data centers
Metadata and Search	NASA JPL, Luca Cinquini	Implement discoverable metadata and ESGF search engine based on Solr5
Node Manager	DOE LLNL, Sasha Ames IS-ENES NSC, Prashanth Dwarakanath	Manage ESGF nodes and node communications
Persistent Identifier Services	IS-ENES DKRZ, Tobias Weigel IS-ENES DKRZ, Stephan Kindermann	Establish fundamental persistent identifier features in preparation for CMIP6
Provenance Capture	DOE PNNL, Bibi Raju	Enable ESGF provenance capture for reproducibility and repeatability
Publication	DOE LLNL, Sasha Ames DOE ANL, Lukasz Lacinski	Enable capability for publishing to ESGF the datasets from CMIP and other projects
Quality Control	IS-ENES DKRZ, Martina Stockhause NCI, Claire Trenham	Provide quality control for data and integrate external information into the ESGF portal
Replication and Versioning	IS-ENES DKRZ, Stephan Kindermann	Create replication tool for moving data from one ESGF center to another. Preserve versioning history of ESGF published datasets
Software Security	NASA GSFC, George Rumney	Implement security measures to identify vulnerabilities in the ESGF software and provide continuous improvement to the ESGF Software Development Life Cycle
Support	IS-ENES DKRZ, Torsten Rathmann DOE LLNL, Matthew Harris	Provide user support and develop frequently asked questions regarding ESGF and housed data
Tracking and Feedback Notification	DOE LLNL, Sasha Ames IS-ENES CMCC, Sandro Fiore	Implement user and node notification of changed data in the ESGF ecosystem
*Note: For a full list of acronyms, please see **Appendix G**, p. 87		

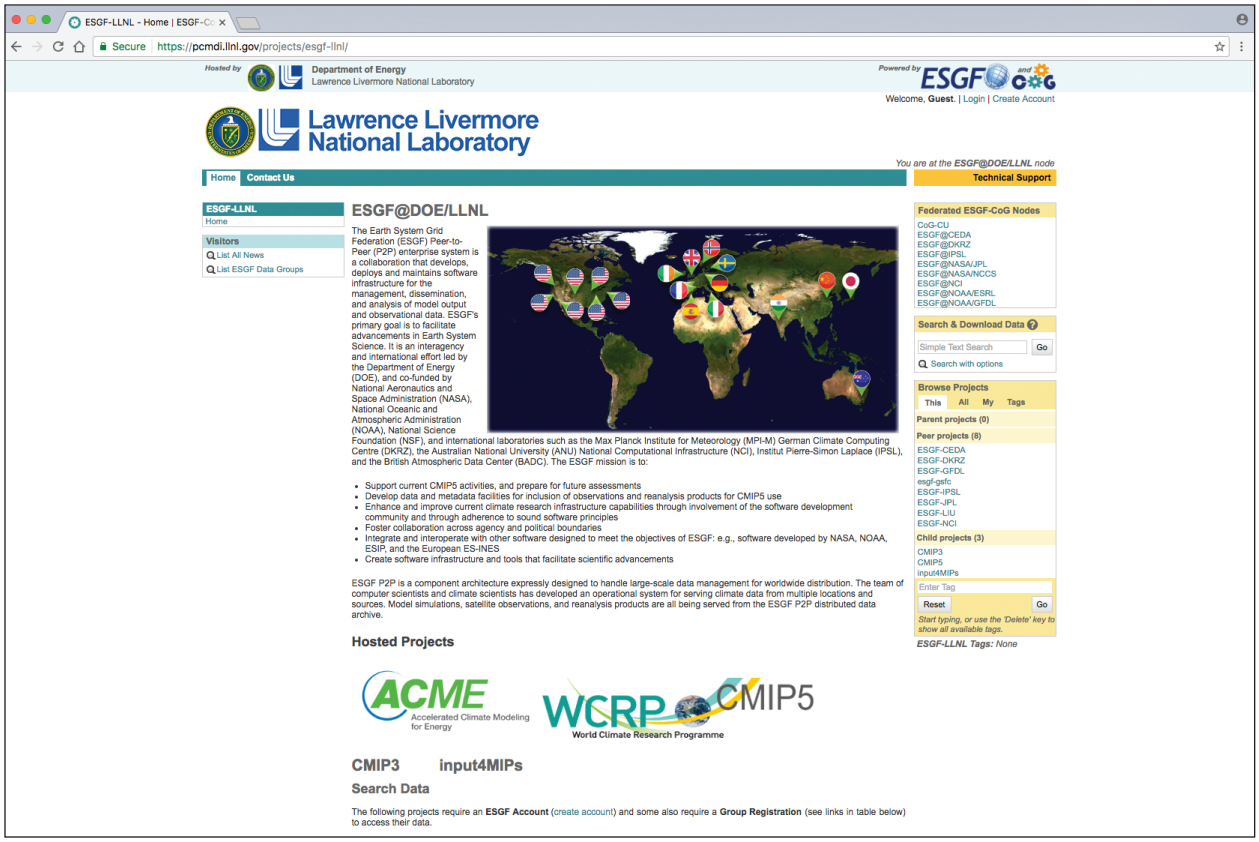


Fig. 1. The ESGF gateway hosted at DOE's Lawrence Livermore National Laboratory provides access to data products for the Accelerated Climate Modeling for Energy and Coupled Model Intercomparison Project.

needs. This design provides an impartial platform for handling intricate but common tasks (e.g., security, logging, configuration management, output handling, and provenance capture), an object-oriented API for developing a new user interface (shown **in Fig. 1**, this page), and integration management for facilitating simpler workflows.

ESGF portals are gateways to scientific data collections hosted at sites around the globe. The portals allow users to register[1] and potentially access all data and services within ESGF. Currently, there are more than 40 portals, including several at LLNL (see **Fig. 1**, this page). Over the course of several years, ESGF has successfully developed and merged dozens of independent software applications to:

- Integrate more than 60 climate model and 30 measurement archives from national and international organizations.
- Share resources across agencies for high-performance computing (HPC) and storage.
- Move tens of petabytes of data across an international network infrastructure.
- Create an infrastructure for national and international model and data intercomparison studies.
- Analyze and visualize large, disparate climate simulation and observational datasets from around the world.

ESGF's primary objectives are to:

- Develop efficient, community-based tools to obtain relevant meteorological and other observational data.
- Develop custom computational models and export analysis tools for climate change simulations, such as those used in IPCC reports.

[1] Users have a shared identity and authentication to all sites; thus, they register only once to gain access.

ESGF enables international climate research teams to work in highly distributed research environments while using unique scientific instruments, petascale-class computers, and extreme volumes of data. Key to ESGF's success is its ability to effectively and securely catalog, disseminate, and analyze research results in a globally federated environment. For example, new results generated by one team member can be made immediately accessible to other team members, who can annotate, comment on, and otherwise interact with those results.

As discussed at the 2016 ESGF F2F conference, governance and use cases determine how project and data requirements affect operations and software development. Therefore, with encouragement from many supporting funding agencies, representatives from numerous projects using ESGF to disseminate and analyze data attended the conference to provide feedback regarding current and future community use cases. Discussions focused on maintaining essential operations while developing new and improved software to handle ever-increasing data variety, complexity, velocity, and volume. This report, along with a series of conference presentations (esgf.llnl.gov/2016-F2F.html), summarizes data use cases for computing and the data science activities that are critical to the community meeting its scientific mission (both as individual projects and as a federation).

Although progress has been made on numerous findings from the 2015 Fifth Annual ESGF F2F Conference Report,[2] a number of the same issues still persist, including data storage, server-side analysis, replication, large data transfers, and software test suites. That said, many other findings, such as routine software security scans, use metrics, and search services that include controlled vocabularies, are well on their way to being resolved. In this year's report, conference participants highlighted their top needs for infrastructure investments:

1. **Handling of CMIP6 Data:** Agreed-upon strategy on whether to compress the CMIP6 data expected to begin arriving in late March 2017. Estimated size of the CMIP6 archive will be 15 petabytes of compressed NetCDF data or 30 petabytes of uncompressed NetCDF data. Also unclear is whether this volume of data can be stored and analyzed at or transmitted to any one node or a series of core nodes.
2. **Tier 1 and Tier 2 Nodes:** Documentation describing Tier 1 and Tier 2 nodes for setting development priorities and eventual CMIP6 operations. Defining these nodes not only will affect short- and long-term ESGF software development, but also the purchase of hardware and network infrastructure over the coming years. In addition, the documentation defines the minimum requirements to participate in the federation.
3. **Software Security:** Continued vigilance following ESGF's apparent overall recovery from its 2015 security challenges. Security efforts are needed to maintain, improve, and work toward a robust, routine automated software security scanning process, which includes risk assessments of the code base.
4. **Server-Side Computing:** Easier process for scientists to download only their needed data portions rather than entire large datasets (i.e., subsetting, averaging, and regridding). These capabilities will be sufficient for 70% to 80% of all user server-side computing requests. In addition, resource management strategies must be in place to accommodate and prioritize the tens of thousands of ESGF users.
5. **Provenance Capture:** Matured capabilities to capture provenance information for component integration and usability by projects and the user community. For example, there is a need for the provenance environment to easily reproduce server-side analyses and products for users requesting the results of work entered into reports or journal articles.
6. **Search and Metadata:** Customized searches on metadata that can be saved and shared across the federation as part of efforts that go beyond including the searching of controlled vocabularies.
7. **Metrics:** Dashboard capable of displaying hard metrics (e.g., number of users, number of downloads, size of current archive). In the 2017 release of ESGF, in time for CMIP6, the dashboard team should incorporate into the dashboard different metrics, such as how many server-side subsets or

[2] U.S. DOE. 2016. 5th Annual Earth System Grid Federation Face-to-Face Conference Report DOE/SC-0181. U.S. Department of Energy Office of Science. DOI: 10.2172/1253685.

averages took place and on which machines, as well as the ability to track user resource allocations.

8. **Survey:** Possibly, more surveys to help capture the needs of projects, operations, and the community more broadly.
9. **Modularity:** ESGF open-source tools and interfaces made available to external groups and projects. Under the DOE Distributed Resources for ESGF Advanced Management (DREAM) project, efforts are being made to develop a component modularity, which can be made accessible to the ESGF community through APIs. The Birdhouse application developed by DKRZ (German Climate Computing Centre) is an example of using ESGF tool components for other purposes.
10. **Installation and Docker:** Easier installation of ESGF software. Continuing to address this challenge are various working teams. Part of the task includes having the software work within a secure software container (e.g., Docker) that meets the federation's software security concerns.
11. **Data Replication and Network:** Substantial progress for data replication in 2017 to prepare for and meet the needs of CMIP6. This effort includes the integration of Synda and the International Climate Network Working Group (ICNWG) for high-speed data transfers between Tier 1 and possibly Tier 2 nodes.
12. **Persistent Identifiers (PIDs):** Development of versioning and citation tools and services (e.g., errata and "unpublish" features). Progress is being made, and these capabilities must be in place for CMIP6.
13. **Test Platform:** ESGF test platform to investigate how tools perform on different nodes and to check performance between nodes before new tools and features go live. In this concept, every new development would run through the test environment before being released.
14. **Documentation:** Needed for all components for software developers, projects, and user communities.

1. Stakeholders' Summary by Steering Committee

The Earth System Grid Federation (ESGF) is a unique international infrastructure that makes very large reference datasets available to scientists around the world. The near-term focus for ESGF is to ensure that the infrastructure is prepared for the sixth Coupled Model Intercomparison Project (CMIP6), which is supported by the World Climate Research Programme (WCRP), as well as other large model intercomparison projects (MIPs). Access to this data is essential to understanding the performance of climate models and their projections for long-term climatic conditions.

New approaches are needed to permit scalable data management and ease of access. With data volumes reaching exabyte scale in the coming years, a federated data system is needed that can handle hundreds of times more data than what is handled today. Such a system would allow data providers to maintain a set of geographically dispersed nodes accessible by the scientific community. Through linkages, nodes would act to harmonize different dataset archives, enabling scientists to access selected datasets, such as monthly averaged surface temperature and precipitation, as if the information were on their own systems.

As data volume and requirements continue to grow, the infrastructure must respond with improved robustness and security. ESGF is meeting this challenge with better practices for federated data archiving and dissemination and an increased focus on both its testing frameworks and state-of-the-art publishing procedures. The increasing need for *in situ* and server-side processing, along with data storage capacity, also provides new challenges in resource management.

Managing and analyzing such large volumes of data present other challenges for the technical infrastructure. As data volumes have overtaken the ability to easily download data, the focus has moved to co-locating computational processing power with the data, within virtual laboratories or cloud infrastructures, and through increasingly intelligent server-side processing and Web Processing Services (WPS). Such an integrated system also can link analysis software, multiscale and multidimensional visualization, and high-performance computing (HPC) as part of the federated data analytics, thus offering a unique collaborative work environment that currently is not widely available. There is an increasing need to enable scientists to rapidly prototype new algorithms by intercomparing various (experimental or observational) data to, for instance, formulate new model representations, develop sophisticated representations of uncertainty quantification (UQ), or perform intelligent pattern recognition.

For the near term, multidimensional data analysis and visualization techniques are needed desperately to study, for example, processes acting simultaneously on small-scale cloud physics and macroscale climate dynamics. Statistical data analytics, machine learning, and inference are central to virtually all scales of data analytics in the biological sciences and climate studies. Incorporation of these analysis methods would allow linkages to UQ and modeling frameworks to enable integrated analysis and comparison of data from multiple modalities and across experimental conditions.

In the future, better approaches will be needed to more easily understand multidata patterns that will emerge. A current limitation is the inability to merge disparate data forms. By using next-generation data analysis, deep learning, and visualization techniques alongside a programming construct for rapid formulation prototyping, ESGF anticipates an acceleration of systems-level understanding of biological and Earth systems. This understanding, for example, could be used to explain how genomic information is translated to functional properties of cells, cellular communities, and plants. The interpretation of multiscale data requires development of truly interactive data analytic and visualization frameworks.

The annual meeting provides an important part of the planning process to carefully review each of these components and develop plans to ensure that ESGF is better prepared for current and future demands, such as in the following areas:

1. The Office of Biological and Environmental Research (BER) within the U.S. Department of Energy (DOE) Office of Science possesses

a large volume of heterogeneous data that, when integrated, will accelerate scientific discovery and become a foundational archive for interdisciplinary data mining, data analysis and visualization, and predictive modeling. Many BER programs support scientific user facilities with integrated experimental and computational functions. However, dramatic improvements in technologies and analytic methodologies during recent years have shifted the bottleneck in scientific productivity from data production to data management, data interpretation, and scientific visualization.

2. Australia's National Computational Infrastructure (NCI) has a large repository of national reference data collections across the full spectrum of climate, weather, water, and solid Earth science. These datasets have been assembled in an integrated high-performance data analysis and HPC facility that includes a dedicated high-performance OpenStack cloud and top-100 supercomputer. As the leading entity in Australia for computation and data research across governmental and academic communities, NCI has an ongoing need to ensure that data management and analysis capabilities developed locally or at international peer centers are made available across the ESGF Tier 1 sites.

3. The U.S. National Aeronautics and Space Administration (NASA) Earth Science Division continues integrating large-scale modeling and observational data into MIP processes to perform server-side data analytics, pattern recognition, and visualization and to constrain Earth system model representations of natural processes. The coupling of HPC with an advanced data analytics platform including tens of petabytes of modeling and observational data allows scientists to quickly analyze experimental results to quantify model uncertainties and correlate small-scale physical processes to macroscale Earth system behaviors.

4. The European Network for Earth System modelling (ENES) encompasses the European community working on climate modeling and studying climate variability and change. Its infrastructure, IS-ENES, supported by two consecutive European projects (2009–2017), supports the European contribution to WCRP coordinated numerical experiments, including CMIP and the Coordinated Regional Climate Downscaling Experiment (CORDEX). IS-ENES, through national and European support, strongly supports ESGF and ES-DOC, which are crucial for the dissemination of WRCP model results. IS-ENES is strongly engaged in ESGF installation, quality control, replication, data citation, and monitoring. To ease provision of data to the climate change impact communities, IS-ENES develops and supports the climate4impact.eu platform services and portal. IS-ENES will continue to support ESGF through collaboration between the European groups, supported by national and some European projects, with coordination by its ENES data task force. The ENES community considers this role essential to ensuring open access to data for climate research but also for climate impacts studies and climate services, in particular through collaboration with the European Copernicus Climate Change Service and the European Open Science Cloud, both under development.

2. User Survey Results (2016)

Prior to the ESGF Face-to-Face (F2F) Conference, the ESGF Executive Committee conducted an online survey of data providers and consumers supported by ESGF. The intent of the survey was to provide the ESGF community of developers with anonymous feedback about how ESGF can improve its core services and to ascertain what scientists believe are the greatest strengths and weaknesses of the ESGF enterprise. The Executive Committee distributed the survey via mailing lists associated with ESGF projects, resulting in a representative sampling of geographically and topically diverse responders. Descriptive results from the global survey attempt to shed light on the data needs of national and international projects. Action items generated from the survey results are intended to bridge the gap between short- and long-term development and operations.

For this survey, the request for feedback went out to (1) several WCRP-endorsed MIPs including CMIP, the Atmospheric Model Intercomparison Project (AMIP), the Observations for Model Intercomparisons (Obs4MIPs) and Input4MIPs; (2) CORDEX; (3) the Accelerated Climate Modeling for Energy (ACME) project; and (4) the Collaborative REAnalysis Technical Environment Intercomparison Project (CREATE-IP). Most questions asked researchers to rate on a scale from 1 to 6 their need for a specific support or service; 1 indicated little or no interest, while 6 indicated high interest or need. Other questions required yes-or-no responses. The survey also presented open-ended questions. Weighted average values were calculated for each question across all responses (e.g., a value of 4.54 for a particular topic would indicate that most participants for that question would rate the topic as being of high or very high interest). Also calculated was the percentage of participants that gave a topic a particular rank (e.g., 37% ranks as a very high response). Merging the weighted average with the percentage of responses gave yet another perspective on the value of the survey response (e.g., 1.49 constitutes a very high community interest, taking into consideration the combined weighted average and the percentage of participants).

Respondents were asked to identify themselves as a data provider, data consumer, or both (see **Fig. 2**, this page). The survey also asked the respondents to best describe their profession (e.g., undergraduate or graduate student, postdoctoral scholar, academic scientist, governmental scientist, or other) and type of affiliation (e.g., governmental agency, university, and the private sector). See **Figs. 3** and **4**, this page and p. 4, respectively. Linux was the most commonly used platform among the respondents, followed by Windows and Mac OSX (see **Fig. 5**, p. 4).

Which of the following best describes your interest in ESGF data?

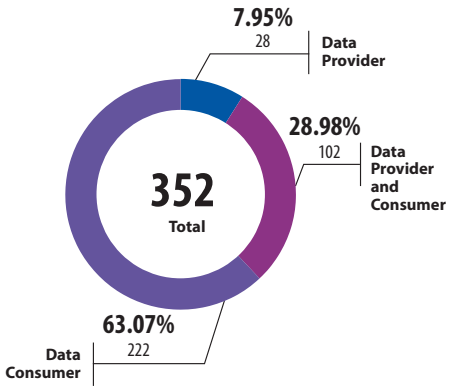


Fig. 2. Survey Question: Data generation and use.

Which of the following best describes you professionally?

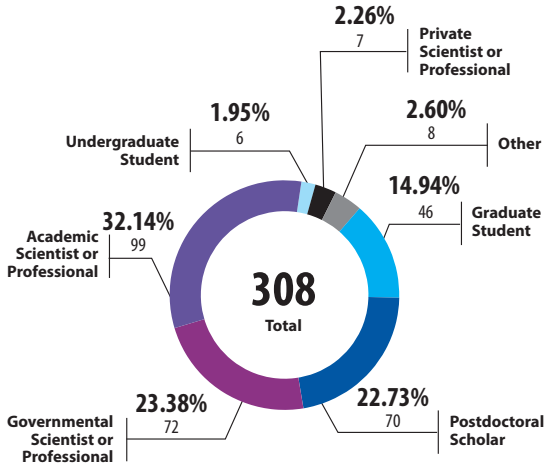


Fig. 3. Survey Question: Profession.

The bulk of the survey consisted of 42 questions listed under several subcategories that asked respondents to rate the importance of the service or tool. These subcategories included:

- User interface (UI) (websites, CoG).
- Ingestion of and access to large volumes of scientific data (i.e., from data archive to supercomputer and server-side analysis).
- Web documentation.
- Improved UI designs and principles to enable easier access to computer and software capabilities (e.g., recommendation systems, more flexible and interactive interfaces).
- Distributed global search.
- Unified data discovery for all ESGF data sources to support research.
- Quality control (QC) algorithms for data.
- Reliability and resilience of resource.
- Data access and usage.
- Remote computing capability.
- Data transport.
- QC issues.

The first step in evaluating the responses was to list the subcategories in terms of need on a scale of 1 to 6:

- 10 of the subcategories earned an average response rating of 4.1 or higher.
- 17 earned between 4.1 and 3.7.
- Remaining responses earned less than 3.6.

Assuming that a higher number of responses translated into a higher priority, the next step was to weight these subcategories by the number of survey responses. For those 10 subcategories with an average rating of 4.1 or higher:

- 6 questions had more than a 30% response rate.
- 21 had less than a 20% response rate.
- Remaining 15 had a 20% to 30% response rate.

Table 3, p. 5, shows the top needs identified in this survey using a weighted metric of the measured need multiplied by the percentage of responses.

Which of the following best describes your affiliation?

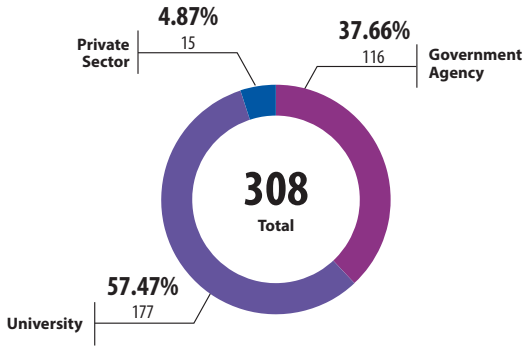


Fig. 4. Survey Question: Affliation.

Which operating system or platform do you use to download and analyze data?

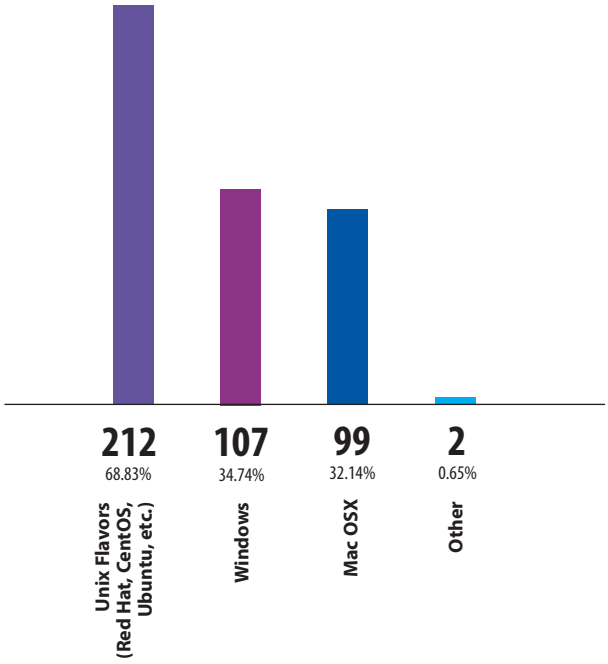


Fig. 5. Survey Question: Operating system. (Respondents could choose more than one answer.)

These results were:

- 3 responses between 1.46 and 1.49.
- 7 responses between 1.20 and 1.28.
- 9 responses between 1.10 and 1.17.

Table 3. Top Needs Identified by the Survey

Question	Question Type	Category	Total Response	Weighted Score	Percentage of Response	Combined Weighted Score and Percentage of Response
Which feature of ESGF do you find most difficult to use? Which needs the most improvement?	Needed ESGF improvements	UI (websites, CoG)	135	3.78	39.36%	1.49
How important are knowledge gathering, managing, and sharing?	Needed capabilities identified by the community	Ingestion of and access to large volumes of scientific data (i.e., from data archive to supercomputer and server-side analysis)	119	4.22	34.69%	1.46
Which feature of ESGF do you find most difficult to use? Which needs the most improvement?	Needed ESGF improvements	Web documentation	129	3.89	37.61%	1.46
How good are human-computer interactions?	Needed ESGF improvements	Improved UI designs and principles to enable easier access to computer and software capabilities (e.g., recommendation systems, more flexible and interactive interfaces)	106	4.13	30.90%	1.28
Which feature of ESGF do you find most difficult to use? Which needs the most improvement?	Needed ESGF improvements	Distributed global search	118	3.67	34.40%	1.26
How important are rapid information retrieval, knowledge-based response, and decision-making mechanisms?	Needed capabilities identified by the community	Unified data discovery for all ESGF data sources to support research	99	4.35	28.86%	1.26
How important are knowledge gathering, managing, and sharing?	Needed capabilities identified by the community	QC algorithms for data	110	3.86	32.07%	1.24

Table continued next page

Table 3. Top Needs Identified by the Survey

Question	Question Type	Category	Total Response	Weighted Score	Percentage of Response	Combined Weighted Score and Percentage of Response
How useful is user support?	Features that the community finds most useful	Data access and usage	101	4.21	29.45%	1.24
Which features of ESGF do you find most useful?	Features that the community finds most useful	UI (websites, CoG)	100	4.22	29.15%	1.23
Which features of ESGF do you find most useful?	Features that the community finds most useful	Distributed global search	91	4.55	26.53%	1.21
Is resource management needed?	Needed capabilities identified by the community	Reliability and resilience of resources	97	4.25	28.28%	1.20
How important are knowledge gathering, managing, and sharing?	Needed capabilities identified by the community	Interoperability: Interfaces that ensure a high degree of interoperability (formats and semantic level) among repositories and applications	107	3.77	31.20%	1.17
How important are rapid information retrieval, knowledge-based response, and decision-making mechanisms?	Needed capabilities identified by the community	Availability of ancillary data products such as data plots, statistical summaries, data quality information, and other documents	104	3.84	30.32%	1.16
Is resource management needed?	Needed capabilities identified by the community	Access to sufficient observational and experimental resources	99	4.01	28.86%	1.16
Is resource management needed?	Needed capabilities identified by the community	Awareness and information availability of these resources	98	4.04	28.57%	1.15
Is resource management needed?	Needed capabilities identified by the community	Access to enough computational and storage resources	99	3.89	28.96%	1.12

Table continued next page

Table 3. Top Needs Identified by the Survey

Question	Question Type	Category	Total Response	Weighted Score	Percentage of Response	Combined Weighted Score and Percentage of Response
How good are human-computer interactions?	Needed ESGF improvements	Environments that support more effective collaboration and sharing within and among science teams (e.g., collaboration tools)	105	3.63	30.61%	1.11
How useful is user support?	Features that the community finds most useful	Data sharing	96	3.96	27.99%	1.11
Which feature of ESGF do you find most difficult to use? Which needs the most improvement?	Needed ESGF improvements	Direct data delivery into ESGF computing systems from distributed data resources	95	3.99	27.70%	1.10
How important are rapid information retrieval, knowledge-based response, and decision-making mechanisms?	Needed capabilities identified by the community	Reproducibility	106	3.57	30.90%	1.10
How useful is user support?	Features that the community finds most useful	Data publishing	95	3.89	27.41%	1.10

This spread indicates that ESGF users have diverse needs and priorities.

Roughly 40% of responses with a combined weighted score of 1.49 indicated that the ESGF UI (i.e., the website or CoG) was the most difficult feature to use and needs improvement. About 35% of responses, with a combined weighted score of 1.46, pointed to the need for sufficient access to large volumes of data with computational resources for server-side (i.e., remote) analysis and visualization. Also notable at a combined weighted score of 1.46 was the emphasis on better, more reliable online documentation. Related to these changes, respondents requested an environment that supports more effective collaboration and sharing within and between science teams (e.g., collaborative tools), at a combined weighted score of 1.11. Of relevance to efforts to design a more integrated data and computing infrastructure was the finding that most respondents access data and compute resources via web interfaces or remote login along with application programming interfaces (APIs).

The question identified as the area of greatest need overall was "How important is knowledge gathering, managing, and sharing?" All questions in this category were rated less than 4.06 but higher than 3.8; no other category had such a high average. The topics included:

- Direct data delivery into ESGF computing systems from distributed data resources—3.99/27.7%.
- Data sharing—3.96/27.99%.
- Web documentation—3.89/37.61%.

- Data publishing (long-tail publishing for individual scientists)—3.89/27.41%.
- QC algorithms for data—3.86/32.07%.
- Ancillary data products (e.g., data plots, statistical summaries)—3.84/30.32%.

A question raising significant interest among the survey participants was "How good are human-computer interactions?" Respondents identified collaborative environments, in particular, as a key requirement (3.63). The new ESGF mandate regarding data management and sharing clearly has penetrated the community and raises questions for many, as evident by high scores for several related survey topics:

- Easy way to publish and archive data using one of the ESGF data centers—3.89/27.41%.
- User support for data access and download—4.21/29.45%.
- Access to enough computational and storage resources—3.89/28.96%.

Other questions focused on the effective use of exascale systems received very mixed results, pointing to a potential need for more community education.

A different independent survey found strong interest in tools that would facilitate and improve analysis. In some ways, this finding is supported by responses from the ESGF survey that indicate analysis tools are currently among the least useful ESGF features (see **Table 4,** this page, from this year's survey). This table shows that the use of the Live Access Server (LAS) analysis and visualization engine for data is least understood (weighted average 4.12).

Datasets submitted to ESGF are subject to stringent procedures before official publication. As the community vets the data, a digital object identifier (DOI) is eventually assigned. If the data need to be changed for any reason, then the ESGF notification service alerts users about the modifications. If users download data from secondary repositories, they are at risk of retrieving old or outdated data without knowledge of any updates. For these reasons, the federation recommends that users access ESGF data directly from ESGF-supported nodes. Encouragingly,

Table 4. Usefulness of ESGF Features

	1 (Least useful)		2		3		4		5		6 (Most useful)		Never Used		Total	Weighted Average
UI (websites or CoG)	10.6%	16	17.88%	27	13.25%	20	14.57%	22	13.25%	20	22.52%	34	7.95%	12	**151**	**3.76**
Distributed global search	8.84%	13	14.29%	21	19.05%	28	10.88%	16	13.61%	20	15.65%	23	17.69%	26	**147**	**3.64**
Web documentation	7.95%	12	11.92%	18	17.22%	26	17.22%	26	11.92%	18	21.85%	33	11.92%	18	**151**	**4.89**
Globus download (currently available only for a few datasets)	2.07%	3	2.76%	4	4.14%	6	4.14%	6	8.28%	12	12.41%	18	66.21%	96	**145**	**4.51**
Synda download client	6.94%	10	3.47%	5	1.39%	2	3.47%	5	4.17%	6	6.25%	9	74.31%	107	**144**	**3.51**
LAS analysis and visualization engine	2.76%	4	2.76%	4	5.52%	8	4.83%	7	4.83%	7	8.97%	13	70.43%	102	**145**	**4.12**
User support	5.37%	8	11.41%	17	10.74%	16	14.77%	22	8.72%	13	10.74%	16	38.26%	57	**149**	**3.68**

Where do you access data published to ESGF?

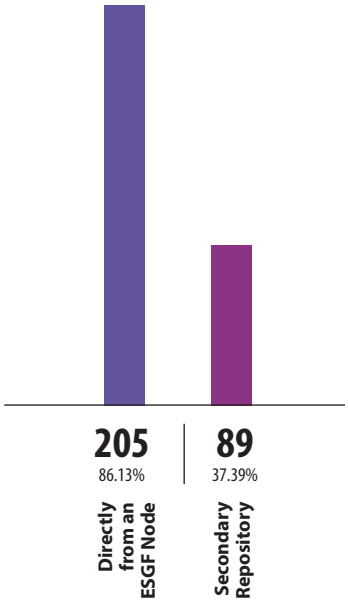


Fig. 6. Survey Question: Data access. (Respondents could choose more than one answer.)

How much ESGF data do you use for your work?

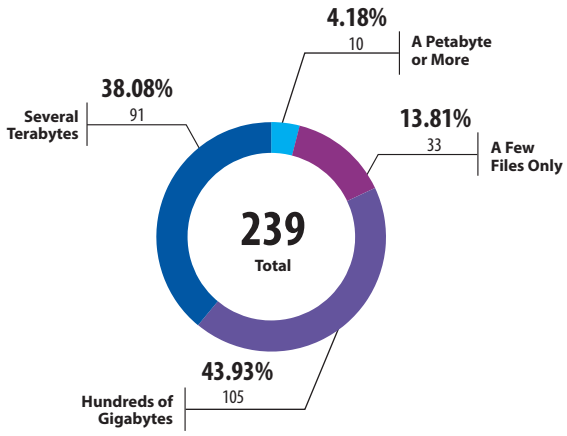


Fig. 7. Survey Question: Data quantity.

the survey shows 86.13% of respondents access ESGF nodes directly for data retrieval (see **Fig. 6**, this page).

Data providers want to know how many people use their data and for what purposes. With so many users, ESGF recognizes the need to monitor its data and count gigabytes, terabytes, and petabytes downloaded and accessed. To understand the ESGF data landscape, a survey question examined how much data users consume for their work. The federation found that the amount of data used for climate research, on average, is in the hundreds of gigabytes nearly 44% of the time and terabytes 38.08% of the time (see **Fig. 7**, this page).

The ESGF data service is designed to help users find data as easily as possible. However, as with all information search tools, knowing how to maximize the system's search functionality is useful. Tutorials and documentation detail how to find data through the ESGF CoG federated websites using both straightforward approaches and more advanced searches. When a user wants to find specific research data, a good place to start is by entering the variable name along with specific attributes. Users may search the variable throughout the federation nodes or on the local node only. If a variable is not found, the search needs further refinement. **Table 5**, this page, shows the average time needed to discover and download data (i.e., about half of the respondents can discover data in a minute, while others took considerably more time. Data are typically downloaded in 1 to 24 hours).

Dividing work into components (i.e., data, computer, storage, and software) is easy enough for a user, but putting together individual submissions to create a workflow is not. Data discovery, compute resource selection, data manipulation, derived data storage site selection, and software selection at each stage of the workflow are challenging at best. Minimizing the time

Table 5. Survey Question: How Long Does it Take You on Average to Discover and Access the Data and Resources You Need?

	Minutes		Hours		Days		Can't Find or Access		Total	Weighted Average
Discover	49.53%	105	36.79%	78	10.38%	22	3.30%	7	**212**	**1.67**
Access or Download	12.92%	27	42.11%	88	41.63%	87	3.35%	7	**209**	**2.35**

Table 6. Survey Question: Which Takes the Longest to Discover and Use?										
	1 (Shortest)		2		3		4 (Longest)		Total	Weighted Average
Data	16.96%	29	22.22%	38	25.15%	43	35.67%	61	**171**	**2.80**
Computer	31.01%	49	49.37%	78	14.56%	23	5.06%	8	**158**	**1.94**
Storage	21.52%	34	36.08%	57	25.95%	41	16.46%	26	**158**	**2.37**
Software	24.53%	39	34.59%	55	20.75%	33	20.13%	32	**159**	**2.36**

Table 7. Survey Question: How important are knowledge gathering, managing, and sharing?														
	1 (Least need for support)		2		3		4		5		6 (Most need for support)		Total	Weighted Average
Ingestion of and access to large volumes of scientific data (i.e., from data archive to super computer)	7.76%	9	10.34%	12	12.93%	15	20.69%	24	18.10%	21	30.17%	35	**116**	**4.22**
Interop-erability: Interfaces that ensure a high degree of interop-erability (formats and semantic level) among repos-itories and applications	2.88%	3	18.27%	19	22.12%	23	25.00%	26	21.15%	22	10.58%	11	**104**	**3.75**
QC algorithms for data	8.41%	9	10.28%	11	21.50%	23	23.36%	25	20.56%	22	15.89%	17	**107**	**3.85**
Provenance capture infor-mation for data	7.29%	7	13.54%	13	30.21%	29	20.83%	20	13.54%	13	14.58%	14	**96**	**3.64**
Reproducibility	8.74%	9	16.50%	17	29.13%	30	16.50%	17	16.50%	17	12.62%	13	**103**	**3.53**

spent finding, using, and storing data is among the more pressing concerns for users when collaborating in ESGF (see **Table 6**, this page).

Capturing a project's most valuable knowledge (asset) and distributing it effectively across the ESGF enterprise is a critical issue for the ESGF help desk, customer support, and information technology departments. **Table 7**, this page, shows the need for QC algorithms, better ingestion of and access to large quantities of data, and better integration of HPC. These features will reduce the need to move large volumes of data over the network.

The importance of data access to users is supported by findings from a 2015 survey of 270 CMIP data users, in which people rated data access and ingestion as the areas of knowledge gathering most in need of support.

Given the importance conveyed by survey responses, ESGF teams already are addressing some of these concerns, and recent conference discussions led to many viable suggestions and action items. For example, the Compute Working Team (CWT) is evaluating server-side solutions for projects that need to conduct remote analysis of large-scale data.

3. Usage Demographics (2016)

ESGF currently has almost 25,000 registered users and manages ~5 petabytes of data. In the past year alone, an estimated 600 scientific publications resulting from analysis of ESGF-delivered data have been written or are under way, and ESGF has assisted in generating more than 2,000 scientific publications during the past 5 years. ESGF's goals this year are to (1) sustain ESGF's successful existing servers; (2) address projected scientific needs for data management and analysis; (3) extend ESGF to support the major international climate assessments; (4) foster new scientific directions in the Earth system modeling community; and (5) support the dissemination of climate science data at leadership-class computing facilities around the world, including the Oak Ridge Leadership Computing Facility (OLCF), Argonne Leadership Computing Facility (ALCF), and the National Energy Research Scientific Computing Center (NERSC). To achieve these objectives, ESGF has been broadened to support multiple types of modeling and observational data, provide high-level (client-side) access and analysis services, enhance interoperability between the federation and common climate analysis tools, and enable an end-to-end simulation and analysis workflow.

Eighteen international research projects use ESGF infrastructure to globally manage and distribute their data. ESGF data scientists selected several of these research communities to help provide focus for the climate ecosystem development effort. One of the targeted communities encompasses global climate modeling groups in dozens of countries that contributed to CMIP5 and now are participating in CMIP6. CMIP researchers typically run the same prescribed set of climate change scenarios on the most powerful available supercomputers to produce datasets containing hundreds of physical variables and spanning tens, hundreds, or thousands of years. Participants in the Obs4MIPs and CORDEX projects use various techniques to simulate Earth's climate system at a higher spatial resolution over more limited areas than CMIP. The CORDEX scientists are working to forecast how Earth's climate may change regionally. To ensure data coordination, Obs4MIPs and CORDEX recently decided to adopt the same ESGF infrastructure as CMIP. More than 70 other MIPs also have adopted ESGF as their *de facto* standard data management and dissemination platform.

Although the images below represent only a fraction of the more than 40 projects as well as the ~70 MIPs supported by ESGF worldwide, they do show the more prominent sites containing most of the datasets.

3.1 Statistics Overview

ESGF supports more than 700,000 datasets from universities as well as national and international laboratories and manages over 4.6 petabytes of data distributed across more than 40 projects along with ~70 MIPs (see **Fig. 8**, this page).

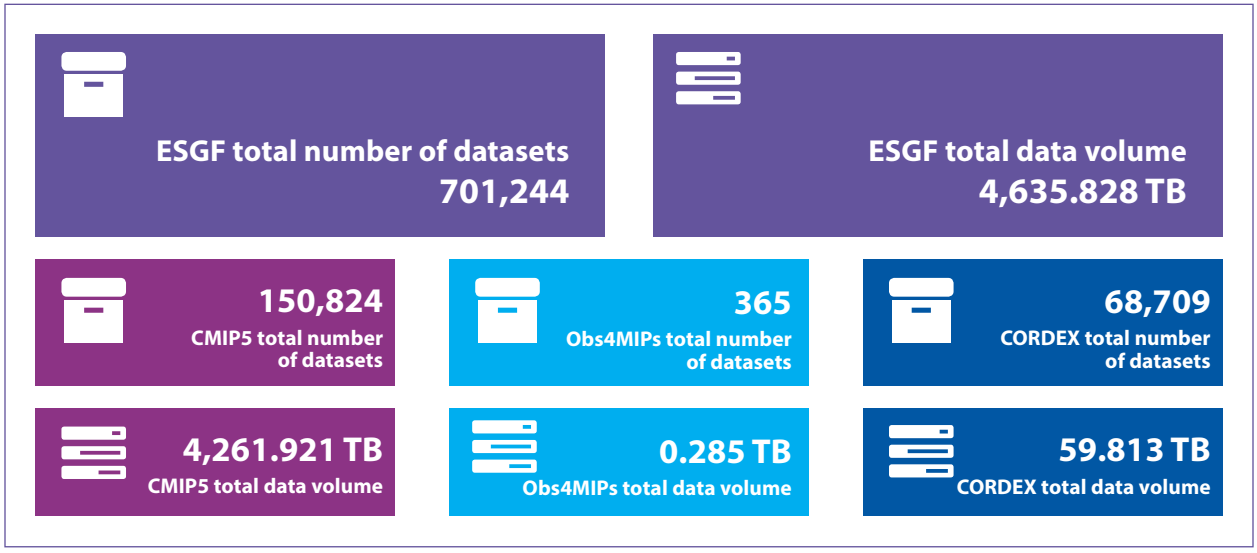


Fig. 8. ESGF data archive.

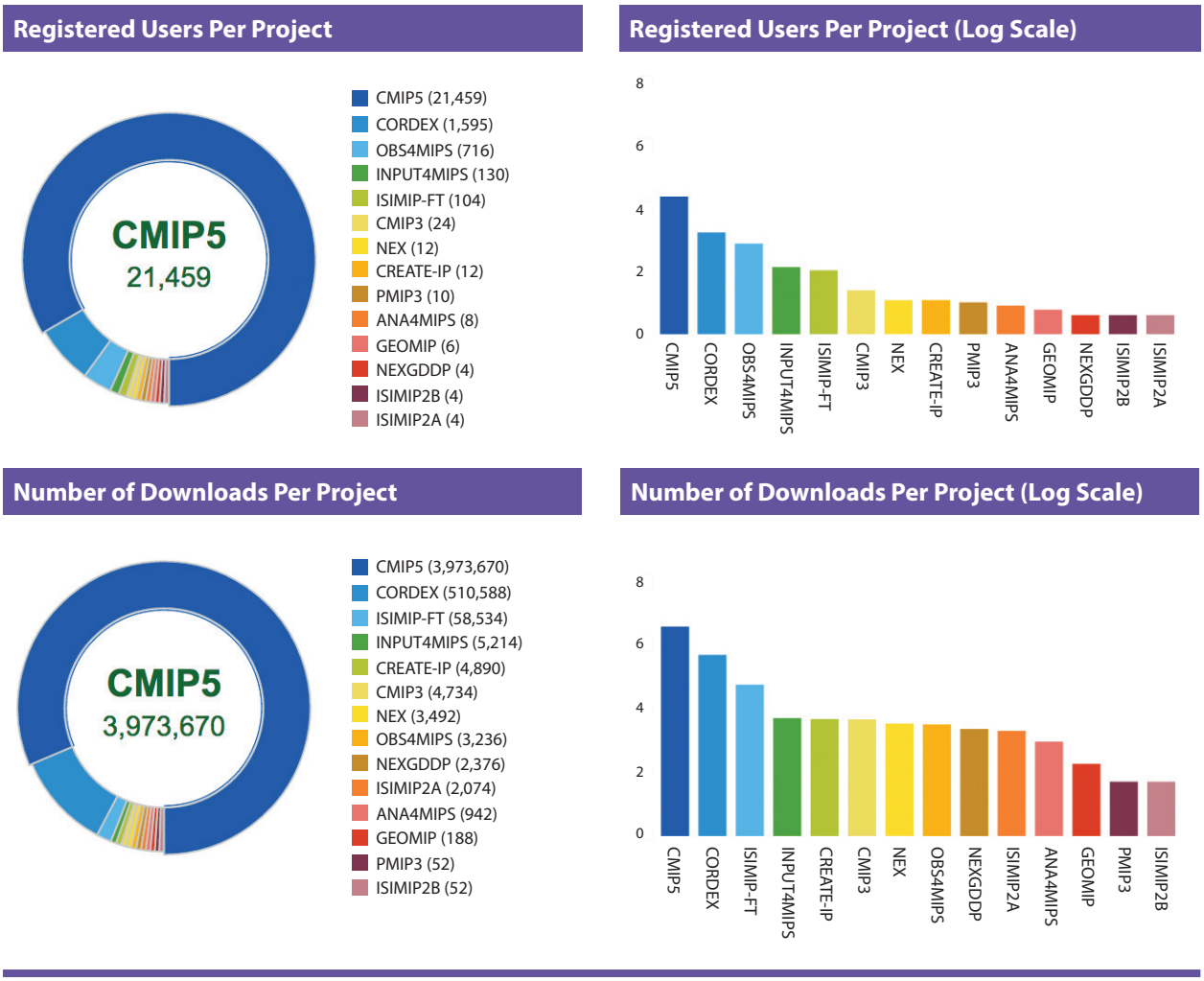


Fig. 9. Registered users and downloads per project.

3.2 Registered Users and Downloads by Project

ESGF supported more than 21,000 active CMIP5 users from universities, national and international laboratories, and private industry via millions of dataset downloads (see **Fig. 9** , this page).

3.3 Downloads by Continent

Millions of downloads come from different continents and countries (see **Fig. 10**, p. 13).

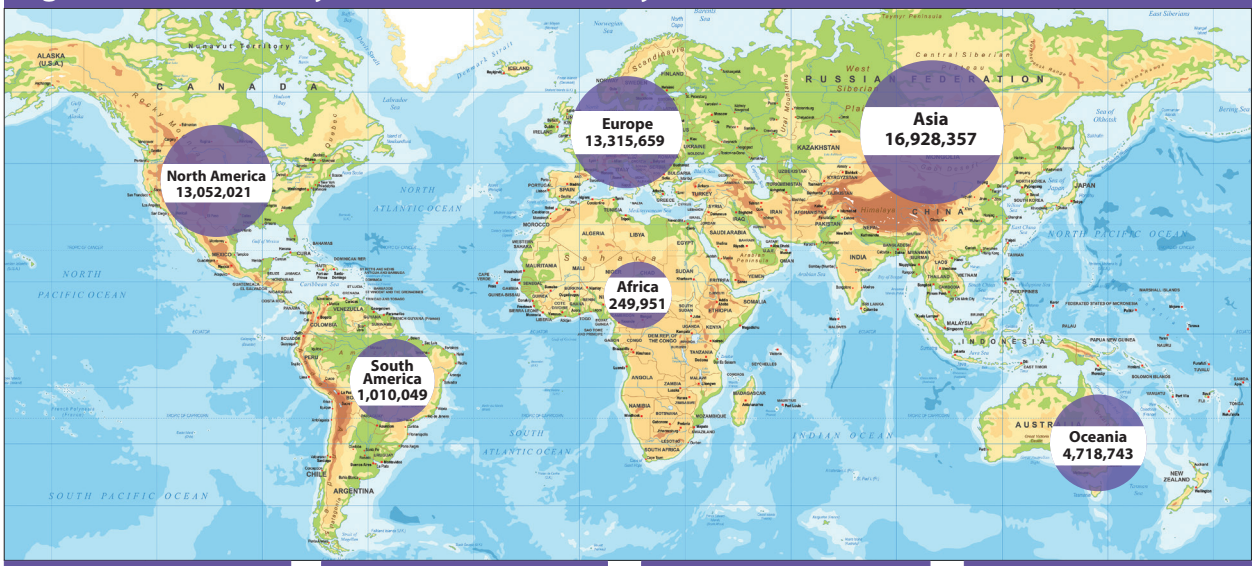
3.4 Registered Users by Continent

Figure 11, p. 14, show the geographic distribution of registered users by continents and by countries, respectively.

3.5 Registered Users by Site

Figure 12, p. 15, shows the geographic distribution of registered users by identity providers (IdPs).

Fig. 10. Downloads by continent and country



North America	13,052,021
United States	12,666,874
Canada	208,352
Trinidad and Tobago	171,291
Mexico	4,432
Jamaica	1,008
Costa Rica	42
Belize	18
Puerto Rico	2
Nicaragua	2

South America	1,010,049
Chile	785,661
Brazil	121,664
Colombia	75,554
Argentina	16,734
Peru	9,886
Ecuador	420
Bolivia	66
Suriname	38
Venezuela	22
Uruguay	4

Africa	249,951
Kenya	172,085
South Africa	59,502
Algeria	4,978
Zambia	2,974
Senegal	2,734
Niger	2,396
Egypt	1,426
Ethiopia	1,150
Ghana	752
Morocco	478
Benin	434
Cameroon	296
Tunisia	248
Nigeria	148
Botswana	130
Congo (Brazzaville)	66
Tanzania	52
Zimbabwe	36
Uganda	28
Sudan	22
Reunion	12
Mali	2
Côte d'Ivoire	2

Oceania	4,718,743
Australia	4,286,555
New Zealand	432,174
New Caledonia	14

Asia	16,928,357
China	12,148,283
Japan	2,666,194
South Korea	696,774
Iran	626,600
Thailand	276,832
India	211,548
Taiwan	118,082
Hong Kong	63,200
United Arab Emirates	29,088
Singapore	28,416
Turkey	19,262
Vietnam	8,498
Israel	7,826
Saudi Arabia	6,886
Indonesia	5,292
Pakistan	3,884
Cyprus	3,880
Malaysia	3,232
Bangladesh	2,926
Oman	504
Philippines	438
Jordan	360
Mongolia	120
Cambodia	92
Qatar	68
Kazakhstan	48
Nepal	14
Sri Lanka	6
Brunei Darussalam	2
Uzbekistan	2

Europe	13,315,659
Germany	5,981,937
Spain	3,029,726
Switzerland	1,318,830
United Kingdom	1,111,048
France	735,278
Netherlands	215,692
Norway	161,670
Italy	134,776
Sweden	112,892
Portugal	102,442
Greece	51,046
Austria	48,388
Denmark	42,584
Belgium	41,454
Finland	37,678
Bulgaria	34,102
No country	32,086
Russian Federation	27,730
Poland	19,476
Croatia	16,606
Czech Republic	14,592
Romania	13,098
Moldova	9,562
Slovenia	8,382
Serbia	4,692
Luxembourg	3,976
Hungary	1,758
Iceland	1,648
Ireland	1,532
Ukraine	686
Lithuania	230
Estonia	62

Fig. 11. Registered users by continent and country

North America	4,032
United States	3,486
Canada	434
Mexico	72
Cuba	10
Costa Rica	10
Jamaica	8
Trinidad and Tobago	4
Guatemala	2
Barbados	2
Panama	2
Nicaragua	2

South America	710
Brazil	304
Colombia	142
Chile	112
Argentina	80
Peru	40
Bolivia	12
Suriname	6
Venezuela	6
Ecuador	4
Paraguay	2
Uruguay	2

Oceania	350
Australia	308
New Zealand	38
Norfold Island	2
New Caledonia	2

Africa	470
Ethiopia	70
South Africa	66
Senegal	42
Kenya	32
Zambia	30
Egypt	26
Ghana	20
Benin	20
Nigeria	20
Cameroon	16
Burkina Faso	14
Tunisia	14
Algeria	14
Zimbabwe	14
Sudan	8
Uganda	8
Morocco	8
Tanzania	8
Madagascar	8
Botswana	6
Lesotho	4
Libya	4
Comoros	2
Côte d'Ivoire	2
Guinea	2
Mozambique	2
Togo	2
Malawi	2
Niger	2
Reunion	2
Congo (Kinshasa)	2

Asia	5,024
China	2,162
India	920
Japan	434
Iran	342
South Korea	270
Thailand	194
Taiwan	92
Indonesia	80
Pakistan	74
Malaysia	58
Israel	42
Vietnam	42
Philippines	40
Singapore	36
Turkey	34
Saudi Arabia	34
Nepal	30
Bangladesh	30
United Arab Emirates	20
Jordan	20
Mongolia	12
Kazakhstan	10
Cyprus	8
Sri Lanka	8
Iraq	6
Palestine	6
North Korea	6
Hong Kong	4
Cambodia	2
Tagikistan	2
Uzbekistan	2
Kuwait	2
Lebanon	2

Europe	3,169
Germany	1,015
France	376
Italy	242
Spain	218
Netherlands	166
Russian Federation	158
Norway	150
Sweden	128
Switzerland	94
Belgium	74
Denmark	72
United Kingdom	70
Greece	70
Portugal	62
Austria	46
Romania	42
Ireland	32
Czech Republic	30
Finland	28
Poland	20
Ukraine	16
Croatia	16
Lithuania	10
Hungary	8
Serbia	4
Belarus	4
Slovenia	4
Slovakia	2
Latvia	2
Bulgaria	2
Malta	2
Moldova	2
Luxembourg	2
Estonia	2

Fig. 12. Geographical Distribution of Registered Users by Identity Providers (IdPs)

1	pcmdi.llnl.gov	14,736
2	esgf-data.dkrz.de	2,132
3	esgf-node.jpl.nasa.gov	1,338
4	esgf-node.ipsl.upmc.fr	378
5	esgf-index1.ceda.ac.uk	198
6	esgf.nccs.nasa.gov	52
7	esgdata.gfdl.noaa.gov	32

4. Conference Findings

The annual ESGF F2F Conference offers development guidance and project prioritization to some 20 working teams within the ESGF developer community. The 2016 meeting brought together a multidisciplinary, multinational group of experts (see **Appendix D**, p. 79) from the computer science, climate science, and research communities to discuss ESGF and review a set of survey results related to ESGF usability and readiness in anticipation of the arrival of CMIP6 data (see **Chapter 2, 2016 User Survey Results**, p. 3). Representatives from each working team presented their team's achievements during the past year, prioritized development, and noted collaborations with other working teams and outside agencies. In addition to report presentations, the conference included town hall discussions to address progress, component interoperability, and roadmaps.

This year's conference highlighted the fact that ESGF's software stack and infrastructure remain incomplete and subject to constant requirements to improve and adapt to project demands, including the need to manage tens of petabytes of data for CMIP6 and related projects. By the end of the conference, many such findings were noted and relayed to the ESGF Steering Committee—the sponsors responsible for driving and funding multiple project needs (See **Chapter 1**, p. 1). Additional findings emerged from conference presentations, town hall meetings, posters, previous reports, expert testimonials, use cases, code sprints, and interoperability discussions. Emerging as a key activity, provenance is becoming more important for easily reproducing products critical for science transparency and validation. Other key findings from conference attendees are summarized below, with an emphasis on the sponsor-funded investments most likely to advance the mission and science goals of numerous community projects.

4.1 CMIP6 Data Archive

ESGF Infrastructure must not only serve a large range of users, but also house data from a variety of communities and projects.

The CMIP6 archive includes a wide range of model, observational, and reanalysis output that adheres to the CMIP data infrastructure standards and conventions adopted by the MIP community for disseminating its projects' data. There are some 70 MIP projects, including Obs4MIPs, Ana4MIPs, C4MIP, GeoMIP, OMIP, PMIP, Input4MIPs, and RFMIP.

The size of the CMIP6 archive remains unknown, with estimates spanning from 25 petabytes to 50 petabytes of uncompressed NetCDF4 data. Most of the data are scheduled for delivery to ESGF between mid-2017 and the end of 2018.

4.2 CMIP6 Tier 1 Node

Tier 1 node sites are expected to run the full suite of ESGF services for data and user management, which can be used to support their own activities and those of Tier 2 node sites. To qualify as a Tier 1 node, a system (1) must have an uptime >98% (i.e., only about 1 week of down time per year), (2) run a 10 gigabit-per-second perfSONAR host (on a physical server if at all possible), (3) deploy a rotating disk high-performance storage unit that can contain preferably 5 petabytes of CMIP6 data, (4) use Synda for data replication between Tier 1 node sites, and (5) maintain core monitoring services.

4.3 CMIP6 Tier 2 Node

Tier 2 nodes are data and modeling centers that typically have fewer physical or staff resources available for ESGF interactions but still distribute a certain (and possibly significant) amount of data to the scientific community. These sites might focus on primary data publication but do not necessarily participate in data replication.

Tier 2 node sites are encouraged to leverage some of the services supported by Tier 1 sites, such as a metadata index and IdP. These nodes focus on supporting local services for data download and possibly analysis, both of which are closely related to ESGF data projects.

4.4 Software Security

Security updates in February 2016 fixed the most recently discovered vulnerabilities. All releases go

through security scans on each component upgrade to check for vulnerabilities. When vulnerabilities are detected, immediate actions are taken to correct the problem and to inform the ESGF node administrators if code is in production.

There are security concerns that software containers, such as Docker, are opaque and not auditable and thus may require stringent vetting before being allowed in the software stack.

The software security team highlighted several security objectives:

- Ensure past security issues do not happen again.
- Perform risk assessment of code base (i.e., Solr, gridFTP, and CDAT).
- Ensure installer and software container issue is investigated closely, including security, maintenance, and installation.
- Address Solr security issues. Going forward, the team sees major security concerns in a lack of risk-based approach with future software development.
- Develop a better working relationship between the software security and installation working teams.
- Ensure all working teams collaborate with the security team during development to make sure their code releases are free from vulnerabilities.

The following security steps should be performed:

- Form a software engineering and security team that has physical, technical, and security insight into the core ESGF components.
- Create documentation for the core ESGF working team members. This will necessitate generation of a current software manifest, architecture, and operation concept of the core components.
- Freeze, if at all possible, all work other than patching on these core modules.
- Perform a risk assessment of these modules, their use, interactions, and interfaces and document the risks. The risk assessment document will provide a baseline (at least for the reviewed components) that shall, at minimum, be maintained and updated annually with component changes (to remain current with the latest threat and vulnerability information).
- Submit the risk assessment to the ESGF Executive Committee for review.
- Acquire an experienced risk executive to assist the Executive Committee in evaluating the risk assessment report, as required by the ESGF security plan.
- Allocate resources to address the identified risks (repair or mitigate) in priority order. The Executive Committee, with risk executive input, shall direct the ESGF Software Development Team to make these allocations.

4.5 Server-Side Computing

Currently, rarely used sever-side calculations are monolithic within ESGF, according to the user survey. Because the CMIP6 data volume will be too large to move, server-side computing is a necessity. To assist in remote computing, the ESGF CWT has developed modularized compute capabilities to allow use of multiple analysis engines within the ESGF framework. Town hall discussions and demonstrations showed that the following analysis tools are potential back-end analysis engines called from the CWT end-user API: Climate Data Analytics Service (CDAS), Ophidia, Power Analytics and Visualization for Climate Science (PAVICS), CAFE, WPS Birdhouse, OPeNDAP, and Ultrascale Visualization–Climate Data Analysis Tools (UV-CDAT). All analysis engines meet the suggested minimum requirements of providing data subsetting, averaging, and regridding to reduce the amount of overall data transfers. These operations are expected to account for more than 70% of all computing requests.

Allocating dedicated hardware for remote computing and managing the resource is also an issue. Therefore, the recommendation is to improve hardware resources at Tier 1 node sites for supporting server-side computing for CMIP6 users. On-demand computing at Tier 1 node sites will open up new ways for managing the complexity of the analytics workflow and will allow scientists who are used to working in isolation to be more open to broad sharing and collaboration. Without the proper resources at Tier 1 nodes, scaling remote computing will not be possible.

Remote computing also will involve resource management across the federation. With so many projects providing their own compute capabilities within the federation, ESGF must make sure only users

identified within a particular project can use designated compute resources that are paid for by that project. Also, resource management must ensure users do not monopolize compute resources and prevent others from accessing them. For this purpose, resource management will confine and set computation limits according to a machine's capacity, use, and project allocation.

4.6 Provenance Capture

ESGF F2F Conference participants identified numerous provenance preferences, from tracking data ingestion through publication and archiving to analytics and big data reduction. With these demands come structural changes in how provenance is put into practice within the ESGF infrastructure and the socialization in how it can help users answer key questions related to data generation and process. These questions include: Who generated the data, from which institution, and which analysis engine was used to reduce the data? Investigating methods for collecting provenance information (i.e., tracking which types of activities are happening when, where, and how) will add a new functionality to ESGF.

Some individual components of ESGF already have provenance integrated into their framework, but many components used by ESGF do not. For tighter integration of provenance and completeness of the overall infrastructure, the federation selected the Provenance Environment (ProvEn). By design, ProvEn is a comprehensive provenance infrastructure that fuses together external provenance standards under one common organization system and enables an application without provenance to incorporate its structure for provenance capture. This satisfies requirements to address provenance at scale and by design. ProvEn uses an open-source stack for its three major components, which also are standards-based approaches: provenance cluster, hybrid store, and API. As currently used by biologists, ProvEn and its user community would benefit ESGF, as well as add to the use of ProvEn for scientific needs. Finally, ESGF will deploy ProvEn for integration into the ESGF software stack in Docker in 2017. This fits well with ESGF's component-based architecture.

Other provenance needs deemed essential at the conference included:

- Integration and capturing of provenance for reproducibility for CMIP6.
- A clearly defined strategy for collecting provenance.
- Implementation of provenance at all sites in the federation.
- Provenance search for captured reproducibility.
- Potential application of provenance capability to climate modeling simulations on different platforms.
- Relationship of provenance to ESGF DOI and errata services.
- Capture of provenance within NetCDF4 files.

4.7 Search and Metadata

Users search metadata information found in the ESGF Apache Solr index. At present, users can search the entire federation, identify a single project or a series of projects, or use the faceted search capability to focus on desired data. These searches on relevant datasets determine temporal and spatial range before submitting the task for data retrieval. Users can check the status of each task and view or download results as NetCDF, text files, or images. ESGF's search component API allows collaborative software to search ESGF's federation for data; an access API enables retrieval. Missing from this process are the abilities to trace data provenance and provide intermediate results to the user's workspace. Also missing is an authentication mechanism needed for secure data access.

The search capability UI was found to be adequate. Mixed results from the survey revealed that 34% of users found searches to be the most difficult part of ESGF, while 27% found searches to be the most useful function. Discussed was the possibility of simplifying the search page—possibly hiding everything that is not necessary for immediate use.

Desired new metadata and search functions:

- Support updated metadata without having to republish data.
- "Retract" data if wrong, but keep the dataset-level metadata.
- Tag datasets for multiple projects.
- Enhance search features.
- Upgrade infrastructure for security.

Upgrades needed to improve metadata and search capabilities:

- Modify search infrastructure to ensure it remains viable in ESGF for years to come.
- Satisfy the Working Group on Coupled Modelling (WGCM) Infrastructure Panel (WIP) and CMIP6 requirements.
- Scale architecture to accommodate much larger volumes of data.
- Test cloud instances to support Solr cloud, enhance search performance, and scale into the future. This also will eliminate some manual tasks.

4.8 Metrics

With the advent of server-side computing to reduce the amount of data transmission, the real question emerged: How does ESGF capture the appropriate metrics that reveal the true worth of the ESGF infrastructure as well as the data and projects it supports? The ESGF Dashboard, newly unveiled at the conference, automatically captures several useful metrics by today's standards (e.g., at the project level or the entire federation).

Among these metrics are:

- Display the total number of datasets and total data volume.
- Display the number of users per project.
- Display the number of datasets downloaded.
- Display the use demographics by continent.

In addition, future metrics work would determine:

- The number of remote compute processing to reduce the data.
- The original size of the data before data reduction takes place.
- The analysis engine used and the computational algorithm(s) employed within the analysis engine.
- The number of datasets moved between federated nodes for remote computing;.
- The number of peer-reviewed journal publications associated with datasets (i.e., associated with DOIs).

More metrics may be added as projects determine which measurements are needed for their community, sponsors, and stakeholders.

4.9 Survey

Quantitative assessments can serve the ESGF enterprise by providing yearly surveys of both the user and data provider communities. These surveys provide key information about usage and problems overlooked by ESGF developers and committees. They also help sponsors and stakeholders evaluate the societal impact of dollars spent to develop and operate the ESGF infrastructure. From the most recent survey results, the user community was most concerned about the inefficient access to the petabytes of data (i.e., for downloading and uploading data), remote processing, federated storage resources, direct data delivery, errata services, and reproducibility. Reliability and resilience of resources were also concerns in addition to motivators for persistent identifier (PID) and DOI usage.

Community feedback from the survey revealed that the ESGF documentation was difficult to use and find. This seemed to be an issue confined to new users of ESGF. Finding the right tutorials also is challenging. Perhaps, providing a short introduction for ESGF beginners may address this.

4.10 Modularity

Modularity is an important software engineering principle discussed throughout the conference. It is a practical application of the principle of separation of concerns by dividing a complex system (such as ESGF) into simpler and more manageable modules. Concepts for the DOE Distributed Resources for ESGF Advanced Management (DREAM) project identified service components to modularize: publishing, search, transfer, computing, analytical, visualization, exploration, monitoring, resource management, network, workflow, and security. Many of these services discussed at the conference are in the beginning stages of modularization. Associated with each service is a Representational State Transfer (REST) API for composition that enables the inclusion of modules in other systems and helps manage the component in the larger ESGF system.

These modular components also are easily extensible; that is, for extensions such as defining properties, controls, input/output, and preprocessing features. In addition, all modular components must be simple to

install as a standalone product or within the greater ESGF software stack.

4.11 Installation

In recent years, a priority has been modifying the many methods of installing the ESGF software stack on predominately Red Hat and CentOS 6.x platforms. There are various methods for installing the ESGF software stack:

- **Manual compilation:** This is the most difficult installation method for non-power users or those unfamiliar with ESGF. It is also the most time-consuming and costly method. All dependencies must be in place along with the knowledge of when to use them.
- **Installer download to put the binary in place:** For security reasons, this is one of the least favorite installation methods. Users often do not know where packages are installed, and they may not be able to install certain packages as root.
- **Compatibility for work with the operating system (OS) package manager:** The software stack should provide Red Hat or CentOS RPMs (i.e., package managers) for as many components as possible but also allow the user to maintain control of the installation procedure without ceding control to the installer.
- **Containerized installation solutions:** This is the likely direction with software containers such as Docker, and it may be the most convenient way for users to install ESGF software and its components. The container finds the needed packages, and the package manager installs them. It also would take care of dependencies and update the package when desired. Now in its beginning stages, the Docker installation has been used for only a few ESGF software components.

In addition to moving to some sort of container solution such as Docker, the ESGF Installation Working Team is rewriting the installer from BASH to Python. This will enable better installation and the building of individual software components. It may be included in PyPI (the Python Package Index) or Conda (Python-agnostic binary package manager) repositories.

4.12 Container Software

ESGF F2F Conference attendees frequently discussed software containers for many ESGF components: in particular, Docker; to some extent, Anaconda. Through DREAM project efforts, the containers represent a lightweight runtime environment with many of the core ESGF components isolated as independent services within the overall federated system, providing an easy way to package and execute specific operations. More specifically, software containers provide the ESGF software stack with the following microservices:

- Independent and separate services.
- Ease of installation for Tier 2 nodes.
- Scalability across platforms.
- Improved fault tolerance that can be isolated.
- Independent development and deployment throughout the federation.
- Elimination of any long-term commitment to the ESGF software stack.

Under the DREAM project, the leading candidate for ESGF's components is Docker. It is an open-source project that automates the deployment and application of software independent of hardware, host OS, and programming languages. Docker has a low overhead for running individual software components independently.

4.13 Data Replication and Test Federation

Easy, automated data sharing between ESGF Tier 1 node sites for replication is a top priority for CMIP6. Needed to meet this goal are additional physical hardware platforms and networks for large data movements. With automated replica data sharing, operations will improve efficiency by keeping Tier 1 node sites in sync and enabling user access to information needed for research, reports, journal articles, and more. Primary concerns identified are the following Tier 1 node sites: ANU NCI, ENES CEDA, ENES DKRZ (German Climate Computing Centre), and DOE's Lawrence Livermore National Laboratory (LLNL). The ENES IPSL site will be heavily involved in the replication development process, but the

federation has not yet decided whether the site will participate as a Tier 1 node.

Conference attendees discussed the following issues:

- **Test federation:** The need for a test federation to verify replication between the Tier 1 sites is crucial. Currently, IPSL has a test federation along with LLNL and the Jet Propulsion Laboratory (JPL). To further the test federation, DKRZ and NCI also will create nodes for the federation's test bed. In addition to completing the test federation, attendees identified personnel or replication leads from each institution and tasked them with completing and maintaining the test federation at LLNL, JPL, IPSL, DKRZ, and NCI. Each individual is responsible for the successful execution of data replication at their respective sites.

  Identified personnel:

  - CEDA—Ruth Petrie
  - DKRZ—Heinz-Dieter Hollweg
  - IPSL—Sébastien Denvil
  - LLNL—Cameron Harr
  - NCI—Ben Evans

Test datasets: IPSL and the International Climate Network Working Group (ICNWG) have test data to move between nodes in the test federation to ensure the replication process is working smoothly. Additional CMIP5 data may be added in the future to test the replication process.

- **Synda:** Tier 1 node sites will identify a server to host and install Synda software for the automated transfer of replicated data. The Synda developers will fully document the Synda workflow replication process from download through publication. Once Synda is up and running on the test federation, they will use the replication group mailing list for coordinating replication activities and the ICNWG mailing list for network and data transfer performance work and tracking.
- **ICNWG and Replication 2017 roadmap:**
  - Deployment of a test federation.
  - Deployment of replication hosts for the test federation.
  - Deployment of test Synda hosts.
  - Publication of a replication test dataset to test federation.
  - Deployment of production Synda replication hosts.
  - Identification of initial production datasets for replication and generation of Synda configuration files.
- **Key dates:**
  - Test federation up by February 15, 2017.
  - End-to-end replication test by April 1, 2017.

4.14 Network

Network services provide ESGF with a fully meshed network topology, resulting in the highest availability of data for the federation. Network providers for the international network team include Australia's AARNet, Germany's DFN, the United States' ESnet, the United Kingdom's JanNEt, and the Netherlands' SurfNet, to name a few. Under ESGF, the ICNWG is working to set up an optimal network infrastructure for ESGF Tier 1 and Tier 2 data transfers. This group's charge is to establish network best practices to effectively transport tens of petabytes of large-scale climate data between sites and to the science community. For the four or five Tier 1 node sites, the working group is looking to achieve a consistent site-to-site data transfer rate ranging from 40 to 80 terabytes a day

Last year's accomplishments include perfSONAR deployment and data transfer node (DTN) deployment at LLNL and DKRZ. Still in progress is deployment of Globus and DTN for the Replication and Versioning Working Team's use of Synda, which are needed to meet CMIP6 performance targets. ESGF needs resource commitments (both people and hardware) to prepare for CMIP6. In particular, the focus of high-performance systems engineering and operations appears to be short on hardware and manpower. The federation must address this issue for successful large-scale dissemination of the CMIP6 petabyte archive, including the need to ensure GridFTP servers are working in production.

Currently, ESGF is not meeting its 2014 network speed goals for CMIP6—moving 1 petabyte a week between the Tier 1 node sites. The decided priority was to get the software in place before tackling the efficient movement and replication of the petabyte archive. The main reason for not achieving the transfer goals is lack of people cycles. This is an expensive process, and most node sites do not have sufficient resources to support it. To help address this issue, ESGF will streamline the process by documenting the needed components for getting the desired performance between particular sites. Afterward, it will focus on each step to make sure the process is working before moving to the next step. The consensus is that ESGF has a year to improve network efficiency before CMIP6 comes into production.

During a conference working lunch, the working groups decided to use Synda for replication, have it installed and tested at all Tier 1 node sites, identify bottlenecks, support GridFTP and Globus for data transfers, and improve network performance after the Synda replication process has been tested in production. An ICNWG and replication Gantt chart 2017 roadmap was developed at the conference to define tasks and personnel at each of the working institutions.

The tasks include:

- Deployment of a test federation.
- Deployment of replication hosts for the test federation.
- Deployment of test Synda hosts.
- Publication of a replication test dataset to test federation.
- Deployment of production Synda replication hosts.
- Identification of initial production datasets for replication and generation of Synda configuration files.

Network connection requirements:

- **Tier 1:** 10 gigabits per second.
- **Tier 2 requirements for network:** 1 to 2 gigabits per second for data provision. For CMIP5, these sites have distributed 10 times more data than they host. If we extend this to CMIP6, we have a high bandwidth requirement for Tier 2 sites.
- **Tapes, single Tier 1:** 20 plus petabytes for long-term archiving. Given the amount of data to be stored, tapes might be very useful, especially for long-term and interim needs.
- **Improved hardware and network capabilities** for future ESGF needs.

4.15 Persistent Identifiers for Data Identification

Established in 2016, the PID Services Working Team addresses fundamental PID features needed in preparation for CMIP6. This effort includes an infrastructure for performance and reliability testing; deployment testing of the PID infrastructure and workflow; and integrating PID services into the publisher, replication, and versioning tools. For the publishing process, data providers could use Climate Model Output Rewriter (CMOR) to create PIDs. This will help data providers adhere to the strict rules for generating PIDs with the federation (i.e., one PID for each dataset; any dataset published to ESGF triggers a PID creation, and any change to a dataset creates a new version of the data and a new PID). Once issued, a PID will not change. This preserves the PID records. Note, there is a distinction between data files and [atomic] datasets. That is, ESGF publication units for CMIP6 consist of all data files that belong to a time series of a single variable. Initially, PIDs assignments for each data file are through CMOR.

The working group reported no milestones missed this year; however, the deployment process and the message queue exchange have not been determined. In addition, stable operations at each site need more resources, including hardware servers. Survey results indicate a strong need within the community for reliability and resiliency of resources to support PIDs.

PIDs will be accessible from the CoG UI (by exposing them through the CoG data cart) and from NetCDF header (metadata) files. Errata services will be available via a PID landing page. Currently, the errata service already uses PIDs for data ownership and creation identification. Errata services rely on PID services for data quality, version tracking, and identification. The PID service must be in place and ready for the CMIP6 community data distribution.

4.16 Digital Object Identifiers

DOIs give users access to datasets from journal articles and conversely. ESGF will provide a data ID service that will enable users to explore published data and associated journal articles. Still under development, this forward-looking service will help make datasets citable and linkable to publications. Linking publications to supporting data clearly is an important next step for public access and reproducible research.

Each data entity suitable for journal publication will be issued a unique DataCite DOI. If data are modified in any way, a new DOI will be issued to the next version of the modified dataset. DOIs also will be used to link data quality and help keep track of versions and replications throughout the federation.

Documentation is needed to describe the DOI registration policy, service, process, and workflow. The federation will implement DOIs, along with provenance, at all sites (i.e., Tier 1 and Tier 2 node sites) where data publication occurs.

4.17 Training and Documentation

Although some ESGF components are well documented for use, others are not. As new and revised components come online, the need for more thorough documentation to assist users will bcome apparent. User support is getting better, but survey feedback suggests that documentation and its improvements are still a source of concern. Besides suggesting that the documentation is difficult to use, the survey indicates that beginners do not know where to start and finding the right tutorial can be a challenge.

In addition to documentation, the Support Working Team suggests production of 2- to 3-minute web videos for specific use cases. For developers, GitHub is still the preferred vehicle for documenting and retrieving information related to extending component development features by outside entities.

The major concern involving support is that there is little to no funding for it. As of March 2017, the lead for the ESGF support team will be unfunded at the same time CMIP6 production is ramping up.

Documentation also is needed for resource management, ESGF services and levels, metrics, download challenges, long-tail storage requests, and end-to-end overall support.

5. Scientific Challenges and Motivating Use Cases

ESGF provides science enablement services to support climate research challenges. A key support area is data management for climate models, together with management of related observations. These datasets are both global (e.g., CMIP and obs4MIPs) and regional (i.e., CORDEX). Services for interdisciplinary research areas such as climate impacts also are becoming available. The expanding diversity of climate research needs currently manifests itself in the climate research questions related to CMIP6, such as responses to forcing, systematic biases, variability, and predictability. Data challenges of CMIP6 are the enablement of multimodel analyses and the sheer volume of the resulting model output.

ESGF must face exascale data management in globally distributed data federations because of future research directions such as:

- Operational decadal predictions comparable to weather and seasonal forecasting.
- High-resolution climate downscaling.
- Exascale computing and next-generation climate missions.

ESGF must support political decision making through knowledge discovery in addition to data discovery, access, and analytics. These directions may be framed as research infrastructure challenges, which are formulated here, from general aspects to those more specific.

5.1 Open Science Cloud Challenge

Funding agencies support diverse research programs in many scientific disciplines. Climate research is just one of them, and even it encompasses a range of different research projects and data types including numerical data (climate models), satellite data (Earth observation), and observational data (monitoring networks). Managing all data types and supporting research activities across scientific disciplines requires a flexible scientific data infrastructure. In the short term, ESGF will support different research activities in a sectorial research infrastructure such as ESGF for climate and environment. In the longer term, the federation expects discipline-specific research infrastructures to form open science clouds. ESGF's data management and online analysis work will be foundational to the eventual existence of these science clouds.

5.2 Data Challenge

Scientific data are isolated, specialized, growing exponentially, and difficult to analyze due to the following factors:

- Lack of standardized data structures and formats.
- Search and discovery difficulties resulting from incomplete and inconsistent metadata.
- The requirement to move more data than necessary from the repository to the analysis platform.

Overcoming these challenges requires a common language across disciplines for enabling data management and standards to achieve benefit from synergies among the copious amounts of data that may be applicable to a particular research problem. Ultimately, the data challenge is to improve the accessibility and usefulness of high-quality research data. The near-term challenges, however, remain the organization, indexing, discovery, and delivery of large data volumes to end users via an efficient and easy-to-use infrastructure.

5.3 Data Integration Challenge

Meeting the data integration challenge requires integrating architectures for complex data-generating systems (e.g., climate models, satellites, and field observations) and high-throughput, on-demand networks. Data collection and management challenges include consistent and complete metadata and quality assessment, both of which would enable cross-disciplinary research data usage and judgment. Data discovery and access will evolve into virtual laboratories. Researchers will investigate cloud storage architectures for transparent data storage across different locations. PIDs assignments to holdings in ESGF will be game changing for the research community in both attribution and provenance of research results.

5.4 Computational Environment Challenge

Data analytics involving terabytes of data motivate integration of HPC facilities and analysis platforms close to data archive nodes. The paradigm of downloading data to an individual researcher's computer eventually will break down if data volumes continue to grow at rates that exceed the growth in network speed and bandwidth. Visualization and intercomparison tools already are in demand at major data repository sites. Next-generation data analyses may involve containerized processing agents that move across data nodes and cloud storage. Community-adapted, modern UIs enable provenance capture, workflow automation, and human-computer interaction. Support for decision control and knowledge discovery is ultimately expected.

6. Computational Environments and Data Analytics

Researchers have developed powerful software systems with flexible frameworks to support complex visualization and data analysis tasks in ESGF's computing environment. These systems can be classified broadly as turnkey applications (e.g., C4I, CAFE, CDAS, CDAT, Ophidia, and PAVICS) and data flow streaming frameworks (e.g., ViSUS). Data flow systems represent computations by directed acyclic graphs typically called pipelines or workflows. A vertex (or module) in a dataflow graph represents an atomic computational task. Connections specify the flow of data in the network—an edge between two modules indicates that the target module consumes the output of the source. Used broadly, data flow networks are the basis for many applications. The flexibility of the computational environment allows users to analyze data employing any of the turnkey applications within the federation.

The computational environment has a common end-user API that represents an important and intuitive visual programming interface. This API will help many non-programmers learn to run common and unique software data analytics. The data analytics developers adopted the API early, and thus it is a well-tested component with test suites and verification files. Most data analysis tasks require many preparation steps, and the steps may change from one data analysis framework to another. The computational environment couples the flexibility required for specifying general analysis tasks with the requirements for programming complex operations.

In the past, similar projects have claimed they would provide tools that the research community could easily adopt; however, after years of painstaking development, many of these tools have not come to fruition. To eliminate this possibility, the consortium (ESGF CWT) included a subset of the climate research community at the outset to help prioritize deployment, lend climate research credibility to the software development, and offer the research community a sense of ownership for the resulting tools. The community will measure ESGF's success by the wide adoption of these tools for MIP research needs and by the number of published papers describing new science results made possible through their use. This chapter separately describes—in alphabetical order—the technologies that will be integrated into the ESGF computational environment, explaining their strengths and weaknesses as they relate to research requirements.

6.1 Framework for Collaborative Analysis of Distributed Environmental Data (CAFE)

As the amount of environmental data expands exponentially on a global scale, researchers are challenged to increase efficiency when analyzing data maintained in multiple data centers. CAFE is a Java-based distributed data management and analysis framework that allows environmental researchers to work efficiently with distributed datasets, frees them from the need to download data from remote servers, and avoids time-consuming local data archiving and preprocessing. The design of CAFE enables it to execute analytic functions near data storage facilities. Multiple nodes can collaborate with each other to perform complex data analysis. A web-based UI allows researchers to search for data of interest, submit analytic tasks, check the status of tasks, visualize analysis results, and download those results. CAFE can deliver both ready-to-use graphs and processed data to end users. In addition, it provides an extensible solution for the deployment of new analytic functions. Compared with similar existing web-based systems, CAFE dramatically reduces the amount of data needing transmission from data centers to researchers. CAFE demonstrates great promise for enabling seamless collaboration among multiple data centers and for facilitating overall research efficiency in scientific data analysis.

6.2 Climate4Impact (C4I)

Easier access to climate data is very important for the climate change impact communities and researchers. To fulfill this objective, the Climate4Impact (C4I) web portal (climate4impact.eu) and services have been developed within the Infrastructure for the European Network for Earth System Modeling (IS-ENES) and

the Climate Information Platform for Copernicus (CLIPC) European Projects, targeting climate change impact modelers, impact and adaptation consultants, as well as other experts using climate change data. It is a platform that provides users harmonized access to climate model data through tailored services. C4I connects with ESGF using ESGF Search, OpenID, X509 PKI, OpenDAP, and THREDDS catalogs. C4I offers user interfaces and wizards for searching, visualizing, analyzing, processing, and easier downloading of ESGF datasets.

C4I exposes open standards such as Web Map Services, Web Coverage Services, and Web Processing Services. Achieved by using open-source tools such as ADAGUC and PyWPS, C4I also can use Birdhouse services. Processing services include country-based statistics and extraction by GeoJSON polygon. Open-source ICCLIM is used for climate indicator calculations and averaging. Provenance integration is executed using the web consortium W3C PROV standard for fully traceable provenance.

C4I's WPS acts as an orchestrator and performs ESGF downloads on users' behalf, extracting needed data and performing calculations on the C4I server and sending the results into the user's C4I basket for further processing, download, or visualization. Current work is enabling C4I's WPS to delegate part of the calculations to the ESGF CWT's API to maximize calculations near the data storage, minimizing data transfers, and make calculations faster. Another outcome of the F2F meeting involves plans to delegate parts of the calculations to the Ophidia platform (described later in **Section 6.5,** p. 29). Active collaborations will continue to support both ESGF CWT and ESGF Identity, Entitlement and Access (IdEA) Working Team.

6.3 Climate Data Analysis Services (CDAS)

Faced with unprecedented growth in climate data volume and demand, NASA has developed the CDAS framework, which enables scientists to execute data processing workflows, combining common analysis operations in a high-performance environment close to the provider's massive data stores. The data are available in standard [e.g., NetCDF and hierarchical data format (HDF)] formats in a Portable Operating System Interface (POSIX) file system and processed using vetted climate data analysis tools [e.g., Earth System Modeling Framework, CDAT, and NetCDF operators. A dynamic caching architecture enables interactive response times. CDAS uses Apache Spark™ for parallelization and a custom array framework for processing huge datasets within limited memory spaces.

CDAS services are accessible via the ESGF CWT API to integrate server-side analytics into the ESGF. The API can be accessed using direct web service calls, a Python script, a Unix-like shell client, or a JavaScript-based web application. Currently, new analytic operations can be developed in Python, Java, or Scala, and eventually in a wide range of programming languages (e.g., FORTRAN, C/C++, and R). Client packages in Python, Scala, or JavaScript contain everything needed to build and submit CDAS requests.

The CDAS architecture brings together the tools, data storage, and HPC required for timely analysis of large-scale datasets, where the data resides, ultimately to produce societal benefits. It currently is deployed at NASA in support of the CREATE project, which centralizes numerous global reanalysis datasets onto a single advanced data analytics platform. This service enables decision makers to investigate climate changes around the globe, inspect model trends and variability, and compare multiple reanalysis datasets.

6.4 Community Data Analysis Tools (CDAT)

Designed to integrate other tools under one application, the CDAT framework supports application and module sharing for computation, analysis, visualization, and management of large-scale distributed data. As an open-source, easy-to-use application based on Python and the Visualization Toolkit, CDAT links disparate software subsystems and packages to form an integrated environment for analysis. Its design and openness permit the shared development of climate-related software by the collaborative climate community. Other DOE projects (e.g., ACME) rely on CDAT to provide visualization and analysis for their research communities.

Specific climate packages such as genutil (developed to promote BER and MIP science requirements)

facilitate day-to-day climate analysis and diagnosis within CDAT. These tools are metadata smart; that is, they retain metadata information after some sort of data manipulation. Genutil tools include statistics, array-growing algorithms that expand a data array before comparing datasets with different numbers of dimensions (e.g., applying a two-dimensional (2D) land or sea mask to a 3D dataset), color manipulation by name, status bars, string templates, selection of noncontiguous values across a dimension, and other related functions. Cdutil is another developed package geared toward climate-specific applications such as time extraction, seasonal averaging, bounds setting, vertical interpolation, variable massager (e.g., preparing data variables such as masking and regridding for comparison), region extraction, and similar functions. Unlike many commonly used analysis tools, CDAT is equipped to process very large datasets resulting from future high-resolution climate model simulations.

6.5 Ophidia

Ophidia is a Centro Euro-Mediterraneo sui Cambiamenti Climatici [Euro-Mediterranean Center on Climate Change (CMCC)] Foundation research effort aimed at providing a big data analytics framework for eScience. Ophidia supports declarative, server-side, and parallel data analysis, jointly with an internal storage model able to deal efficiently with multidimensional data and a hierarchical data organization to manage large data volumes, or datacubes. The project relies on a strong background of high-performance database management and Online Analytical Processing (OLAP) systems to manage large scientific datasets. It also provides a native workflow management support to define processing chains and workflows with tens to hundreds of data analytics operators to build real scientific use cases. The software stack includes an internal workflow management system, which coordinates, orchestrates, and optimizes the execution of multiple scientific data analytics and visualization tasks. A graphical UI also supports execution of real-time workflow monitoring. To address the challenges of the use cases, the implemented data analytics workflows include parallel data analysis, metadata management, virtual file system tasks, map generation, rolling of datasets, and import or export of datasets in NetCDF format. Implementation of a native input/output server provides in-memory analytics.

The Ophidia server exposes a WPS-compliant interface able to accept and manage WPS requests related to all three WPS methods. The WPS implementation of Ophidia relies on the Python Web Processing Service (PyWPS) module, an Apache-embedded Python module enabling a WPS interface. Specifically, Ophidia is an implementation of the server-side Open Geospatial Consortium (OGC) WPS standard that allows users to easily activate a WPS interface on top of a set of processes and define their features and behavior (compliant with the WPS specification), exploiting the Python language. An implementation of the ESGF CWT specification is also ongoing.

Ophidia is being exploited and extended in the context of several European projects. In particular, two Horizon 2020 cloud-centric projects, INDIGO-DataCloud and EuBra-BIGSEA, aim at extending Ophidia to enable, respectively, (1) multimodel analytics experiments in ESGF and (2) QoS-based elastic and dynamic analytics scenarios in cloud environments.

6.6 Power Analytics and Visualization for Climate Services (PAVICS)

Increasingly, climate observations and simulations are informing investment, management, and conservation decisions. However, the raw data is not what reaches decision makers; instead, tailor-made climate products aggregate and synthesize terabytes of data for specific applications. Coined "climate services," this processing of raw data into usable, actionable information is undergoing a surge of interest worldwide. Climate service providers are under pressure to serve more users while ensuring the information they provide is high quality, well documented, and reproducible.

The PAVICS project is an effort begun in 2016 to build a computational platform for facilitating the work of climate service providers and other scientists analyzing climate data. Through an OGC WPS interface, PAVICS exposes a number of data processing operations that can be assembled into workflows. These operations include dataset selection and retrieval, spatial and temporal subsetting, bias-correction and

simple statistical downscaling, ensemble averaging and regridding, climate indices computation, and visualization. By abstracting the complexity of climate analyses and its underlying computing architecture, the objective is to make climate analyses more accessible to a wide range of users, as well as to facilitate the work of climate scientists. PAVICS is similar in scope and intent to the European Climate4Impact portal.

In addition to these services, PAVICS offers a web front end to access, monitor, and archive processes. This front end uses the web framework, React, and leverages the mapping capability of Geoserver, which also is used to store and serve the polygons used for spatial subsetting. For the back end, PAVICS relies heavily on Birdhouse components. A demonstration server will be available in 2017, and the software will enter production in 2018.

6.7 Visualization Streams for Ultimate Scalability (ViSUS)

Modern climate datasets are growing massive due to the increased resolution and temporal granularity needed to more accurately simulate natural phenomena. These new datasets are virtually unattainable to all but the most well provisioned users. Furthermore, the amount of data becomes a burden even for an organization, and transferring and processing these data will be increasingly more cumbersome, hindering scientific investigation. To empower exploration and quick comparison of these huge datasets, scientists can use an alternate data format called IDX to facilitate the streaming of multiresolution data processing. The ViSUS software system implements a streaming dataflow based on the multiresolution IDX format, so scientists need not compromise the fidelity of their simulations to perform rapid analysis and visualization. While providing the means for full-resolution computations is still important, the ViSUS streaming framework and lossless IDX data format provide a rich method for performing rapid analyses for asking preliminary questions and enabling cursory exploration of otherwise intractable data. Scientists can use the results of these initial *ad hoc* investigations to determine the best parameters for full-scale processing operations and to share early results with colleagues in the field.

Used for interactive comparison of multiple remote climate simulations, the ViSUS framework performs ensemble analyses such as computing the variance or correlation between the results of different climate models. The embedded scripting system enables flexible, interruptible analyses by implementing a generic data processing type without explicit specification of location or resolution of the underlying data, as well as by showing incremental results for ongoing computations. By abstracting resolution and loop order, and using the multiresolution IDX format, the system can select the best settings to maintain interactivity while producing incremental results that rapidly converge toward the final solution.

A demonstration server installed at LLNL enables on-demand access to the existing hosted NetCDF climate datasets in the multiresolution IDX format. The CMIP5 class of datasets can be converted to the IDX format in near real time. A multiresolution cache of the high-resolution, 7 kilometer GEOS-5 Nature Run simulation from NASA is also available. It includes all full-resolution, 2D aerosol and climatology fields and a selection of 3D fields, while enabling on-demand streaming access to all other fields using OPeNDAP. In addition, instrumentation of the CDMS library used by many other tools will enable native multiresolution access to IDX datasets.

6.8 Growth Areas

The federation must address numerous requirements of climate projects to fulfill the growing visualization and analysis needs of relevant research communities. General solutions to these requirements involve one or more of the above-mentioned emerging technologies that have provided individual solutions to climate data and analysis needs, while other technologies have met the needs of communities outside climate. Combining these technologies could usher in a new era within the climate research community. The consortium's software (under ESGF CWT's guidance), woven together under a common architecture, will address the multifaceted requirements of general research. These requirements include streaming and parallel visualization and analysis (exploiting parallel input/output); distance visualization and data access; comparative visualization; statistical analyses; robust tools for regridding, reprojection, and aggregation;

support for unstructured grids and nongridded observational data, including geospatial data common for many observational datasets; and workflow analysis and provenance.

Other ESGF CWT efforts, some of which may rely on other ESGF working teams' development, include tasks such as security, resource management, provenance, publication, and infrastructure testing. Other issues include end users' concerns with documentation, ease of use, and independent or separate installation.

Short-term goals that emerged from the F2F conference include immediate efforts such as the following: (1) providing a base server using ZeroMQ (a high-performance asynchronous messaging library) as a communication tool between the server and implementation kernels; (2) hardening the caching capabilities; (3) using THREDDS v4.6 (eventually, THREDDS v5) to serve results; (4) ensuring that a common end-user API can call federated ESGF CWT servers; (5) implementing an automated testing infrastructure; and (6) determining and implementing a set of additional core services for specific climate research projects.

The following compute capabilities will form the initial set of core services in time for the release of CMIP6 data: subsetting, aggregation, averaging, and regridding.

7. Technology Developments

The 2016 F2F conference was, as usual, an opportunity for all ESGF working groups to highlight their work over the past year and report on progress and future plans. Overall, four broad goals guided ESGF work and development in 2016:

- Bring the federation back to operational status from the 6-month shutdown after the security incident that occurred in 2015.
- Prepare to serve the upcoming CMIP6 data collection, together with other high-profile collections such as CMIP5, CORDEX, and Obs4MIPs, with particular emphasis on optimizing the end user's experience.
- Improve and modernize the ESGF infrastructure to guarantee its longevity and collaborative nature as the federation moves further into the era of big data.
- Expand the array of available services operating on ESGF data; for example, improve data transfer performance and enable server-side computation.

This chapter summarizes the major technical developments that took place in several functional areas, often spanning more than one working group. For a more in-depth description of each task, please see the working group reports posted online at esgf.llnl.gov/media/pdf/2015-ESGF-Progress-Report.pdf and esgf.llnl.gov/media/pdf/2016-ESGF-Progress-Report.pdf.

7.1 Installation

Despite great effort and progress, installing and maintaining an ESGF node remain challenging even for expert system administrators. Because of this, two separate efforts began in 2016, both aimed at improving the installation process and the maintainability of its software.

- **Conversion to Python.** A new task undertaken by LLNL aimed to replace the current system of bash scripts with a more maintainable set of Python modules. The Python-based installation software would be more modular, easier to upgrade, and allow separate installation of each ESGF component, as opposed to running the whole bash script each time.
- **Conversion to Docker.** As a separate effort, JPL began to investigate creation of a new installation paradigm, whereby ESGF creates components as Docker images that run as interacting containers on Docker-enabled hosts. In addition to having the same advantages of the Python installer (i.e., modularity, readability, and maintainability), the Docker installer would enable easy deployment of multiple system architectures (e.g., a full ESGF node, an index node, and a data node) simply by using a different configuration file and would allow scaling applications with multiple containers and into multiple hosts. By using Docker Swarm, ESGF services could take further advantage of built-in features such as load balancing, fault tolerance, and rolling updates.

7.2 Publishing Services

The Publication Working Team worked throughout 2016 to improve the reliability of the publishing software, expand its functionality, and tighten the accuracy of the resulting metadata.

On the client side, the team undertook a major effort to enforce conformance of the generated metadata to a set of controlled vocabularies (CVs) that a group of experts maintains. This will guarantee a more homogenous metadata archive and, ultimately, more accuracy for data search operations. Additionally, the new features of the enhanced client software include esgprep, which includes functionality for map-file generation, vocabulary check, and retrieval of initialization file.

On the server side, the publishing services were augmented with two major features:

- **Atomic metadata updates:** Adds, updates, or deletes any metadata field of an already published dataset, without having to republish it from scratch.
- **Data retraction:** Retracts datasets, resulting in the data being unavailable for download, but retaining some metadata (at the dataset level) in the archives for long-term reference.

Security-enabled REST APIs support both of the above operations.

7.3 Search Services

Besides the server-side publishing functionality already described above, the Metadata and Search Working Team has focused on implementing a few additional requirements as presented by the WIP in preparation for CMIP6. These include tagging datasets for multiple projects, searching datasets for older versions (for citation purposes), and enabling geospatial searches.

Additionally, the team has researched and prototyped the use of a new Solr Cloud-based search architecture on and has benchmarked its performance up to 1,000 times the size of the current metadata archives. Solr Cloud is the enterprise version of the Solr search engine, which ESGF already has adopted, and would come with additional benefits such as automatic distributed indexing and searching, fault tolerance, load balancing, and horizontal scaling. Using simulated data, the search team demonstrated that Solr Cloud would allow the ESGF search infrastructure to scale from the currently supported datasets of about 100 K to those of 100 M. The roadblock to adoption lies in the fact that the Solr Cloud architecture comprises a cluster of internal hosts, managed by a single organization, and is not designed for a distributed environment such as ESGF. A possible solution already in progress is to use the current distributed Solr servers as "repeaters" that are harvested into a high-performance search cluster based on Solr Cloud.

7.4 User Interface

As part of the ESGF 2.0 "reboot" that occurred after the 2015 security incident, ESGF switched the UI to CoG, a Django-based web application that was first developed under separate funding from the U.S. National Oceanic and Atmospheric Administration (NOAA) to support governance, interaction, and documentation of scientific projects. Throughout the first part of 2015, the ESGF team worked at installing, configuring, and operating this new system of CoG web portals as a front end to the distributed array of ESGF data services. Now operated by Tier 1 institutions, several interconnected CoG portals allow users access to the ESGF data and metadata archives.

In addition, the major CoG development efforts that took place this past year resulted in seven major CoG releases (from 3.2.0 in January 2016 to 3.8.0 in November 2016) and eight minor releases. The main focus has been to keep improving the CoG UI to the ESGF search services—supporting retracted datasets, new search options, import and export of search configuration, multiple selection options for each facet, and generation of massive Wget scripts via POST requests. Also, the federation has extended integration with the Globus platform to support downloading of restricted datasets in addition to the public datasets enabled last year. Finally, ESGF addressed several security issues in response to a code analysis performed by NOAA NCDC, including constantly evolving the CoG software to base it on the latest (and most secure) version of Django.

7.5 Security

In ESGF, user authentication and authorization, while necessary for several reasons such as licensing and metrics, remain hindrances for many users and constant sources of user questions and complaints. Additionally, the ESGF security infrastructure is based on protocols and standards (such as the authentication protocol, OpenID 2.0, and the Security Assertion Markup Language, SAML) that are either being discontinued or are falling out of fashion. Consequently, the security team has spent considerable time drafting an evolutionary path that will enable ESGF to upgrade its core infrastructure while avoiding disruption of the current operational system. **Chapter 8**, p. 37, describes this roadmap in detail. In terms of development, two major milestones already have occurred:

- CEDA has implemented an OAuth2-based IdP, which is a Django-based application destined to replace, when possible, the current Java-based OpenID 2.0 IdP.
- Argonne National Laboratory has prototyped the upgrade of the CoG web interface to use OAuth2 client libraries to authenticate users with the CEDA OAuth2 IdP.

7.6 Data Transfer, Network, and Replication

During 2016, several groups worked together to improve ESGF's ability to efficiently transfer data between data centers and to end users.

The ICNWG has demonstrated how the use of the GridFTP protocol, combined with the setup of proper science DMZs and network tuning, can achieve data transfer speeds of 10 to 15 gigabits per second between continental DOE data centers. Now the team is ready to work with the individual ESGF data centers to improve data transfer performance to levels between 500 megabits per second and 4 gigabits per second, speeds which are necessary to execute efficient replication across CMIP6 Tier 1 sites. The current limitation is in obtaining a time commitment from the data centers to prioritize this task.

At the same time, the Replication and Versioning Working Team (RVWT) has worked to establish the infrastructure for replicating CMIP6 major data collections when they become available in 2017. The team has installed and prototyped Synda servers at the core sites, including configuration with Globus, replication policies, and optimization for large file transfers. Also executed were several replication tests between sites.

The Data Transfer Working Team (DTWT) has focused on enabling faster data downloads to the end user. The team has developed procedures and documentation for adding Globus end points to all published data. This addition has been implemented at several sites such as the Program for Climate Model Diagnosis and Intercomparison (PCMDI), NASA JPL, DKRZ, and NCI. DTWT also worked on further integration of Globus with the CoG UI, which now allows Globus downloads for both restricted and unrestricted datasets. The team currently is working on updating the CoG interface to use OAuth2 for user authentication. This program will enable users to skip retyping their credentials when requesting a data transfer to Globus. Finally, the team also developed a monitoring infrastructure for HTTP, GridFTP, and Globus file transfer endpoints for compiling a daily report on the status of the federation.

7.7 Computing Services

Enabling efficient computing services to operate on the distributed ESGF data archive is key to the program's successful handling of CMIP6 (and other) data collections. To this goal, the CWT has worked to enable deployment of production-level services by the second half of 2017, when a significant portion of CMIP6 data are expected to become available. In 2016, the team finalized a general API for server-side operations, based on WPS with specific CWT extensions for domain and variable. The API implementation is purposely left open ended so that different groups may implement it in different languages. In fact, several implementations already are in the works: Scala based by NASA GSFC, Birdhouse based by Ouranos , and Ophidia based by CMCC. LLNL PCMDI also has deployed a test server. Next, the team plans to work on several other tasks including user authentication and resource management, naming conventions, and cross-validation of results across different API implementations.

7.8 Metadata Services

DKRZ and IPSL are developing several metadata services to support publishing and documentation operations for CMIP6 data management.

- The **PID service** is intended to assign PIDs to several objects (e.g., collections, datasets, and files) during publication without impeding publication in any way. This service later will allow easy data search and identification of constructs by PID and will become part of data citations. During 2016, a full prototype of the PID service environment (composed of a RabbitMQ server, PID client, and web front end) was developed and installed at DKRZ.
- The **Early Citation Service** is intended to allow citation of datasets used in scientific publications before any DOIs are assigned to the data (assignments can take up to a few years). During 2016, the team formulated the relevant use cases involving a data creator, data user or article writer, and data reviewer or article reader, and it worked on prototype integration with the CoG UI. The team is ready to start working on end-to-end implementation and full integration into the CoG master branch.
- The **Errata Service**, under development at IPSL as part of the ES-DOC project, aims to provide a central repository for reporting and accessing documentation related to problems with the data. The full system comprises an issue client, a web service, and a front end, plus a GitHub OAuth client and a handle service. In 2016, the team finalized an end-to-end workflow that involves all of the above

components. The team now is preparing to bring the entire infrastructure into production to meet the CMIP6 timeline.

7.9 Provenance Capture, Integration, and Usability

ESGF uses an open-source scientific provenance management system developed at DOE's Pacific Northwest National Laboratory (PNNL) to support data exploration and visualization. Provenance systems traditionally are used to capture and automate ordered repetitive tasks disclosed, for example, from desktop applications to support reproducibility. However, in applications that are exploratory in nature—such as parallel or distributed simulations, streaming data analysis, and interactive visualization—change is the norm. In these situations, provenance typically is disclosed asynchronously, in parallel from multiple software or human agents. As an engineer or scientist generates and evaluates hypotheses about data under study, they use real-time provenance feedback (e.g., performance optimization, quality controls (QCs), and anomaly detection or trending) for recommendations to steer the interactive process. In effect, this process creates a series of different, albeit related, workflows. The ProvEn used for ESGF was designed for ease of use and to manage and integrate reproducible events contained within ESGF.

A distinguishing feature of ProvEn is its comprehensive provenance infrastructure, which fuses a high availability in-memory data grid, graph database, and time series database for maintaining detailed history information about the steps followed and data derived during an exploratory task. ProvEn maintains the provenance of data products (data flows), the workflow history (process flows) that derives these products, and their executions. This information, which persists as a federated partitioned graph and time-series database, allows users to navigate workflow versions in an intuitive way. Because provenance disclosure uniquely identifies individual messages, errors can be removed without unintentionally altering other results. Users also can visually examine and compare different provenance workflows, their results, and the actions that led to a specific result.

Although not currently used in the climate community, ProvEn naturally supports complementary functionality to many of the ESGF components. In particular, it provides seamless workflow integration, a REST API, use of the W3C PROV ontology, and comparative visualization functionality. The client API is written in Java, harvesters have been developed, and the REST API offers a variety of solutions for applications to disclose provenance. ProvEnalso is capable of incorporating or citing provenance managed by the ESGF community. For instance, the PID service will be relied upon to identify ESGF data products in the provenance. The system, written in Java, relies on open-source technology and with its modular design is easy to integrate with ESGF.

7.10 Services

The Dashboard and Stats Working Team has been developing a complete rewrite of the ESGF metrics services, supporting both coarse-grained statistics for older logging data (before the 2015 security incident), finer-grained statistics for newer logging data, and federated statistic views that span all nodes in the federation. The architecture consists of a dashboard back end to be installed at each data node, a REST API, and a dashboard front end to be installed at the index nodes, all of which will present graphical information constructed by using the REST API to query the back ends. The team has completed a beta version of all software components and is in the process of executing local and federation-wide tests to validate all queries and the information presented before integrating the software with the ESGF installation process.

8. Roadmap

During the F2F conference, the ESGF working teams listened to and shared feedback and requirements from the user community, sponsors, some major scientific projects, and other teams. By the end of the meeting, the working teams in collaboration with the ESGF Executive Committee had established a new roadmap. This roadmap is divided into two time horizons: CMIP6 preparedness (short-term ESGF longevity, 0 to 2 years) and longer-term longevity (2 to 5 years).

8.1 Short-Term Plans (0 to 2 Years)

To successfully handle and distribute CMIP6 data, the ESGF developer community and Executive Committee have identified the following short-term gaps and priorities (in approximate order of critical need).

8.1.1 Replication

Based on a collective assessment, many accomplishments are needed to bring the ESGF replication infrastructure to an operational status in time for CMIP6. The goal is to have the four major Tier 1 CMIP6 centers (i.e., LLNL, DKRZ, CEDA, and NCI—possibly IPSL) ready to replicate data across continental and intercontinental boundaries within 6 months after the 2016 F2F meeting (i.e., by June 1, 2017). Toward this purpose, the ESGF Executive Committee, in collaboration with the major players, has developed the following plan:

- January 2017. The major Tier 1 CMIP6 centers will begin to devote all necessary resources for setting up a test federation of nodes that will be fairly stable and be used to demonstrate replication. This federation needs to be ready by mid-February 2017.
- January 2017, or immediately thereafter. The centers begin to install and test the Synda software, first using HTTP for back-end transfer protocols and then switching to Globus and GridFTP. This task is set for completion by the end of February 2017.
- After February 2017. When data start moving between sites, center representatives will work with the ICNWG and system administrators at their sites to tune network and settings to maximize data transfer performance.

The team decided to schedule the first replication tests for March 1, 2017, with subsequent tests in the April to May time frame, with the goal of being fully operational by June 1, 2017.

8.1.2 Documentation and Training for Data Publishers

Another gap identified during the F2F meeting was the lack of a systematic approach in identifying the federation's state of readiness along with its software tools and the procedures of the several groups and data centers that will be publishing CMIP6 data. Toward this goal, the Coupled Model Intercomparison Project Data Node Operations Team will assume responsibility for the following tasks:

- Compile a list of software tools that need to be installed at each publishing center (e.g., CMOR, ESGF PrePARE, QC checkers, and ESGF Publisher).
- Verify that each group has acquired the necessary credentials and privileges to publish to its corresponding index node.
- Create a set of instructions and tutorials describing the publishing operations.

8.1.3 Software and Operations Security

The conference noted that ESGF still has margin for improvement in the way it handles software security at each site and across the federation. A need raised was for creation of a best practices document that will instruct site administrators how best to configure and monitor the ESGF software stack, for example, network configuration, port restrictions, and SELinux installations. Additionally, ESGF needs to develop a model for periodic risk assessments of both software code and operations and to establish a remediation procedure for fixing any vulnerabilities in a timely manner.

8.1.4 PID Service

Although strictly not required to execute publishing operations, the new PID service is considered an

essential part of the CMIP6 publishing process because it will enable a higher level of identification and tracking services than was available for CMIP5. Therefore, establishing an operational PID service from the beginning of the CMIP6 data flow will be important. This involves bringing the PID software to full release status, deploying PID services at the three intended sites (i.e., IPSL, DKRZ, and LLNL), and executing publication tests that include PID assignments.

8.1.5 Basic Data Reduction and Analysis Operations

As data volumes increase at an ever-faster pace, ESGF needs to enable users to download only portions of data in which they are interested and, possibly, to execute some preliminary data processing on the server side. Unfortunately, although many ESGF groups are making great progress toward enabling a wide array of server-side computations, there is doubt that a fully fledged and validated software stack will be ready for installation within the next 6 months. Therefore, the ESGF collaboration has decided to follow a conservative timeline that includes:

- Deployment of a TDS OPeNDAP server and publication of OPeNDAP endpoints at all data nodes. The fact that CMIP6 data will not require the downloading of any user authentication will guarantee that standard OPeNDAP clients can request basic subsetting operations. Also needed are basic documentation and tutorials.
- Deployment, if possible, of a first WPS-compliant server at each data node, supporting basic operations such as averaging, regridding, and zonal means. A mature implementation such as UV-CDAT or the Birdhouse suite should provide the back-end libraries. Deployment of the server as a Docker image would facilitate installation.

8.1.6 User Authentication and Authorization

The Software Security Working Team (SSWT) is planning a gradual, major upgrade of the ESGF infrastructure, which will take place over the next few years. The ultimate goal is to use OpenID-Connect as the authentication protocol and OAuth2 for all authorization purposes such as delegation. Specific milestones in this roadmap include the following objectives (in chronological order of implementation):

- Replace the OpenID 2.0 IdP with a new OAuth2 IdP, and later with a new OpenID Connect IdP.
- Enable CoG to authenticate users instead of OAuth2 IdP, and later use the OpenID Connect IdP.
- Similarly, evolve the current OpenID Relying Party (ORP) to authenticate users versus the OAuth2 IdP, and later move to the OpenID Connect IdP.
- Replace the current MyProxy server with CEDA's new REST-based server for proxy certificates.
- Upgrade the Wget web downloader scripts to contain an embedded certificate that does not require user credentials.

When completed, the security upgrade will enable a superior user experience while providing the opportunity to authorize server-side computations through user-delegated credentials.

8.1.7 Other Short-Term Priorities

Other, less critical priorities identified for CMIP6 preparedness include bringing the errata services to operational status, capturing provenance through the data lifecycle, and possibly supporting older but easier ways for users to access data, such as via FTP or rsync.

8.2 Longer-Term Plans (2 to 5 Years)

Users of ESGF have directed long-term services, giving community projects the flexibility they want while reducing unmet needs for deliverables and community-based services and supports. The federation is making efforts to expand such services and components under ESGF, including those supported by the federation and its Executive Committee and through the evaluation of proposals and grants. Challenges related to costs, staffing and organizational issues, new infrastructure requirements, and resistance from stakeholders often hinder these efforts. The ESGF community has developed a number of successful strategies for overcoming these challenges, even in financially trying times, as part of a longer-term vision. These experiences offer valuable insights, guidance, and encouragement to the community for completing the longer-term, user-directed service expansions described below.

8.2.1 Server-Side Computation

One of the most critical requirements for ESGF in the near future is to progressively move the computation to the data. That is, the federation should accomplish a technological—and social—paradigm shift, whereby scientists do not download massive amounts of data to their desktop but, rather, take advantage of proven and reliable computational facilities offered on servers with direct access to the data storage. Within ESGF, the distributed nature of the collective archive compounds the problem. For instance, scientists might orchestrate computations across different sites, possibly involving data movement, partial results, more data movement, and finally computation.

During this F2F conference, several groups reported great progress in developing computation suites that are deployed and run on the servers; UV-CDAT, Birdhouse, and Ophidia are a few prominent examples. The Compute Working Team has provided a unification layer by defining a server-side API (based on WPS, with custom extensions) that different groups are implementing. This API will allow interoperability between the different packages. Additionally, ESGF is developing a client-side API to facilitate the task of formulating correct processing requests (which can have quite a complex syntax) and sending them to the server.

Meeting participants formulated two action items with regard to server-side computation:

- Provide cross-validation across the different WPS implementations—that is, guarantee that the same operation ("average," for example) will yield the same result on a given dataset independent of the site at which it is performed.
- Develop "orchestrators" that are able to coordinate server-side operations at multiple distributed data centers, within the same analytical workflow.

Development and improvement of server-side computational engines will be strong driving forces of ESGF development for the near future.

8.2.2 Installation

The long-term longevity of ESGF is not guaranteeable without drastic changes to the installation process. The ESGF software stack needs easier installation and maintenance to enable a wider range of configuration architectures as well as scaling on multiple hosts and commodity clusters.

As mentioned before, during the F2F meeting, there were presentations of two separate efforts that aim to address these problems: the conversion of the installer to Python and the adoption of Docker as enabling containerization framework. These two approaches are compatible; for instance, ESGF could run modular Python scripts as part of the Docker recipes for creating images.

Meeting participants raised a few security concerns about the Docker model—namely, how to guarantee that Docker images are promptly updated across the federation in response to newly discovered vulnerabilities, either in the server kernel or in the upper software layers. By the end of the meeting, the collaboration decided to further evaluate the Docker approach while analyzing the security concerns, and it proposed a "threat response" model that would be acceptable to the SSWT.

8.2.3 Cloud Computing

A general trend in science and business information technology is an increased reliance on commercial cloud services offered by technology giants such as Amazon, Google, Microsoft, and others. During this meeting, several groups and sponsors presented their long-term data strategies, which invariably involve some form of cloud computing. For example, NASA GSFC is planning to move more and more of its data assets to the cloud, starting with observational data, then model output, for general access by users. The European community has funded the Open Science Cloud project, which aims to deploy a full-fledged IT infrastructure for scientific discovery and analysis on the cloud.

To stay relevant and competitive, ESGF also must be able to make use of cloud services either for permanent deployment or, perhaps, for enabling short periods of intensive "data bursting." Converting the ESGF software stack to the Docker platform is a move in this direction, because Docker images could be easily instantiated on the cloud and Docker itself offers several tools for automatic cloud deployment, monitoring, and scaling (e.g., Docker Swarm and Docker

Data Center). Independently, the ESGF Executive Committee has formulated cloud deployment as one of the critical requirements for the long-term evolution of ESGF.

8.2.4 Programmatic Access to Data

More and more, experts and power users are moving beyond using a simple web portal to search for data and create scripts for massive data download; instead they are requesting that ESGF enable APIs that would allow them to access data directly from their client and applications.

ESGF already offers a very stable search API, which has been used by several client applications and has been a cornerstone of ESGF success for several years. In the future, the ESGF collaboration might consider pairing the more climate specific ESGF API with the Solr API. Enabling direct access to the Solr servers (in read-only mode) would enable clients to execute a wider range of queries and take advantage of advanced Solr features such as statistics and other packages.

Even more importantly, users need access to the data objects beyond the simple files. They need the ability to subset, aggregate, decimate, and operate on data. ESGF recognizes the need to enable such direct APIs, which might be coalesced into the one WPS API formulated by the Compute Working Team or expressed as a portfolio of APIs from which to choose.

9. Community Developments and Integration

Partnerships and the intent to collaborate reflect close relationships to a wide variety of data, science, and technology efforts that position ESGF to make a major impact on the progress of science in several areas: (1) CMIP6 and Intergovernmental Panel on Climate Change's (IPCC) Sixth Assessment Report (AR6), (2) ACME, (3) CORDEX, (4) MIPs, and (5) many more inclusive funded research projects such as DOE's DREAM and Europe's Copernicus Climate Change Service. In most cases, these projects would be developing tools and technologies for the ESGF software stack; in some cases, there is an interest in generalizing and enhancing ESGF-developed technologies and disseminating them to a larger audience. For example, the federation will rely on the DREAM project to enhance, support, and develop well-defined APIs for component layers of core ESGF services (e.g., publishing, search, transfer, computation, resource, and exploration) and for the construction of scalable science services (e.g., service framework, service provisioning, and service monitoring). In addition, visualization and analytics capabilities will rely on CDAT to integrate, package, and deliver visualization and analysis products to the climate community.

For many on the collaboration list, this project will be vital for their national and international programs and projects to reach their goals. For this reason, ESGF's well-positioned team members will overlap in services and institutions and represent the project as liaisons among many of the disparate organizations.

Table 8, this page, provides an impressive list of potential collaborators and other community development and integration opportunities in alphabetical order. Each community development (i.e., collaboration) must have an agreed-on schedule of technology deliverables that meets the federation's requirements for success.

Table 8. Integrated Community Collaborations

Title (Type)	Collaborator Type	Lead Institution and PI	Areas of Collaboration
Accelerated Climate Modeling for Energy (ACME)	Relies on ESGF to reach their goals	DOE LLNL, Dave Bader	Computational resources
BASEJumper	Relies on ESGF to deliver their products to climate science community	DOE LLNL, Sam Fries	Publication and access of HPSS datasets
Birdhouse	Relies on ESGF to deliver their products to climate science community	ENES DKRZ, Stephan Kindermann	Provisioning web processing services
Community Data Analysis Tools (CDAT)	ESGF relies on collaborator for developing tools and technology	DOE LLNL, Charles Doutriaux	Analysis and visualization toolkit
Community Diagnostics Package (CDP)	ESGF relies on collaborator for developing tools and technology	DOE LLNL, Zeshawn Shaheen	Framework for creating new climate diagnostics
Climate Information Platform for Copernicus (CLIPC)	Relies on ESGF to reach their goals	ENES CEDA, Martin Juckes	Climate data store for reanalysis, satellite data (ESA CCI) seasonal forecasts, and projections (global/ or regional)
Climate Model Output Rewriter (CMOR)	Relies on ESGF to reach their goals	DOE LLNL, Denis Nadeau	Generates CF-compliant metadata conventions and incorporates a controlled vocabulary API

Table continued next page

Table 8. Integrated Community Collaborations			
Data Reference Syntax	ESGF relies on collaborator for developing tools and technology	DOE LLNL PCMDI, Karl Taylor	Structured set of conventions to facilitate the naming of data entities within the data archive and of files delivered to users
DREAM: Distributed Resources for ESGF Advanced Management	ESGF relies on collaborator for developing tools and technology	DOE LLNL, Dean N. Williams	Providing a host of underlying services that can be adopted in part or as a whole by other science domains
Earth System Documentation (ES-DOC)	ESGF relies on collaborator for developing tools and technology	ENES IPSL, Mark A. Greenslade	Supports Earth system documentation, creation, analysis, and dissemination
Globus	ESGF relies on collaborator for developing tools and technology	DOE ANL, Ian Foster	Large-scale data movement and delivery (i.e., transfer)
Program for Climate Model Diagnosis and Intercomparison (PCMDI)	Relies on ESGF to reach their goals	DOE LLNL PCMDI, Karl Taylor	IPCC AR6, CMIP, MIPs, PCMDI Metrics Package (PMP)
Provenance Environment (ProvEn)	ESGF relies on collaborator for developing tools and technology	DOE PNNL, Eric Stephan	Provenance capture, integration, and connection
Synda	ESGF relies on collaborator for developing tools and technology	ENES IPSL, Sébastien Denvil	Bulk data movement and publication of replica data
THREDDS Data Server: OPeNDAP	ESGF relies on collaborator for developing tools and technology	Unidata, Sean Arms	Browseable metadata cataloging confirmation to THREDDS

10. Report Summary and Development for 2017

With the data from CMIP6 and many related MIP projects now being served or close to being served, ESGF is working to improve various aspects of the system by adding new capabilities that should better meet the needs of its users. However, how the community's needs will be met is unclear should any of the contributing funding agencies end or have their funding levels reduced—a concern mentioned repeatedly at the conference.

Should financial support continue for the near future, the following improvements are planned for 2017:

- **Fast Analytics:** Will enable users to connect, manipulate, and visualize data in minutes, with 10 to 100 times faster analytics and more capabilities than the LAS solution. With online visualization streaming and other analysis components, these newly introduced capabilities will enable users to make quick inspection and comparison of datasets from multiple locations. More importantly, an enhanced capability to perform server-side data reduction and calculations will reduce the volume of data transferred to users via the Internet.
- **Ease of Use:** Will allow users to more easily search, access, and analyze data with ESGF's intuitive UI. This includes a simpler scripting method for downloading files (using Globus) and notification services for advising users when errors are found in datasets. Users will not need to program; however, an API will be available for external application developers who wish to connect to ESGF internal and external services.
- **Big Data:** Will enable ESGF users to more easily explore any data, from managed software containers (e.g., Docker) to databases to Hadoop to cloud services. Enhancements will include straightforward methods to report errors discovered in the data and to provide feedback to the modeling groups about their simulations (e.g., ES-DOC). General system enhancements related to scaling to hundreds of millions of datasets and tens of petabytes of data volume—in accordance with the CMIP6 projected archive—also are under consideration.
- **Smart Dashboards:** Will allow users to combine multiple views of data to gain richer insights into the data (i.e., a dashboard capability showing the worldwide system metrics of users, downloads, and published data). Folded into the UI will be best practices for visual displays.
- **Automatic Update:** Will enable users to obtain the most up-to-date data with Synda and Globus connections, which replicate data among Tier 1 node sites at specified scheduling times. This enhancement will include the ability to move large amounts of data at high speeds among sites critical to CMIP6 goals. The critical Tier 1 sites are LLNL PCMDI (U.S), BADC (U.K.), DKRZ (Germany), and ANU NCI (Australia). An extended list may include IPSL (France).
- **Immediate Sharing (in seconds):** Will allow users, with a few clicks, to securely publish long-tail individual data to the federated archive and to share the information with designated colleagues on the web or via ESGF's external API services.
- **Accumulated Data History:** Will provide users with provenance information that is necessary for reproducible results in a shared environment. Distinguishing features of the provenance infrastructure within ESGF will include maintenance of detailed history information about steps and procedures occurring during data publishing, processing, and movement. Plugins will provide support for transparently gathering provenance information derived through the scientific discovery process.
- **Error Reporting:** Will make accessible to all researchers a centralized repository in which to record errors found in data files. There also should be a mechanism to report errata or concerns to the modeling centers and data publishers by the end-user community. To this end, ESGF will make errata tools available through ES-DOC and the ESGF PID service, which enable viewing of file histories and linkage to errata reports based on file identifier.

ESGF is by no means a finished product because climate modeling is an ever-evolving activity of increasing complexity and sophistication—and

endless data. Ensuring the optimal design of ESGF's data infrastructure for developing new, validated, and verified capabilities with proven technology is just as important as the climate models that produce the data. The federated data and computing infrastructure must undergo constant, rapid development with assessment of new scientific modules to provide a testing-to-production environment for simulation and evaluation (i.e., metrics, diagnosis, and intercomparison) of observational and reanalysis data. Scientific challenges and requirements, along with a diverse set of climate use cases, drive the development and use of the overall enterprise and individual components as stand-alone systems with specific APIs. Though some tools are specific to a particular project, wherever possible the development teams identify common methods and similar APIs and establish tools that satisfy the requirements of many projects.

To achieve individual project and community goals, the ESGF team will continue to further develop and enforce standards and promote the sharing of resources, such as NetCDF, CF metadata conventions, OPeNDAP, UV-CDAT, ES-DOC, Data Reference Syntax, 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 to base the data infrastructure. ESGF continues to build 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.

The climate modeling and observational communities of tomorrow will have a tremendous need for ESGF (or its descendants). For the reasons listed above, ESGF must be fully functional, lightweight, fast, flexible, and accurate enough to meet the demands of virtually any big data project. ESGF exists to ensure that this will be the case.

Appendices

Appendix A. Conference Agenda 47
Appendix B. Presentation, Demonstration, and Poster Abstracts 55
Appendix C. ESGF's Current Data Holdings 77
Appendix D. Conference Participants and Report Contributors 79
Appendix E. Awards 83
Appendix F. Acknowledgments 85
Appendix G. Acronyms 87

Appendix A. Conference Agenda

2016 Earth System Grid Federation Face-to-Face Conference (Washington, D.C.)

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

Time	Topic
Monday, December 5, 2016	
2:00 p.m. – 4:00 p.m.	Pre-conference registration: Jr. Ballroom Salons 1 and 2
5:00 p.m. – 6:00 p.m.	Social Activity: Meet and Greet (no host) at Cuba Libre, 801 9th St., NW A, Washington, D.C.
Tuesday, December 6, 2016	
7:30 a.m.– 8:30 a.m.	Registration: Jr. Ballroom/Salons 1 and 2
8:00 a.m. – 8:30 a.m.	Meet and greet
8:30 a.m. – 8:35 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 2016 ESGF F2F 6th Annual Conference
8:35 a.m. – 8:45 a.m.	DOE opening comments — Gary Geernaert, director of the Climate and Environmental Sciences Division of DOE's Office of Biological and Environmental Research (BER)
8:45 a.m. – 9:00 a.m.	State of the Earth System Grid Federation — Dean N. Williams
ESGF Steering Committee (A note from our sponsors)	
9:00 a.m. – 10:45 a.m.	**ESGF Steering Committee** *Session Discussion Lead — Dean N. Williams* **9:00 a.m. – 9:20 a.m.** Justin Hnilo — DOE Office of Biological and Environmental Research (BER) Data Management **9:25 a.m. – 9:45 a.m.** Sylvie Joussaume — Infrastructure for the European Network of Earth System Modelling (IS-ENES2) Coordinator **9:50 a.m. – 10:10 a.m.** Tsengdar Lee — National Aeronautics and Space Administration (NASA) Headquarters High-End Computing Program **10:15 a.m. – 10:35 a.m.** Ben Evans — National Computational Infrastructure (NCI) **Questions** • What infrastructure strategies should be established to accelerate progress in Earth system modeling/ observation and understanding? • What are the key things that are difficult to do today and are impeding scientific progress or productivity and the sharing of data? • What is your timeline for data production and distribution from climate model and observations, high-performance computer, network, and storage facilities needs and investments? • What is the estimated size of your distributed archive? • What are your common developments, sharing of expertise, and accelerated developments? • What are the administrative/sponsor requirements that arise from each project (basically, metrics collection and reporting)? • What are your expected strategic roadmaps and ESGF funding levels for the short term (1 to 3 years), mid term (3 to 5 years), and long term (5 to 10 years)? • What is the political landscape to be made aware of? **Homework assignment** • The homework assignment before the conference is to convert all known science drivers to use cases.
10:45 a.m. – 11:00 a.m.	**Break**

Table continued next page

Time	Topic
11:00 a.m. – 11:30 a.m.	**Steering Committee Town Hall Discussion** *Session Discussion Lead — Dean N. Williams* Town Hall Panel — Justin Hnilo, Sylvie Joussaume, Tsengdar Lee, and Ben Evans • What is working, and what is not? • What are the key challenges to your programs? • What data services would address the identified challenges? What exists already today? What do we still need? What are the key characteristics that these services need to have to be successful (i.e. integrated, easy to customize, etc.)? • What are the key impediments (on the data provider/service provider side) in delivering these services? • Which services should be developed with the highest priority, and what would be their measurable impact on science/programs?
11:30 a.m. – 5:30 p.m.	**ESGF Progress and Interoperability** *Session Discussion Lead — Dean N. Williams* ESGF working teams quickly report out on meeting 2016 projects requirements (work achieved over the past year, prioritized development, collaborations with other agencies, etc.) **11:30 a.m. – 11:40 a.m.** CoG User Interface Working Team — Luca Cinquini, NASA/JPL **11:45 a.m. – 11:55 a.m.** Metadata and Search Working Team — Luca Cinquini, NASA/JPL **12:00 noon – 1:30 p.m.** Lunch **1:30 p.m. – 1:40 p.m.** Publication Working Team — Sasha Ames, DOE/LLNL **1:45 p.m. – 1:55 p.m.** Node Manager and Tracking/Feedback Working Team — Sasha Ames, DOE/LLNL **2:00 p.m. – 2:10 p.m.** Stats and Dashboard Working Team — Alessandra Nuzzo, ENES/CMCC **2:15 p.m. – 2:25 p.m.** Identity Entitlement Access Management Working Team — Phil Kershaw, ENES/CEDA **2:30 p.m. – 2:40 p.m.** Compute Working Team — Charles Doutriaux, DOE/LLNL **2:45 p.m. – 2:55 p.m.** Errata Service — LEVAVASSEUR Guillaume, ENES/IPSL **3:00 p.m. – 3:10 p.m.** Quality Control Working Team: Data Citation Service for CMIP6 — Status and Timeline — Martina Stockhause, ENES/DKRZ **3:15 p.m. – 3:25 p.m.** Installation Working Team — Prashanth Dwarakanath, ENES/Liu **3:30 p.m. – 3:45 p.m.** Break **3:45 p.m. – 3:55 p.m.** Docker for ESGF — Luca Cinquini, NASA/JPL **4:00 p.m. – 4:10 p.m.** International Climate Network Working Group — Eli Dart, DOE/ESnet **4:15 p.m. – 4:25 p.m.** Data Transfer Working Team — Lukasz Lacinski, DOE/ANL **4:30 p.m. – 4:40 p.m.** Security Working Team — George Rumney, NASA/GSFC **4:45 p.m. – 4:55 p.m.** Replication and Versioning Working Team — Stephan Kindermann, ENES/DKRZ **5:00 p.m. – 5:10 p.m.** Persistent Identifier Services — Tobias Weigel, ENES/DKRZ **5:15 p.m. – 5:25 p.m.** User Working Team — Torsten Rathmann, ENES/DKRZ
5:30 p.m.	**Adjourn Day 1**
6:00 p.m. – 7:00 p.m.	Awards ceremony and live entertainment — Jr. Ballroom Salons 1 and 2

Table continued next page

Time	Topic
Wednesday, December 7, 2016	
8:00 a.m. – 8:30 a.m.	Meet and greet
8:30 a.m. – 9:30 a.m.	**ESGF Progress and Interoperability Town Hall Discussion** *Session Discussion Lead — Dean N. Williams* **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 ESGF integrated 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 that are needed for the working teams in respect to the many projects? • What level of service, monitoring, maintenance, and metrics is needed for each of the working team data services and tools? • What do working teams want to see from others? • What do the scientists want to have access to with regard to the working teams? • What standards and services that needs to be adopted within the compute environment that will allow projects to participate in multi-agency data initiatives discussed on the first day? • What is needed for data sharing across the multi-international agencies?
9:30 a.m. – 11:30 a.m.	**Advanced Computational Environments and Data Analytics** *Session Discussion Lead — Robert Ferraro* **9:30 a.m. – 9:40 a.m.** Overview of the Compute Working Team and Target Milestones — Daniel Duffy, NASA/GSFC; Charles Doutriaux, DOE/LLNL **9:45 a.m. – 9:55 a.m.** Compute Working Team (CWT) End-User Application Programmer's Interface (API) — Jason Boutte, DOE/LLNL; Charles Doutriaux, DOE/LLNL **10:00 a.m. – 10:10 a.m.** The Climate Data Analytic Services (CDAS) Framework — Thomas Maxwell; Dan Duffy, NASA/GSFC **10:15 a.m. – 10:25 a.m.** Ophedia big data analytics framework — Sandro Fiore, ENES/CMCC **10:30 a.m. – 10:40 a.m.** PAVICS: A Platform to Streamline the Delivery of Climate Services — David Huard; Tom Landry; Blaise Gauvin-St-Denis; David Byrns, CRCM **10:45 a.m. – 11:00 a.m.** Break **11:00 a.m. – 11:10 a.m.** Server-side Computing Services provided by IS-ENES through the climate4impact Platform — Christian Page; Wim Som De Cerff; Maarten Plieger; Manuel Vega; Antonia S. Cofino; Lars Barring; Fokke De Jong; Ronald Hutjes; Sandro Fiore, ENES/Copernicus **11:15 a.m. – 11:25 a.m.** CAFE: A framework for collaborative analysis of distributed environmental data — Hao Xu, China/Tsinghua University **11:30 a.m. – 11:40 a.m.** Embedded Domain-Specific Language and Runtime System for Progressive Spatiotemporal Data Analysis and Visualization — Cameron Christensen; Shusen Liu; Giorgio Scorzelli; Ji-Woo Lee; Peer-Timo Bremer; Valerio Pascucci, University of Utah

Table continued next page

Time	Topic
9:30 a.m. – 11:30 a.m.	**Questions** • What are the key challenges that scientists encounter? • What capabilities would address the identified challenges? What exists already today? What do we still need? • What are the impediments for resource providers and software developers to provide these missing capabilities? • Which requirements need to be addressed with the highest priority and what would be their measurable impact on science? • What is the overall integration plan? • What are the key things that are difficult to do today and are impeding scientific progress or productivity? **Homework assignment** • The homework assignment before the conference is to convert all known data center drivers to use cases.
11:45 a.m. – 12:10 p.m.	**Computational Environments and Data Analytics Town Hall Discussion** *Session Discussion Lead — Robert Ferraro* Town Hall Panel — Charles Doutriaux, Daniel Duffy, Jason Boutte, Thomas Maxwell, Sandro Fiore, Maarten Plieger, David Huard, Christian Page, and Cameron Christensen • Define 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 • Identify key gaps, identify benefitting communities, and prioritize
12:10 p.m. – 1:30 p.m.	**Lunch**
1:30 p.m. – 5:45 p.m.	**Coordinated Efforts with Community Software Projects** *Session Discussion Lead — Sébastien Denvil* **1:30 p.m. – 1:40 p.m.** CMIP6 Standards Enabling Management, Search and Interpretation of Model Output — Karl Taylor, DOE/LLNL **1:45 p.m. – 1:55 p.m.** CMIP6 ESGF Tier 1 and Tier 2 Nodes — Sébastien Denvil, ENES/IPSL; Michael Lautenschlager, ENES/DKRZ **2:00 p.m. – 2:10 p.m.** CMIP6 "Impact" on Scientific Community — Sergey Nikonov; V. Balaji; Aparna Radhakrishnan; Daniele Schneider; Hans Vahlenkamp, NOAA/GFDL **2:15 p.m. – 2:25 p.m.** Control Vocabulary Software Designed for CMIP6 — Denis Nadeau; Karl Taylor; Sasha Ames, DOE/LLNL **2:30 p.m. – 2:40 p.m.** Developing a Vocabulary Management System for Data Reference Syntax using Linked Data Technologies in the Climate Information Platform for Copernicus (CLIPC) Project — Ruth Petrie; Phil Kershaw; Ag Stephens; Antony Wilson, ENES/CEDA **2:45 p.m. – 2:55 p.m.** DKRZ ESGF Related Infrastructure and CMIP6 Services — Stephan Kindermann; Michael Lautenschlage; Stephanie Legutke; Katharina Berger; Martina Stockhause, ENES/DKRZ **3:00 p.m. – 3:10 p.m.** The IPCC DDC in the context of CMIP6 — Martina Stockhause; Michael Lautenschlager; Stephan Kindermann, ENES/DKRZ **3:15 p.m. – 3:25 p.m.** Persistent Identifiers in CMIP6 — Merret Buurman; Tobias Weigel; Stephan Kindermann; Katharina Berger; Michael Lautenschlager, ENES/DKRZ **3:30 p.m. – 3:45 p.m.** Break

Table continued next page

Time	Topic	
1:30 p.m. – 5:45 p.m.	**3:45 p.m. – 3:55 p.m.**	ES-DOC and ES-DOC Services (Atef Ben Nasser, Mark Greenslade, ENES/IPSL
	4:00 p.m. – 4:10 p.m.	National Computational Infrastructure's Research Data Services: Providing High-Quality Data to Enable Climate and Weather Science — Claire Trenham; Kelsey Druken; Adam Steer; Jon Smillie; Jingbo Wang; Ben Evans, NCI/ANU
	4:15 p.m. – 4:25 p.m.	Automating Data Synchronization, Checking, Ingestion, and Publication for CMIP6 — Ag Stephens and Alan Iwi, ENES/CEDA
	4:30 p.m. – 4:40 p.m.	Input4MIPs: Boundary Condition and Forcing Datasets for CMIP6 — Paul J. Durack, DOE/LLNL; Karl Taylor, DOE/LLNL; Sasha Ames, DOE/LLNL; Anthony Hoang, DOE/LLNL
	4:45 p.m. – 4:55 p.m.	Update on the ESGF Needs for Obs4MIPs — Peter Gleckler, DOE/LLNL
	5:00 p.m. – 5:10 p.m.	Recent Climate4impact Developments: Provenance in Processing and Connection to the CLIPC Portal — Maarten Plieger; Wim Som de Cerff; Andrej Mihajlovski; Ernst de Vreede; Alessandro Spinuso; Christian Page; Ronald Hutjes; Fokke de Jong; Lars Barring; Antonio Cofino; Manuel Vega; Sandro Fiore; Alessandro D'Anca, ENES/KNMI
	5:15 p.m. – 5:25 p.m.	Federated Data Usage Statistics in the Earth System Grid Federation — Alessandra Nuzzo, Maria Mirto, Paola Nassisi, Katharina Berger, Torsten Rathmann, Luca Cinquini, Sébastien Denvil; Sandro Fiore; Dean N. Williams; Giovanni Aloisio, ENES/CMCC
	5:30 p.m. – 5:40 p.m.	Large-Scale Data Analytics Workflow Support for Climate Change Experiments — Sandro Fiore; Charles Doutriaux; D. Palazzo; Alessandro D'Anca; Zeshawn Shaeen; Donatello Elia; Jason Boutte; Valentine Anantharaj; Dean N. Williams; Giovanni Aloisio, ENES/CMCC
	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 are you funded for longevity?	
5:45 p.m.	**Adjourn Day 2**	

Table continued next page

Time	Topic
Thursday, December 8, 2016	
8:00 a.m. – 8:30 a.m.	Meet and greet
8:30 a.m. – 10:15 a.m.	**Coordinated Efforts with Community Software Projects** *Session Discussion Lead — Sébastien Denvil* **8:30 a.m. – 8:40 a.m** THREDDS Data Server: OPeNDAP and Other Tales from the Server-Side — Sean Arms, Unidata **8:45 a.m. – 8:55 a.m.** A Hybrid Provenance Capture Approach to Scientific Workflow Reproducibility and Performance Optimization — Todd Elsethagen; Eric Stephan; and Bibi Raju, DOE/PNNL **9:00 a.m. – 9:10 a.m.** QA/QC at the DKRZ — Heinz-Dieter Hollweg, ENES/DKRZ **9:15 a.m. – 9:25 a.m.** Web Processing Services and ESGF: the Birdhouse System — Stephan Kindermann; Carsten Ehbrecht; Nils Hempelmann, ENES/KNMI **9:30 a.m. – 9:40 a.m.** Synda (synchro-data) — Sébastien Denvil, ENES/IPSL **9:45 a.m. – 9:55 a.m.** Globus Update — Rick Wagner, University of Chicago and DOE/ANL **10:00 a.m. – 10:10 a.m.** BASEJumper: Publishing HPSS datasets via ESGF — Sam Fries; Sasha Ames; and Alex Sim, DOE/LLNL
	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 are you funded for longevity?
10:15 a.m. – 10:45 a.m.	**Community Software Projects Town Hall Discussion** *Session Discussion Lead — Sébastien Denvil* Town Hall Panel — John Caron, Todd Elsethagen, Maarten Pileger, Ag Stephens, Denis Nadeau, Sam Fries, A. Nuzzo, Cameron Christensen, Sandro Fiore, and Denis Nadeau **Questions** • What standards and services need to be adopted within the environment that will allow projects to participate in multi-agency data initiatives? • How should these tools and services be made available in ESGF's future in an integrated way?
10:45 a.m. – 11:00 a.m.	**Break**
11:00 a.m. – 12:00 noon	**Live Demonstration Session** *Session Discussion Lead — Dean N. Williams*
12.00 noon – 1:30 p.m.	**Lunch**

Table continued next page

Time	Topic
2:30 p.m. – 3:00 p.m.	**Poster Session** *Session Discussion Lead — Luca Cinquini* **Posters** **1.** ADAGUC open source visualization in climate4impact using OGC standards — Maarten Plieger and Ernst de Vreede, ENES **2.** Community Data Management System (CDMS) — Denis Nadeau; Charles Doutriaux; and Dean N. Williams, DOE/LLN **3.** Community Diagnostics Package — Zeshawn Shaheen; Charles Doutriaux; Samuel Fries, DOE/LLNL **4.** ESGF Compute Working Team End-User Application Programmer's Interface — Jason Jerome Boutte and Charles Doutriaux, DOE/LLNL **5.** Earth System Model Development and Analysis using FRE-Curator and Live Access Servers: On-demand analysis of climate model output with data provenance — Aparna Radhakrishnan; V.Balaji; Roland Schweitzer; Serguei Nikonov; Kevin O'Brien; Hans Vahlenkamp; and Eugene Francis Burger, NOAA/GFDL **6.** Toward a high-performance data analysis platform for impact analysis — Wim Som de Cerff; Sandro Fiore; Maarten Plieger; Alessandro D'Anca; Giovanni Aloisio; KNMI; CMCC Foundation, ENES/CMCC **7.** Web Processing Services and ESGF: the birdhouse system — Stephan Kindermann; Carsten Ehbrecht; and Nils Hempelmann, ENES/CEDA **8.** Climate4Impact Portal — Maarten Plieger, KNMI **9.** ACME Workflow — Sterling Baldwin, DOE/LLNL **10.** HPSS connections to ESGF — Sam Fries, DOE/LLNL **11.** Distributed Resource for the ESGF Advanced Management (DREAM) — Dean N. Williams and Luca Cinquini, DOE/LLNL **12.** Community Data Analysis Tools (CDAT) — Charles Doutriaux; Sam Fries; Aashish Chaudhary; Dean N. Williams, DOE/LLNL **13.** Visual Community Data Analysis Tools (VCDAT) — Matthew Harris and Sam Fries, DOE/LLNL **14.** Climate Forecast (CF) Convention — Karl Taylor, DOE/LLNL **15.** ES-DOC — Mark Greenslade, ENES/IPSL **16.** Agreement on Data Management and Publication Workflow — Sasha Ames, DOE/LLNL **17.** Data Citation Service — Martina Stockhause, ENES/DKRZ **18.** PCMDI's Metrics Package — Paul Durack, DOE/LLNL **19.** DOE UVCMetrics — Jim McEnerney and Jeff Painter, DOE/LLNL **20.** ESMValTool — Stephan Kindermann, ENES/DKRZ **21.** CMIP6 Errata as a New ESGF Service — Guillaume Levavasseur, ENES/IPSL **22.** A NASA Climate Model Data Services (CDS) End-to-End System to Support Reanalysis Intercomparison — Jerry Potter, NASA/GSFC **23.** CAFE: A framework for collaborative analysis of distributed environmental data — Eric Xu, China/Tsinghua University **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 Discussion and Cross-Team Discussions** • Poster session feedback • Open discussion
5:00 p.m.	**Adjourn Day 3**

Table continued next page

Time	Topic
Friday, December 9, 2016	
8:00 a.m. – 8:30 a.m.	Meet and greet
8:30 a.m. - 10:00 a.m.	**ESGF XC and WIP Breakout Meeting** • Discuss of the construction of the annual report • Meeting location and time of the next ESGF F2F meeting **Working Teams Meeting** • All working teams discuss conference findings for their area for the annual report
10:00 a.m. – 10:15 a.m.	**Break**
10:15 a.m. - 12:00 noon	**ESGF Development Teams Report Back on Conference Findings** *Session Discussion Lead — Dean N. Williams* • ESGF Team Leads findings on conference feedback • Open discussion
12:00 noon	**Adjourn Day 4**
Conclusion of the 6th Annual ESGF F2F Conference	
1:30 p.m. – 5:00 p.m.	**General Code Sprint (optional)** • Working Teams and Leads

Appendix B. Presentation, Demonstration, and Poster Abstracts

Day 1: Tuesday, December 6, 2016

ESGF Steering Committee

The State of the Earth System Grid Federation

Dean N. Williams (DOE/LLNL), Williams13@llnl.gov

The Earth System Grid Federation (ESGF) has assisted the scientific research community and their projects for over a decade with the dissemination and management of climate simulation and observational data products. Our federated "data cloud" infrastructure houses millions of files and annually transfers many petabytes of data to the community for large-scale knowledge discovery. The large-scale use of the infrastructure has enabled us to amass a great deal of intelligence about the state of our software stack, housed data, and needed capabilities for customer satisfaction. This information enables us to effectively organize, plan, and prioritize the next steps in ESGF software development. To help in our planning phase, we have conducted a survey study around user practices and ESGF node performance and capabilities. From the survey, we also ascertain the extent to which a large sample of projects, regardless of the national or international funding agency, use different types of data quality control, gathering, managing, or sharing methods.

The state of ESGF will also feature federated data usage statistics generated by the ESGF dashboard and desktop and highlighted ESGF Executive Committee documents, such as the ESGF Policies and Guidelines, ESGF Strategic Roadmap, ESGF Software Security Plan, ESGF Implementation Plan, and ESGF Root Certificate Authorization Policy and Certificate Practices Statement.

Department of Energy Office of Biological and Environmental Research Data Management

Justin Hnilo (DOE/BER), Justin.Hnilo@science.doe.gov

The Climate and Environmental Sciences Division (CESD) within DOE's Office of Biological and Environmental Research (BER) focuses on advancing a robust, predictive understanding of Earth's climate and environmental systems by exploiting unique modeling, observational, data, and infrastructure assets, which BER develops and manages. CESD's strategic plan includes five goals, each of which contains a modeling, observational, and data management component. Within this plan, there is a special emphasis on leading the nation in developing highly efficient modeling architectures, testbeds, data analytics, and analysis tools to support the broad climate science community within the context of DOE's mission. CESD's Data Management activity represents a highly coordinated set of data-oriented research activities, with a goal to provide the CESD scientific community with easy and efficient access to all necessary databases to study increasingly complex scientific challenges. Research in support of this activity involves metadata compatibility from disparate research projects; fusion of data derived from laboratory studies, field observatories, and model-generated output; server-side analysis; and development of multimedia analytical tools, including multidimensional visualization and efficient storage. Current and future investments will be highlighted.

Infrastructure for the European Network for Earth System Modelling

Sylvie Joussaume (ENES/CNRS-IPSL), sylvie.joussaume@lsce.ipsl.fr

The European Network for Earth System Modelling (ENES) integrates the European community working on climate modeling. Its infrastructure project, IS-ENES, supports the European contribution to ESGF and ES-DOC for WCRP-coordinated experiments for global and regional models, CMIP and CORDEX. The ENES data infrastructure contributes to the development of the ESGF software stack, data quality control, data identification and data citation, data replication and cache maintenance, and dashboard, as well as development of the metadata tools. It provides support to both data users and data providers. With the associated European climate modeling groups, ENES is preparing for CMIP6 and its large data volume.

IS-ENES also aims to facilitate access to model results for the climate impact community by easing the interface to ESGF data through the Climate4Impact portal. This platform provides tools to explore data, compute indices, perform analyses, and offer guidance to users. The ENES community is also engaged in providing access to global projections for the new Copernicus Climate Change Service.

National Aeronautics and Space Administration High-End Computing Program

Tsengdar Lee (NASA), tsengdar.j.lee@nasa.gov

CMIP datasets have been fundamental and critical for climate trend analysis and resiliency applications. As climate models are becoming more comprehensive Earth system models, requirements for the datasets and data systems are also changing. Under the CMIP6 framework, there are 21 endorsed "MIPs." How much do we know about the different analyses that will be done for these CMIP6 experiments? How do we enable the research community to interrogate a dataset that is easily PBs in size?

We will present current model development efforts at NASA and our approaches to constrain the models. In addition, we will discuss how NASA is supporting America's National Climate Assessment using the CMIP data.

National Computational Infrastructure

Ben Evans (NCI/ANU), Ben.Evans@anu.edu.au

How ready is ESGF for the next stage of CMIP activities? With model data ramping up starting in 2017, key questions revolve around production stability, including service and pre-service benchmark tests and acceptance, data publishing confirmation, ensuring data and service quality processes are in place, status report cards, confirmation of software deployment processes and stable and tested release cycle, data management and replication processes, and preparedness for managing user and model group questions. Many nodes will also provide data-intensive environments and fully featured data services that allow users to probe and analyze the data *in situ*. These services can be extremely valuable, but are they uniform even across the key nodes? Are our loggings and processes in place to ensure that the data is located where it needs to be in response to user demand? I will also touch on some open questions regarding technical architecture and challenges, which we need to consider, in addition to important and immediate issues for data service.

Day 1: Tuesday, December 6, 2016

ESGF Progress and Interoperability

CoG User Interface Working Team

Luca Cinquini (NASA/JPL), Luca.Cinquini@jpl.nasa.gov

Over the past 12 months, the CoG team has worked with the rest of the ESGF collaboration to deploy CoG as the new web front end of the next-generation ESGF software stack. CoG instances are now operational and federated across the ESGF system. Additionally, we have been focusing on implementing several new requirements in support of the upcoming data distribution for the Coupled Model Intercomparison Project, Phase 6 (CMIP6), which will remain our main focus for the next year.

Metadata and Search Working Team

Luca Cinquini (NASA/JPL), Luca.Cinquini@jpl.nasa.gov

During this past year, the ESGF Search and Publishing services have been expanded to include new functionality needed to support a growing federation of nodes and the upcoming CMIP6 massive data volumes. New features include publishing to a local (non-shared) index, supporting atomic metadata updates, searching on datasets with date greater or less than a given value, retracting datasets, and improvements in Wget downloads.

Additionally, we have been experimenting with using Solr Cloud and a new topology architecture.

Publication Working Team

Sasha Ames (DOE/LLNL), ames4@llnl.gov

The Publication Working Team has been on track in improving the publisher software and tools to aid in the publication process. The addition of ESGprep marks a major overhaul in the workings of the esg-publisher component, and CMIP6 requirements have necessitated added features for controlled vocabulary

and quality checking. We will present details of these efforts and additional features, including progress in the ingestion service application programming interface (API) and ideas for refreshing the current publisher implementation.

Node Manager and Tracking/ Feedback Working Team

Sasha Ames (DOE/LLNL), ames4@llnl.gov

The "old" node manager component that previously ran under Tomcat in the ESGF v1.X is no longer deployed in ESGF v2.X. We have a testable replacement service, based on a two-tier architecture for managing node communication that implements the "registration.xml" for dashboard interoperability and also exports several JSON-based RESTful APIs for gathering node information. The node manager has now been tested at several sites, and we plan to deploy a production version in early 2017. The "Tracking/Feedback" effort focuses on a workflow that comprises several new software service modules whose initial purpose is user notification in the event of an update or retraction of a previously downloaded dataset. We will give a brief overview of the tracking and feedback architecture and software implementation progress. Also, we will discuss several additional features, namely notification for "saved" searches and helpful dataset prediction.

Stats and Dashboard Working Team

Alessandra Nuzzo (ENES/CMCC), alessandra.nuzzo@cmcc.it; Maria Mirto (ENES/CMCC), maria.mirto@cmcc.it

The Stats and Dashboard Working Team has been on track to improve the metrics modules for ESGF experiments (mainly, but not only, CMIP5, Obs4MIPS, and CORDEX). Since September 2016, weekly meetings have focused mainly on requirements for validation, testing, and feedback from the testing sites for both the back and front ends as well as preparation for the coming releases. A short overview about the main results achieved, status of working group activities, and plan for the next months will be presented.

Identity Entitlement Access Management Working Team

Philip Kershaw (ENES/BADC), philip.kershaw@stfc.ac.uk; Rachana Ananthakrishnan (DOE/ANL), ranantha@uchicago.edu

This year, the Identity Entitlement Access Management Working Team (IdEA) team focused on integration of OAuth 2.0 for user delegation. Earlier in the year, the live access server (LAS) was linked up with the Centre for Environmental Data Analysis's (CEDA's) OAuth service to implement a delegation flow. The CEDA service has been updated over the course of the year; reviewing the latest Python packages, the decision was made to port the service to use OAuthLib. This decision allows the service to be deployed as a standard Django package and will facilitate future migration to OpenID Connect, which OAuthLib now supports. In addition, the service has been packaged using Ansible to facilitate its integration into ESGF. This will enable easier installation and subsequent roll out by other IdPs in the federation. Further implementation steps involved updating dependent ESGF components so that they can use the new service, including CoG and ORP in the Data Node. Work also has been done to investigate the steps needed for integration with Globus and the Compute Node. More advanced use cases also are being explored, including two-stage delegation for IS-ENES2 Climate4Impacts Portal (KNMI) with the downscaling portal (University of Cantabria).

Compute Working Team

Charles Doutriaux (DOE/LLNL), doutriaux1@llnl.gov; Daniel Duffy (NASA/GSFC), daniel.q.duffy@nasa.gov

ESGF's main goal for the Compute Working Team (CWT) is to facilitate advancements in Earth system science with a primary mission of supporting CMIP activities. In preparation for future climate assessments, the CWT has been working toward a goal of providing server-side analytics capabilities through the development of server-side APIs and client-side (end user) APIs. This talk will provide an overview of the CWT, current status, and future goals. In addition, we will describe advances made by the CWT on APIs along with various implementations made over the last year. An overview of projects by NASA, Ouranos, and Euro-Mediterranean Center on Climate Change will be provided. In addition, a demonstration of how

the Python client API can be used to access analytics services will be shown.

Errata Service

Guillaume Levavasseur (ENES/IPSL), glipsl@ipsl.jussieu.fr; Atef Ben Nasser (ENES/IPSL), abennasser@ipsl.jussieu.fr; Mark A. Greenslade (ENES/IPSL), momipsl@ipsl.jussieu.fr; Merret Buurman (ENES/DKRZ), buurman@dkrz.de; Sébastien Denvil (ENES/IPSL), sebastien.denvil@ipsl.jussieu.fr; Katharina Berger (ENES/DKRZ), berger@dkrz.de; Martina Stockhause (ENES/DKRZ), stockhause@dkrz.de

Recording and tracking the reasons for dataset version changes is important, due to the inherent complexity of experimental protocols for projects such as CMIP5/6. The currently established system makes it impossible for scientists to know easily whether the data in hand is deprecated and/or replaced and corrected by a newer version. Also, accessing a description of this issue is difficult.

IPSL is finalizing a new ESGF Errata Service to:

- Provide timely information about known issues. Within the ES-DOC ecosystem, the errata web service front end displays the whole list of known issues. The list can be filtered by several useful parameters, such as issue severity or status. Three tabs describe each issue, providing (1) information details, (2) graphics or pictures to illustrate the issue, and (3) list of affected datasets.
- Allow identified and authorized actors to create, update, and close an issue. We developed a piece of software that enables interaction with the Errata Service. It can be used to create, update, close, and retrieve issues. The client is aimed to be used by publishing teams, so that they can directly describe problems as they are discovered.
- Enable users to query about modifications and/or corrections applied to the data in different ways. The errata web service provides an API to query the issue database. The end users can submit one or several file or dataset identifiers to receive all annotations related to each corresponding issue. This search API also is able to retrieve the issues that affect an MIP variable or experiment.

To succeed, the Errata Service exploits the Persistent IDentifier (PID) attached to each dataset during the ESGF publication process. The PIDs enable requests of the Handle Service to get the version history of files/datasets. Consequently, IPSL is working closely with DKRZ on the required connections and APIs for the two services. The ESGF implementation of the citation service is coordinated by the ESGF-QCWT. [Errata Service development deployment (http://test.errata.es-doc.org/).]

Quality Control Working Team: Data Citation Service for CMIP6—Status and Timeline

Martina Stockhause (ENES/DKRZ), stockhause@dkrz.de; Katharina Berger (ENES/DKRZ), berger@dkrz.de; Guillaume Levavasseur (ENES/IPSL), glipsl@ipsl.jussieu.fr

The review of the CMIP6 data citation procedure resulted in the requirement of a citation possibility prior to long-term data archiving in the IPCC DDC (Data Distribution Centre) hosted at DKRZ. The presentation will give an overview of the Data Citation concept, with emphasis on technical requirements and dependencies on developments by other teams. Implementation of the different components will be reviewed to provide a service development status and timeline toward its operability. ESGF implementation of the citation service is coordinated by the ESGF-QCWT.

Installation Working Team

Prashanth Dwarakanath (ENES/Liu), pchengi@nsc.liu.se

Installation and maintenance of an ESGF node is, unfortunately, still a laborious, risky, and over-complicated process. Additionally, the current installation software is not modular enough and difficult to evolve in the long-term. On the other hand, easy installation and upgrading of the underlying software stack is critical to ESGF adoption and success. This talk will present a brief analysis of the major shortcomings of the ESGF installation model and outline alternatives for transitioning to a more robust and reliable framework.

Docker for ESGF

Luca Cinquini (NASA/JPL), Luca.Cinquini@jpl.nasa.gov

Docker is becoming a mainstream technology for packaging, deploying, and operating complex applications on multihost environments, including in the cloud. This talk will report on an exploratory effort to run an ESGF node as a set of interacting Docker containers, each running a specific ESGF service. Docker could be useful to ESGF in many respects in that it would:

- Greatly simplify installing and upgrading an ESGF node.
- Make the installer software much more modular, maintainable, and upgradable.
- Allow scaling up of services, as needed, and deployment in the cloud.
- Make adding new services (e.g., nginx, other Python apps, and other Java web apps) a much simpler process.

Is Docker the future of ESGF?

International Climate Network Working Group

Eli Dart (DOE/ESnet), dart@es.net

The International Climate Network Working Group (ICNWG) is dedicated to improving data transfer performance between the major climate data centers, and from climate data centers to the users of climate model data worldwide.

This talk will discuss ICNWG efforts in 2016 and in 2017 and beyond.

Data Transfer Working Team

Lukasz Lacinski (DOE/ANL), lukasz@uchicago.edu

The Data Transfer Working Team (DTWT) has worked on both improving data transfer performance and adding new features that simplify transferring datasets to and from the data node. Two new features that can be optionally enabled have been added to the ESGF authorization callout: sharing and write access for users with the publisher role. The DTWT has worked with the CoG User Interface Working Team and Publication Working Team to add Globus URLs as a new method of accessing datasets. The latest ESGF installer provides all the aforementioned features. DTWT also added a new transport method to Synda that uses Globus Transfer Service.

Security Working Team

George Rumney (NASA/GSFC), george.rumney@nasa.gov

The ESGF Software Security Working Team (SSWT) was established in the wake of the multisite compromise in 2015. Recovery focused on remediation of identified flaws, short-term correction of some software engineering methods, and creation of a software security plan (http://esgf.llnl.gov/media/pdf/ESGF-Software-Security-Plan-V1.0.pdf). Progress since then has been modest, and significant challenges remain. This talk will highlight the current challenges and near-term goals.

Replication and Versioning Working Team

Stephan Kindermann (ENES/DKRZ), kindermann@dkrz.de; Tobias Weigel (ENES/DKRZ), weigel@dkrz.de

Together with the Data Transfer Working Team and International Climate Network Working Group, the Replication and Versioning Working Team worked on replication tests between sites as well as on improvements of replication-related software components (e.g., Synda). Overall progress in 2016 was hindered mainly by the new setups of data transmission network-related hardware infrastructure at sites. A short status of the situation and plans for 2017 will be discussed.

In close collaboration with the Publication Working Team, the versioning procedure for CMIP6 was improved. The PID Services Team will give an update on PID-related versioning aspects.

Persistent Identifier Services

Tobias Weigel (ENES/DKRZ), weigel@dkrz.de; Stephan Kindermann (ENES/DKRZ), kindermann@dkrz.de; Katharina Berger (ENES/DKRZ), berger@dkrz.de

The PID Services Team will give a brief update on milestones reached since the last meeting, current status, and next activities. Past activities that will be reported on include development of the necessary software components for registering and managing

PIDs for CMIP6. The PID Services Team collaborated with the Publication and Errata Service teams on this and coordinated necessary changes with the Climate Model Output Rewriter (CMOR) development team. Upcoming action items include the full operational rollout at multiple sites and development of dedicated user tools and services or integration of PIDs into existing solutions.

User Working Team

Torsten Rathmann (ENES/DKRZ), rathmann@dkrz.de; Matthew Harris (DOE/LLNL), harris112@llnl.gov

Help pages were completely moved to CoG and were extended by the new authorization for ESGF data access and OPeNDAP). The Wget tutorial was completely revised.

Operative support for users has been continued via the mailing list (esgf-user@lists.llnl.gov). Compared to the first half of 2015, the number of user questions decreased by 47 % (January-June 2016). More results from support statistics will be shown.

Day 2: Wednesday, December 7, 2016

Advanced Computational Environments and Data Analytics

Overview of the ESGF Compute Working Team and Target Milestones

Charles Doutriaux (DOE/LLNL), doutriaux1@llnl.gov; Daniel Duffy (NASA/GSFC), daniel.q.duffy@nasa.gov; Jason Boutte (DOE/LLNL), boutte3@llnl.gov; Thomas Maxwell (NASA/GSFC), thomas.maxwell@nasa.gov; Tom Landry (CRCM), tom.landry@crim.ca; S. Fiore (ENES/CMCC), sandro.fiore@cmcc.it; Dean N. Williams (DOE/LLNL), Williams13@llnl.gov

ESGF's main goal is to facilitate advancements in Earth system science with a primary mission of supporting CMIP activities. In preparation for future climate assessments, CWT has been working toward a goal of providing server-side analytics capabilities through the development of server-side APIs and client-side (end user) APIs. This talk will provide an overview of the CWT, current status, and future goals. In addition, we will describe advances made by the CWT on APIs along with various implementations made over the last year. An overview of projects by NASA, Ouranos, and Euro-Mediterranean Center on Climate Change will be provided. In addition, a demonstration of how the Python client API can be used to access analytics services will be shown.

Compute Working Team End-User Application Programming Interface

Jason Boutte (DOE/LLNL), boutte3@llnl.gov; Charles Doutriaux (DOE/LLNL), doutriaux1@llnl.gov

The ESGF Compute Working Team end-user API was created to leverage the power of the Web Processing Service (WPS) interface standard. A WPS server can expose large-scale computational processing to users that are location agnostic, allowing computations to be performed where the data resides, thus saving bandwidth and time. To execute a WPS process, a user would normally be confronted with lengthy and intricate URLs. To simplify the task of using WPS processes, a well-defined climatology specific API was planned and an object-oriented Python end-user API was created. With the API, users are eased into the use of these WPS processes, allowing them to easily harness the power they provide.

The Climate Data Analytic Services Framework

Thomas Maxwell (NASA/GSFC), thomas.maxwell@nasa.gov; Daniel Duffy (NASA/GSFC), daniel.q.duffy@nasa.gov

Faced with unprecedented growth in climate data volume and demand, NASA developed the Climate Data Analytic Services (CDAS) framework. This framework enables scientists to execute data processing workflows combining common analysis operations in a high-performance environment close to the massive data stores at NASA. The data is accessed in standard formats (e.g., NetCDF and HDF) in a POSIX file system and processed using vetted climate data analysis tools (e.g., ESMF, CDAT, and NCO). A dynamic caching architecture enables interactive response times. CDAS utilizes Apache Spark for parallelization and a custom array framework for processing huge datasets within limited memory spaces.

CDAS services are accessed via a WPS API being developed in collaboration with the ESGF CWT

to support ESGF server-side analytics. The API can be accessed using direct web service calls, a Python script, Unix-like shell client, or JavaScript-based web application. New analytic operations can be developed in Python, Java, or Scala. Client packages in Python, Scala, or JavaScript contain everything needed to build and submit CDAS requests.

The CDAS architecture brings together the tools, data storage, and high-performance computing required for timely analysis of large-scale datasets, where the data resides, to ultimately produce societal benefits. It is currently deployed at NASA in support of the Collaborative REAnalysis Technical Environment (CREATE) project, which centralizes numerous global reanalysis datasets onto a single advanced data analytics platform. This service enables decision makers to investigate climate changes around the globe, inspect model trends and variability, and compare multiple reanalysis datasets.

The Ophidia Big Data Analytics Framework

S. Fiore (ENES/CMCC), sandro.fiore@cmcc.it; C. Doutriaux (DOE/LLNL), doutriaux1@llnl.gov; J. Boutte (DOE/LLNL), boutte3@llnl.gov; D. Elia (ENES/CMCC), donatello.elia@cmcc.it; A. D'Anca (ENES/CMCC), alessandro.danca@cmcc.it; C. Palazzo (ENES/CMCC), cosimo.palazzo@cmcc.it; D. N. Williams (DOE/LLNL), williams13@llnl.gov; G. Aloisio (ENES/CMCC), giovanni.aloisio@unisalento.it

The Ophidia project is a research effort on big data analytics facing scientific data analysis challenges in the climate change domain. 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 ("datacubes"). The project relies on a strong background in high-performance database management and OLAP systems to manage large scientific datasets.

The Ophidia analytics platform provides several data operators to manipulate datacubes,and array-based primitives to perform data analysis on large scientific data arrays. Metadata management support also is provided.

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) with a complete set of commands.

The presentation will give an overview of the new, recently released, in-memory analytics engine, which allows fast data analysis on large amounts of data, outperforming the previous approach based on MySQL servers.

Ongoing activities of the ESGF CWT Working Team also will be presented.

PAVICS: A Platform to Streamline the Delivery of Climate Services

David Huard (CRCM), Huard.David@ouranos.ca; Tom Landry (CRCM), tom.landry@crim.ca

Ouranos is a Montreal-based consortium on regional climatology playing the role of a catalyst for climate adaptation. Beyond creating simulation ensembles with the Canadian Regional Climate Model, we also work on translating climate science into services and products tailored to the needs of decision makers and scientists from other disciplines. As demand for climate services grows, we felt the need to develop software to speed up and standardize the production of climate scenarios, both for our own needs and those of the climate research community. With funding from the CANARIE research software program, we launched the PAVICS project, one objective of which is to create a web platform to facilitate data distribution, streamline standard climate analyses, and serve as a backbone for various tailored web applications and services. Ouranos works closely with CRIM, an IT Applied Research Centre focusing on innovation and collaborative development.

In the spirit of the ESGF Compute Working Team vision, we are working to co-locate the heavy number crunching close to the data stores on the Calcul-Québec (HPC) infrastructure. The system architecture is based on Birdhouse, a collection of independent WPSs manageable as workflows. Birdhouse bundles THREDDS, ncWMS, and OCGIS, as well as identity providers and data sources key to ESGF. PAVICS implements data harvesting, crawling, and updates to Solr. It also offers search capabilities found in ESGF Search's RESTful API. Additionally, we are integrating geospatial management and processing capabilities

from GeoServer. This server is used to store region definitions (Bukovski regions, countries and states, watersheds), apply geometrical transformations through WPS (union, buffer, polygonize, rasterize), and request base layers or specific records (WMS, WFS). Beyond a public list of common regions, users will be able to upload custom regions and maintain their own collection for later use. Most of the user interfaces rely on selected WPS services and workflow schema to automatically create the necessary widgets to hold inputs for the climate analyses and show their results. Web-based tools and widgets are developed with modern web frameworks such as React-Redux, OpenLayers 3, Cesium, and Plotly. Services required to build climate scenarios will be created (e.g., bias correction and spatial analogs), as well as tools to build, archive, and run workflows.

Server-Side Computing Services Provided by IS-ENES Through the Climate4Impact Platform

Christian Pagé (ENES/IPSL), christian.page@cerfacs.fr; Maarten Plieger (ENES/KNMI), maarten.plieger@knmi.nl; Wim Som De Cerff (ENES/KNMI), wim.som.de.cerff@knmi.nl; Manuel Vega (ENES/University of Cantabria), manuel.vega@unican.es; Sandro Fiore (ENES/CMCC), sandro.fiore@cmcc.it

Within the FP7 European projects IS-ENES/IS-ENES2 that work with the European climate model data infrastructure, a web portal [called Climate 4Impact (C4I)] tailored for climate change impact communities is being developed. It has evolved from a climate web portal to a platform offering standard Open Geospatial Consortium (OGC) services that can be used to build targeted and specific climate data portals.

One of the services made available by C4I is server-side computing of climate indices and simple statistics through use of the python package icclim developed within the IS-ENES2 and CLIPC European projects. Accessing icclim services is done using OGC WPS processes. This enables users to perform first-step or final analyses and data reduction on the C4I server prior to download and/or visualization.

The aim is stronger integration among ongoing developments within IS-ENES/C4I/icclim, the Copernicus CLIPC project, and the API being developed within the ESGF Compute Working Team. Some possible future integration with EUDAT Services also will be discussed.

CAFE: A Framework for Collaborative Analysis of Distributed Environmental Data

Hao Xu (China/Tsinghua University), xuhao13@mails.tsinghua.edu.cn; Sha Li (China/Tsinghua University), lis14@mails.tsinghua.edu.cn; Wenhao Dong (China/Tsinghua University), dongwh12@mails.tsinghua.edu.cn; Wenyu Huang (China/Tsinghua University), huangwenyu@tsinghua.edu.cn; Shiming Xu (China/Tsinghua University), xusm@tsinghua.edu.cn; Yanluan Lin (China/Tsinghua University), yanluan@mail.tsinghua.edu.cn; Bin Wang (China/Tsinghua University), wab@tsinghua.edu.cn; Fanghua Wu (China/Tsinghua University), wufh@cma.gov.cn; Xiaoge Xin (China/Tsinghua University), xinxg@cma.gov.cn; Li Zhang (China/Tsinghua University), zhangli@cma.gov.cn; Zaizhi Wang (China/Tsinghua University), wzz@cma.gov.cn; Tongwen Wu (China/Tsinghua University), twwu@cma.gov.cn; Yuqi Bai (China/Tsinghua University), yuqibai@mail.tsinghua.edu.cn

As the amount of information about our environment expands exponentially on a global scale, researchers are challenged to remain efficient when analyzing data maintained in multiple data centers. We present a new software package named Collaborative Analysis Framework for Environmental Data (CAFE). CAFE is dedicated for collaborative analysis of large volumes of distributed environmental data. It is designed to execute analytic functions on the node where the data are stored. Multiple nodes can collaborate with each other to perform complex data analysis. A web-based user interface allows researchers to search for data of interest, submit analytic tasks, check the status of tasks, visualize analysis results, and download those results. Compared with existing web-based environmental data analysis systems, CAFE dramatically reduced the amount of data that had to be transmitted from data centers to researchers. CAFE demonstrates great promise for enabling seamless collaboration among multiple data centers and facilitating overall research efficiency in scientific data analysis.

Embedded Domain-Specific Language and Runtime System for Progressive Spatiotemporal Data Analysis and Visualization

Cameron Christensen (University of Utah), cam@sci.utah.edu; Shusen Liu (DOE/LLNL), liu42@llnl.gov; Giorgio Scorzelli (DOE/LLNL), scorzelli2@llnl.gov; Ji-Woo Lee (DOE/LLNL), lee1043@llnl.gov; Peer-Timo Bremer (DOE/LLNL), bremer5@llnl.gov; Valerio Pascucci (University of Utah), pascucci@sci.utah.edu

As our ability to generate large and complex climate simulation datasets grows, accessing and processing these massive data collections is increasingly becoming the primary bottleneck in scientific analysis. Challenges include retrieving, converting, resampling, and combining remote and often disparately located data ensembles with only limited support from existing tools. In particular, current solutions predominantly rely on extensive data transfers or large-scale remote computing resources, both of which are inherently offline processes with long delays and substantial repercussions for any mistakes. Such workflows impede scientific discovery by severely limiting the flexible exploration and rapid evaluation of new hypotheses that are crucial to the scientific process.

We present an embedded domain-specific language (EDSL) specifically designed for the interactive exploration of large-scale, remote data. Our EDSL allows users to express a wide range of data analysis operations in a simple and abstract manner. The underlying runtime system transparently resolves issues such as remote data access and resampling while at the same time maintaining interactivity through progressive and interruptible computation. This allows, for the first time, interactive remote exploration of massive datasets, such as the 7-km NASA GEOS-5 Nature Run simulation, which previously have only been analyzed offline or at reduced resolution.

Day 2: Wednesday, December 7, 2016

Coordinated Efforts with Community Software Projects

CMIP6 Standards Enabling Management, Search and Interpretation of Model Output

Karl Taylor (DOE/LLNL/PCMDI), taylor13@llnl.gov; Paul J. Durack (DOE/LLNL/PCMDI), pauldurack@llnl.gov; Denis Nadeau (DOE/LLNL), nadeau1@llnl.gov; Sasha Ames (DOE/LLNL), ames4@llnl.gov

As specifications for CMIP6 model output and metadata become finalized, ESGF requirements have become clearer. A brief review of the CMIP6 requirements will emphasize areas perceived to be possibly problematic. Various ESGF software services will rely on certain global attributes that identify and describe essential aspects of model simulations. These global attributes must be drawn from CVs, which have been defined, for example, for model and experiment names, grid descriptions, and frequencies. An overview will be provided describing how the global attributes will be used to construct file names and directory structures, as well as their use in defining the CMIP6 Data Reference Syntax, which enables faceted searches and links to model and experiment documentation. Despite the growth of CMIP (nearly 250 experiments are now planned), the strictly enforced data requirements, expanded capabilities of the CMIP-supporting infrastructure, and increased emphasis on transparency (e.g., via web-based services for sharing code and exposing issues) promise to serve an expanding community of scientists and stakeholders with an interest in climate and climate change.

CMIP6 ESGF Tier 1 and Tier 2 Nodes

Sébastien Denvil (ENES/IPSL), sebastien.denvil@ipsl.jussieu.fr; Michael Lautenschlager (ENES/DKRZ), lautenschlager@dkrz.de

The ESGF Executive Committee tasked Michael Lautenschlager and Sébastien Denvil with collecting and discussing ESGF Tier 1 and Tier 2 data node requirements. After discussions and iterations within the ENES Data Task Force, we came up with an initial plan. We collected feedback from groups like CDNOT

(CMIP Data Node Operations Team) and will present the main outcomes of this process.

Tier 1 data node requirements for the ESGF infrastructure will cover the level of service (90% to 95% uptime), installation of the full software stack, contribution to development and maintenance, support for Tier 2 data nodes, and support for data providers. The level of service needs to be at the core of those requirements. For example:

- NAGIOS-like monitoring for certificates (host and Globus certificates).
- NAGIOS to monitor ESGF nodes and guard against
- Expired certificates
- http/https endpoints unavailable
- GridFTP endpoints unavailable
- Tier 1 node will be responsible for monitoring data node publishing.
- Tier 1 node will self monitor.

For Tier 1, requirements for data projects (i.e., CMIP6) will cover the following major items:

- Spinning disks and compute resources contribution to the Data Project for data replication and analysis purposes.
- Tier 1 will have to optimize nominal bandwidth of 10 GBit/s that will result in 30–50% for real replication bandwidth. This, together with specification of the core dataset, defines the CMIP6 replication strategy.
- Tier 2 will have to warranty a bandwidth of 12 GBit/s for data provision. CMIP5 experience shows that each data node provides 10 times the data it hosts over a period of 4 years, and the average available network bandwidth should cover this.
- Single Tier 1: about 20 PB for long-term archiving of reference data from the CMIP6 data (volume not yet clear).
- Tier 1: tapes to fill the storage gap in case of insufficient disc space for initial data publication and data replication.

There is an ongoing proposition to enable ESGF to exclude a data node that does not satisfy all the CMIP6 requirements or a data node that will degrade the federation usability. Implementation is currently under discussion, but we can anticipate that when the governance is set, it will be the responsibility of Tier 1 nodes to enforce the decision and make it operable.

CMIP6 "Impact" on Scientific Community

Sergey Nikonov (NOAA/GFDL), serguei.nikonov@noaa.gov; V. Balaji (NOAA/GFDL), balaji@princeton.edu; Aparna Radhakrishnan (NOAA/GFDL), aparna.radhakrishnan@noaa.gov; Hans Vahlenkamp (NOAA/GFDL), hans.vahlenkamp@noaa.gov

The results of resource estimations for the forthcoming CMIP6 are shown. The analysis is done based on an XML database designed and populated with MIPs requests by Martin Juckes (CEDA). The main goal is to show impact of CMIP6 on both sides of climate community—data producers and data analyzers. The results characterize the output volume and corresponding efforts demanded for publishing planned experiments. The total amount of generated data from all participating modeling centers was estimated and compared with volume of CMIP5. There was also an attempt to assess scientific human resources being spent for QC of published data and analyzing and utilizing the CMIP6 outcome.

Control Vocabulary Software Designed for CMIP6

Denis Nadeau (DOE/LLNL), nadeau1@llnl.gov; Karl Taylor (DOE/LLNL), taylor13@llnl.gov; Sasha Ames (DOE/LLNL), ames4@llnl.gov

CMIP6 contains more activities and many more experimentations than its predecessor CMIP5. To compare this model output increase, a standard was created to ensure information homogeneity. This standard creates the ability to understand and exchange data between different Earth science groups. The Climate Forecast (CF-1) compliance already insures interoperability between different visualization and analysis software, but an extension of CF-1 is necessary to accommodate CMIP6 outputs. Variable names, global attributes, and variables attributes need to be set to facilitate comparison between similar geophysical variables coming from different provenances. The CMIP6_CV Python program insures control of the different attributes needed before publication of CMIP6 and distribution of resulting NetCDF model outputs. CMIP6_CV ensures that all required

attributes are present, and if one is missing, the program will not allow publication of the file. Other missing attributes or wrong attributes can sometime be created or replaced automatically by the CMIP6_CV Python program and warn users about the changes that have been made. CMIP6_CV establishes common ground between model outputs, which facilitates analyses for scientists studying climate change. CMIP6_CV is flexible and can be used by similar projects that necessitate a control vocabulary.

Developing a Vocabulary Management System for Data Reference Syntax Using Linked Data Technologies in the Climate Information Platform for Copernicus Project

Ruth Petrie (ENES/CEDA), ruth.petrie@stfc.ac.uk; Phil Kershaw (ENES/CEDA), philip.kershaw@stfc.ac.uk; Ag Stephens (ENES/CEDA), ag.stephens@stfc.ac.uk; Antony Wilson (ENES/CEDA), antony.wilson@stfcac.uk

CEDA host data centers manage a large and varied archive of climate and Earth observation data. CEDA is the lead partner in the Climate Information Platform for Copernicus (CLIPC) project. One aim of the CLIPC project is to be a single point of access for a variety of climate data records.

Within CLIPC, many highly heterogeneous datasets were published through ESGF, such as satellite and *in situ* observational data and climate impact indicators. Within each of these communities, different descriptive metadata is required to construct a useful Data Reference Syntax (DRS) when compared with the traditional model-based data published through ESGF. The European Space Agency (ESA) Climate Change Initiative (CCI) project is the most mature of these, having a dedicated ESA-CCI Open Data Portal. A CCI DRS was developed to provide a single authoritative source for cataloguing and searching the CCI data, and this has been successfully deployed for the ESA-CCI Open Data Portal and the CLIPC portal. Use of the Simple Knowledge Organization System (SKOS) and Web Ontology Language (OWL) to represent the DRS are a natural fit, providing controlled vocabularies as well as representing relationships between similar terms used in different communities.

The CLIPC portal supports data discovery based on the OGC CSW specification and ESGF's powerful faceted search. These services provide complementary content at different levels of granularity, and therefore a common data model was needed. Key terms are defined in vocabularies serialized in SKOS and OWL and are accessible from a central vocabulary server, which can be queried from applications consuming metadata content.

Exploiting the vocabulary service, it has been possible to develop an innovative solution tagging ISO19115 records for CSW with the equivalent vocabulary terms used for the ESGF faceted search system. SKOS provides a tool to manage CVs with semantic relationships and arbitrary tagging of datasets. In this way, it has been possible to create enhanced metadata records and a search interface, combining CSW and ESGF search results driven by a faceted search interface managed and populated from the vocabulary server.

DKRZ ESGF-Related Infrastructure and CMIP6 Services

Stephan Kindermann (ENES/DKRZ), kindermann@dkrz.de; Michael Lautenschlager (ENES/DKRZ), lautenschlager@dkrz.de; Legutke (ENES/DKRZ), legutke@dkrz.de; Katharina Berger (ENES/DKRZ), berger@dkrz.de; Martina Stockhause (ENES/DKRZ), stockhause@dkrz.de

The DKRZ will coordinate German ESGF-related activities, as well as the national CMIP6 contribution. Besides hosting ESGF nodes and providing support for CMIP6 data ingest, data publication, long-term archiving, and data citation, the DKRZ in engaging in a set of new activities to support the national and international climate community, including

- Integration of CMOR with CDO to support climate modelers in generating CMIP6-compliant data.
- Establishment of a national CMIP data pool acting as a replica cache of often-needed CMIP5- and CMIP6-related data, which can be exploited for efficient data analysis and evaluation.
- Development of a generic data QA tool supporting CMIP6 data quality checking (going beyond "pure CMIP6 convention compliance checking" (e.g., CF compliance checking and outlier detection).
- Establishment of a persistent identification infrastructure integrated with ESGF and supporting CMIP6.

- Extension of the data citation service on long-term archived data, with a citation possibility for the evolving CMIP6 data.
- Integration of the CMIP data pool and ancillary metadata into the IPCC DDC AR6 reference data archive and improvements in the integration of the IPCC DDC in IPCC's assessment process.
- Development of a Web Processing Service framework to support future data processing service provisioning, including supporting conda sw packaging and Docker.

The talk will summarize the current status of these activities as well as next steps and plans.

The IPCC DDC in the Context of CMIP6

Martina Stockhause (ENES/DKRZ), stockhause@dkrz.de; Michael Lautenschlager (ENES/DKRZ), lautenschlager@dkrz.de; Stephan Kindermann (ENES/DKRZ), kindermann@dkrz.de

The CMIP6 data underlying the IPCC AR6 of WG1 will be transferred to the long-term archive of the IPCC DDC (Data Distribution Centre) at DKRZ to build the Reference Data Archive of the global climate model output. Apart from the data, different pieces of data-related information are to be integrated in the archive to enrich data documentation for interdisciplinary long-term use. The second task of the DDC within CMIP6 is the support of IPCC authors by opening DKRZ's CMIP Data Pool for IPCC authors. The CMIP6 data subset in the CMIP Data Pool will be the source for the AR6 Reference Data Archive.

The presentation will give an overview of the different connections between IPCC DDC and CMIP6. A detailed description of the transfer of data and metadata from ESGF and repositories of ancillary metadata (e.g., errata, citation, and model descriptions) will be given, with special emphasis on requirements for ancillary metadata providers.

Persistent Identifiers in CMIP6

Merret Buurman (ENES/DKRZ), buurman@dkrz.de; Tobias Weigel (ENES/DKRZ), weigel@dkrz.de; Stephan Kindermann (ENES/DKRZ), kindermann@dkrz.de; Katharina Berger (ENES/DKRZ), berger@dkrz.de; Michael Lautenschlager (ENES/DKRZ), lautenschlager@dkrz.de

All CMIP6 files and datasets in ESGF will receive a persistent identifier (PID). A PID is a string that can be resolved to a landing page that shows some minimal metadata (e.g., information about data storage locations and checksums). In CMIP6, PIDs will also be used to record relationships between data objects (e.g., which dataset version consists of which file sets, or which dataset version is replaced by which new version). Also, information on replication sites and dataset errata (see abstract on QCWT Errata service by Levavasseur et al.) will be stored. These metadata will be available even after un-publication of the data from the ESGF data nodes, so researchers can find metadata on a data object they have been using even if the data were outdated; in particular, they can find out if the data object is outdated and which new version replaces it.

Technically, the system behind the CMIP6 PIDs is the Handle System (www.handle.net/), which also underlies the Digital Object Identifier (DOI) system but does not aim for citability. The handles are registered at the ESGF Handle Service during the ESGF publication process, using the esgfpid library (github.com/IS-ENES-Data/esgf-pid) called by the ESGF publisher. To cushion temporary publication peaks, a message queuing system ensures that the publication process is not delayed and that no PID registration request is lost.

Users can view and access the PIDs from the CoG front end, where PIDs are displayed for every dataset and file. Furthermore, the PID strings are contained in the files' NetCDF headers. Permanently bound to the file, they will help researchers find information about data they have used, found, and received for years to come, potentially beyond the scope of ESGF. PIDs also will provide a sustainable foundation for data management tools and "intelligent" client-side tools (e.g., exploiting the versioning and replication information).

The talk will outline the current status of the technical PID infrastructure as well as its integration with the ESGF publisher and first test results. How PIDs

will support CMIP6 versioning and CMIP6 errata annotation also will be discussed.

ES-DOC and ESGF Errata Services

Atef Ben Nasser (ENES/IPSL), abennasser@ipsl.jussieu.fr; Mark Greenslade (ENES/IPSL), momipsl@ipsl.jussieu.fr

The Earth System documentation (ES-DOC) started in the documentation of the CMIP6 project, putting into good use experience gained from CMIP5. With further formalization and a clear set of use cases, the process has been streamlined and rendered less of a burden as a large chunk has been automated, a beta period scheduled, and each and every step thoroughly documented.

The ES-DOC is ready for community review as of November 2016, and the beta testing phase will occur during October 2016-February 2017. The full community release is scheduled for March 2017.

In the context of overseeing data quality, the ESGF Errata service was encapsulated in the ES-DOC structure and built on top of the Handler service that will be deployed in the next release cycle. Consuming PIDs from Handler Service, the ESGF Errata service is guided by a specifically built algorithm that extracts metadata regarding issues that may or may not affect the quality of datasets and files and cause newer versions to be published. This new structure has been designed keeping in mind usability by end users specialized in the publishing process or other scientists requiring feedback on reliability of needed data.

The expected outcome from both ES-DOC and the Errata service project is to increase data quality. Providing this critical information for end users requires a well-defined process and is ensured by exploring incoming features of the ESGF ecosystem.

National Computational Infrastructure's Research Data Services: Providing High-Quality Data to Enable Climate and Weather Science

Claire Trenham (NCI/ANU), claire.trenham@anu.edu.au; Kelsey Druken (NCI/ANU), kelsey.druken@anu.edu.au; Adam Steer (NCI/ANU), adam.steer@anu.edu.au; Jon Smillie (NCI/ANU), jon.smillie@anu.edu.au; Jingbo Wang (NCI/ANU), jingbo.wang@anu.edu.au; Ben Evans (NCI/ANU), Ben.Evans@anu.edu.au

The National Computational Infrastructure (NCI) hosts over 10 PB of broad-based, nationally significant research data collections, including climate and weather, water and ocean, satellite Earth observations, reanalysis, elevation and bathymetry, geodetic, other geosciences, astronomy, social sciences, and bioinformatics. We have a particular focus on Earth systems data (including CMIP) as part of the National Environmental Research Data Interoperability Platform (NERDIP). The data is, where possible, stored in a standard format (NetCDF) in clear directory structures. Quality is assured by compliance to metadata standards and tested against common tools and protocols, and data management plans are created to provide full data collection metadata and provenance information.

Open datasets are published through our Research Data Services, including THREDDS, and also are available on NCI's high-performance file system. These data services are accessible from anywhere in the world to allow broader access than just within NCI's high-performance environment, and data therefore is available to a wider community of scientists and visualization specialists via remote file access, download, or OGC-compliant services. NCI is an ESGF node for the publication and replication of CMIP and other international climate data, which enables bulk data transfer for greater access to CMIP data by the Australian climate research community and distribution of their modeled data.

A powerful use of these data facilities is to provide high-performance access to data-enabling advanced virtual laboratories, particularly the Climate and Weather Science Laboratory as well as other community virtual laboratories and portals. NCI also makes the data available via our interactive Virtual Desktop Infrastructure (VDI) and Raijin

supercomputer. The VDI enables climate science, including multimodel intercomparison and detection and attribution work, by providing access to the data collocated with programming, analysis, and visualization tools (including Python, UV-CDAT, and VisTrails). VDI also provides remote batch submission capability to the HPC infrastructure.

We provide a metadata catalog of our data holdings, through which anyone can search our data collections and datasets and find information on how to access the data. In particular, the catalogue information shows location on NCI's file systems and via web data services as available.

NCI supports the widespread use of these datasets, without the need for scientists to move data to their local workstations, by enabling the data to be accessed remotely from anywhere via standard web protocols or, for Australian researchers, directly at NCI in a value-added virtual environment equipped with the typical datasets and tools a climate or weather scientist is likely to need.

Automating Data Synchronization, Checking, Ingestion, and Publication for CMIP6

Ag Stephens (ENES/CEDA), ag.stephens@stfc.ac.uk; Alan Iwi (ENES/CEDA), alan.iwi@stfc.ac.uk

CEDA is responsible for providing the ESGF United Kingdom data node, which involves publication of Met Office Hadley Centre datasets provided for CMIP6. Following lessons learned from CMIP5, we have developed an automated system for remotely synchronizing the contents of the Met Office MASS tape archive to CEDA. This system builds upon a RabbitMQ message service that prompts for actions such as "publish-to-ESGF" and "withdraw-from-ESGF."

The CEDA ingestion pipeline is complex because it requires access to multiple services running across a range of platforms. To automate the pipeline, we have developed a simple client-server architecture in which a collection of distributed workers query a centralized database for instructions to manage their own processes and workloads. This approach allows each independent worker to run his or her own controller under a different user identification with access to specific resources relevant to its stage in the processing chain (e.g. "sync," "validate," "ingest," and "publish"). Individual (client) workers have no knowledge of other workers because all states and decisions about which controller should be run on each dataset are managed through the database (server).

The data model uses the ESGF dataset as its unit of currency, and the system records each "do" (and, when problems occur, "undo") event that takes place across all platforms. A Django web application provides queryable views of important components such as files, ESGF datasets, and events, as well as a "Global Pause" feature that can be activated to quickly halt all clients for an important fix or change. This modular architecture allows pipelines to be added or modified without redesign of the underlying framework, making the tool ideal for a range of automated processing chains in big data management.

Input4MIPs: Boundary Condition and Forcing Datasets for CMIP6

Paul J. Durack (DOE/LLNL/PCMDI), pauldurack@llnl.gov; Karl E. Taylor (DOE/LLNL/PCMDI), taylor13@llnl.gov; Sasha Ames (DOE/LLNL), ames4@llnl.gov

Input4MIPs (input datasets for Model Intercomparison Projects) is an activity to make available via ESGF the boundary condition and forcing datasets needed for the sixth Coupled Model Intercomparison Project (CMIP6). Various datasets are needed for pre-industrial control, AMIP, and historical simulations, and additional datasets are needed for many of the 17 CMIP6-endorsed MIP experiments. Earlier versions of many of these datasets were used in CMIP5.

Unlike model data generated from CMIP6 experiments and standardized using CMOR, the formats of these contributed datasets vary and often test ESGF infrastructure limits. This presentation highlights some of the use cases encountered during collation and publishing of the Input4MIPs data and provides some insights into how the publishing step was augmented to deal with these highly variable data formats. Considering these use cases will be helpful as the ESGF system further evolves to address requirements of Obs4MIPs and other large international projects.

An Update on ESGF Needs for Obs4MIPs

Peter Gleckler (DOE/LLNL/PCMDI), gleckler1@llnl.gov

Obs4MIPs have advanced considerably since their inception nearly 5 years ago and are now formally recognized as a WCRP project, with oversight provided by the WCRP Data Advisory Council (WDAC) Task Team. There are currently seven ESGF nodes serving Obs4MIPs data, contributed by 15 institutions, with a current inventory of over 80 large-scale gridded observational products. Scientists recently proposed adding 100 new datasets to Obs4MIPs. The WDAC Task Team has identified several ways that ESGF could be enhanced to greatly facilitate the planned Obs4MIPS expansion, which would ensure the inclusion of a broader observational community. This presentation describes two enhancements that the Obs4MIPs WDAC Task Team wants to convey to the ESGF community.

Recent Climate4Impact Developments: Provenance in Processing and Connection to the CLIPC Portal

Maarten Plieger (ENES/KNMI), maarten.plieger@knmi.nl; Christian Pagé (ENES/IPSL), christian.page@cerfacs.fr; Sandro Fiore (ENES/CMCC), sandro.fiore@cmcc.it

The aim of Climate4Impact is to enhance the use of research data and support other climate portals. Climate4Impact was developed within the European projects IS-ENES, IS-ENES2, and CLIPC. Climate4Impact is connected to ESGF using certificate-based authentication, ESGF search, OpenID, OPeNDAP, and THREDDS catalogs. Climate4Impact offers web interfaces for searching, visualizing, analyzing, processing, and downloading datasets.

Climate4Impact exposes open standards like WMS, WCS, and WPS using open-source tools. Processing services include climate indicator calculations, country-based statistics, and polygon extraction by GeoJSON. Provenance integration is achieved using the W3C PROV standard for fully traceable provenance. The PROV document is stored in NetCDF files and can be visualized. The provenance module traces data usage statistics in a database, which is interesting for data providers.

Climate4Impact has a personal basket where users can upload their own data and do research with the provided tools. The basket supports formats like NetCDF, GeoJSON, and CSV. The basket has an access token mechanism to make data sharing and command line access to web services easier, enabling client-side scripting of the Climate4Impact portal and making it possible to connect third-party portals, like the European Union's FP7 CLIPC portal. The CLIPC portal uses web services from Climate4Impact and has an appealing front end built in openlayers3 and targeted to boundary workers.

This presentation details web services, provenance integration, and connection with the CLIPC portal.

Federated Data Usage Statistics in the Earth System Grid Federation

A. Nuzzo (ENES/CMCC), alessandra.nuzzo@cmcc.it; M. Mirto (ENES/CMCC), maria.mirto@cmcc.it; P. Nassisi (ENES/CMCC), paola.nassisi@cmcc.it; K. Berger (ENES/DKRZ), berger@dkrz.de; T. Rathmann (ENES/DKRZ), rathmann@dkrz.de; L. Cinquini (NASA/JPL), Luca.Cinquini@jpl.nasa.gov; S. Denvil (ENES/IPSL), sebastien.denvil@ipsl.jussieu.fr; S. Fiore (ENES/CMCC), sandro.fiore@cmcc.it; D. N. Williams (DOE/LLNL), williams13@llnl.gov; G. Aloisio (ENES/CMCC), giovanni.aloisio@unisalento.it

Monitoring ESGF is challenging. From an infrastructural standpoint, two components (Dashboard and Desktop) provide the proper environment for capturing (1) usage metrics and (2) system status information at the local (node) and global (institution and/or federation) level.

All the metrics collected by the ESGF monitoring infrastructure are stored in a system catalog that has been extended to support a large set of information about data usage statistics. More specifically, the Dashboard provides coarse- and fine-grained data usage statistics. Regarding the coarse-grained statistics, information like the data downloaded (GB/TB), number of downloads, number of distinct files and users, downloads by user and IdP, and client statistics (country and continent distribution) are provided. Fine-grained statistics are related to (1) cross projects, such as the number of downloads per project and host and by time and (2) specific projects such as CMIP5 and Obs4MIPs download data. In this case, the number

of downloads, number of successful downloads, downloaded data, and timeframe are provided with the possibility of knowing the top 10 datasets, experiments, and variables, grouped by (1) experiment/model and (2) experiment/model by time.

The fine-grained statistics are available for single and federated data nodes. A specific protocol allows ESGF to gather metrics from each data node [classified as leaf node (responsible for each site) and collector node (gathers the data from registered leaf nodes)]. To this end, several data marts have been created to allow fast access to this information. Project-specific views provide a deep insight about related statistics.

Large-Scale Data Analytics Workflow Support for Climate Change Experiments

S. Fiore (ENES/CMCC), sandro.fiore@cmcc.it; C. Doutriaux (DOE/LLNL), doutriaux1@llnl.gov; C. Palazzo (ENES/CMCC), cosimo.palazzo@cmcc.it; A. D'Anca (ENES/CMCC), cosimo.palazzo@cmcc.it; Z. Shaheen (DOE/LLNL), shaheen2@llnl.gov; D. Elia (ENES/CMCC), cosimo.palazzo@cmcc.it; J. Boutte (DOE/LLNL), boutte3@llnl.gov; V. Anantharaj (DOE/ORNL), anantharajvg@ornl.gov; D. N. Williams (DOE/LLNL), williams13@llnl.gov; G. Aloisio (ENES/CMCC), giovanni.aloisio@unisalento.it

Defining and implementing experiments with hundreds of data analytics operators can be a real challenge in many practical scientific use cases, such as multimodel analysis, climate indicators, and processing chains for operational environments. This is usually done via scripts (e.g., bash) on the client side and requires climate scientists to implement and replicate workflow-like control logic aspects (which also may be error-prone) in their scripts, along with the expected application-level part.

High-level solutions leveraging workflow-enabled big data analytics frameworks for e-science could help scientists in defining and implementing workflows related to their experiments by exploiting a more declarative, efficient, and powerful approach.

This talk presents key needs and challenges regarding big data analytics workflow management for e-science and provides insights about real use cases implemented in some European projects (e.g., BIGSEA, CLIPC, and INDIGO).

All the proposed use cases have been implemented exploiting the Ophidia big data analytics framework. The software stack includes an internal workflow management system, which coordinates, orchestrates, and optimizes the execution of multiple scientific data analytics and visualization tasks. Real-time workflow monitoring execution is also supported through a graphical UI. The provided data analytics workflow engine supports conditional sections, parallel loops, and massive statements for high-throughput experiments.

Specific emphasis will be devoted to a large-scale climate model intercomparison data analysis experiment (e.g., precipitation trend analysis) performed in the context of the H2020 INDIGO-DataCloud project. The use case exploits the INDIGO capabilities in terms of software framework deployed on cloud, UV-CDAT for data visualization, and Ophidia to run multimodel data analysis.

Day 3: Thursday, December 8, 2016

Coordinated Efforts with Community Software Projects

THREDDS Data Server: OPeNDAP and Other Tales from the Server-Side

Sean Arms (NSF/Unidata), sarms@ucar.edu

This talk is geared toward informing ESGF on the status of TDS data services. OPeNDAP is discussed in terms of both DAP2 and DAP4. Other TDS services, such as ncWMS and the NetCDF Subset Service, are discussed, highlighting how they may benefit ESGF users.

A Hybrid Provenance Capture Approach to Scientific Workflow Reproducibility and Performance Optimization

Todd Elsethagen (DOE/PNNL), todd.Elsethagen@pnnl.gov; Eric Stephan (DOE/PNNL), Eric.Stephan@pnnl.gov; Bibi Raju (DOE/PNNL), bibi.raju@pnnl.gov

As HPC infrastructures continue to grow in capability and complexity, so do the applications that they serve. HPC and distributed-area computing (DAC;

e.g., grid and cloud) users are looking increasingly toward workflow solutions to orchestrate their complex application coupling and pre- and post-processing needs. To gain insight and a more quantitative understanding of workflow performance, the Provenance Environment (ProvEn) architecture includes not only the capture of traditional provenance information, but also the capture and integration of system environment metrics helping to give context and explanation for a workflow's execution. This presentation describes how ESGF will use ProvEn to support reproducibility, data lineage, and performance optimization.

QA/QC at the DKRZ

Heinz-Dieter Hollweg (ENES/DKRZ), hollweg@dkrz.de

Between the states of climate datasets being compliant or rejectable there exists a grey range of being almost compliant to project rules. Thus, it is important to provide annotations describing compliance deviations in a brief but helpful way. Just as important is to enable modelers to perform pre-checks before submitting a large amount of data, whether by a locally installed QA tool or via WPS. The QA-DKRZ tool is presented with some technical aspects: installation and running as WPS, respectively, selection of subsets out of rather large data volumes, an interface to run external components or tools within QA-DKRZ (e.g., the CMOR checker), and the annotation model used to generate concise results. Also discussed are some experiences gained during CMIP5 and CORDEX regarding the "interaction" between submitters of data and QA results.

Web Processing Services and ESGF: The Birdhouse System

Stephan Kindermann (ENES/DKRZ), kindermann@dkrz.de; Carsten Ehbrecht (ENES/DKRZ), ehbrecht@dkrz.de; Nils Hempelmann (ENES/IPSL), nils.hempelmann@lsce.ipsl.fr

Provisioning web processing services near large ESGF sites support efficient future data analysis activities. To support the exposure of data analysis and data evaluation code in the form of OGC WPS, a modular set of easily installable and deployable components is being developed and bundled in the "birdhouse" framework (birdhouse.readthedocs.io/en/latest/). Individual processing services as well as generic infrastructural services (the "birds") are supported by a generic "birdhouse," providing a generic installation and deployment solution (e.g., supporting Docker-based hosting solutions).

The current system status is described and an overview is provided of existing services and services in active development, especially supporting the climate impact community (a short demo is given). There is special emphasis on the unique aspects and open issues that supporting efficient computing at sites providing large ESGF replica caches entails.

Synda (Synchro-Data)

Sébastien Denvil (ENES/IPSL), sebastien.denvil@ipsl.jussieu.fr; Raciazek Jérôme (ENES/IPSL), jripsl@ipsl.jussieu.fr; Guillaume Levavasseur (ENES/IPSL), glipsl@ipsl.jussieu.fr

Synda is a command line tool to search and download files from the ESGF archive. Since its inception in 2011, Synda essentially has been used for the replication use case when a large bulk of data needs to remain in sync between a local archive and the ESGF system. New features have been added to support a broader set of use cases including the first: easily grabbing a small or a large number of files or datasets from ESGF.

Synda can easily download files from the ESGF archive, 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 walks through Synda's main features and supported use case. We also expose how we plan to support an automatic replication workflow for CMIP6.

Current main features are listed as follows:

- 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, history stored in a database.

- GridFTP-enabled, fallback position to HTTP when needed hooks available for automatic publication upon datasets download completion.

Globus Update

Rick Wagner (University of Chicago, DOE/ANL), rick@globus.org

ESGF uses Globus for managed data transfer and sharing, both for replication between nodes and for users to transfer the data. Globus is software-as-a-service for research data management and provides high-speed and secure data transfer, data sharing directly from existing storage systems, and data publication. Developed and operated by the University of Chicago, Globus has become a preferred service for moving and sharing data between and among a wide variety of storage systems at research laboratories, campus computing resources, and national facilities across the United States. This presentation covers Globus integration with ESGF components and services, relevant recent updates to Globus, and potential methods for leveraging these new features by ESGF.

BASEJumper: Publishing HPSS Datasets via ESGF

Sam Fries (DOE/LLNL), fries2@llnl.gov; Sasha Ames (DOE/LLNL), ames4@llnl.gov; Alex Sim (DOE/LBNL), asim@lbl.gov

The capacity of hard disk space has not kept pace with the volume of output created by climate models. To store model output, High Performance Storage Systems (HPSSs) are required. These tape archives are notoriously slow, and getting permission to access them can be tricky and time consuming. To facilitate climate modelers and consumers of model output, the Analytics and Informatics Management Systems team at LLNL allows archived data to be requested and retrieved via ESGF. This system (BASEJumper, named after the Berkeley Archive Storage Encapsulation library) uses existing ESGF services, provides all the normal metadata required for a dataset, and uses ESGF's access control mechanisms to safeguard the data. It uses a two-stage design, allowing it to move around firewalls and many layers of security to prevent denial of service attacks on HPSS resources.

Day 3: Thursday, December 8, 2016

Live Demonstration Session

ESGF Ingestion Service Overview

Lukasz Lacinski (DOE/ANL), lukasz@uchicago.edu

This talk presents an overview of the Ingestion Service, which provides a remote interface to the ESGF publication command line tools. The interface is provided as a RESTful API, integrated with ESGF authentication and authorization. Additionally, the API can manage dataset transfers through the Globus Transfer Service and reorganize dataset files on the data node before publication. The API supports three different publication workflows independent of the location of dataset files being published: the local data node, a remote file system accessible through a Globus endpoint, and a remote storage with dataset accessible through HTTP.

Compute Working Team Server-Side Demonstration

C. Doutriaux (DOE/LLNL), doutriaux1@llnl.gov; Jason Boutte (DOE/LLNL), boutte3@llnl.gov

Over the last year, the ESGF CWT made significant progress. An API was established to communicate with ESGF CWT's WPS servers. Additionally, an end-user Python-based API was developed. In this talk, we demonstrate how the user API can be used to call various ESGF CWT WPS servers, all implementing a similar workflow in various fashions. Specifically, a multimodel ensemble average will be computed on the server(s)-side and employed by the end user.

Live Demo of Visualization and Processing Services in the Climate4Impact Portal

Maarten Plieger (ENES/KNMI), maarten.plieger@knmi.nl

The aim of Climate4Impact is to enhance the use of climate research data and support other climate impact portals. This live demonstration shows how Climate4Impact enables researchers to use climate data in their research.

Researchers are spending considerable amounts of ttime on data gathering, conversion, integration, and interpretation. Parts of this process have already

been done before and do not need to be repeated or re-invented. Climate4Impact facilitates this process and lifts the burdens from researchers, thus increasing the time available for real research.

To facilitate this process, Climate4Impact offers several web processing services and wizards, including an averager, subsetter, regridder, reformatter, combine tool, and tool for polygon subsetting.

The following topics are demonstrated:

- Login: Using OpenID to access Climate4Impact and ESGF data nodes.
- Discovery: Faceted search using the ESGF search API.
- Visualization: Visualize data using OGC web map services.
- Convert and subset: Transform data into other formats and geographical projections, using OCG web processing services.
- Import the obtained data in a GIS system (e.g., QGIS).
- Other processing services (e.g., averaging, polygon subsetting, and combining).

Climate4Impact is developed through the IS-ENES, IS-ENES2, and CLIPC projects, which receive funding from the European Union's Seventh Framework Programme for research, technological development, and demonstration.

Day 3: Thursday, December 8, 2016

Poster Session

ADAGUC Open-Source Visualization in Climate4Impact Using OGC Standards

Maarten Plieger (ENES/KNMI), maarten.plieger@knmi.nl; Ernst de Vreede (ENES/KNMI), ernst.de.vreede@knmi.nl

ADAGUC is an open-source geographical information system to visualize NetCDF, HDF5, and GeoJSON over the web. The software consists of a server-side C++ application and a client-side JavaScript application. The software provides several features to access and visualize data over the web; it uses the OGC WMS and WCS standards for data dissemination. Web clients like Google Maps, OpenLayers, and Leaflet are supported and can directly use the exposed web services. ADAGUC is used in projects like Climate4Impact to visualize datasets stored in ESGF.

ADAGUC can visualize remotely published NetCDF files by adding the OPeNDAP resource as a parameter to the web service request. This enables direct visualization of any OPeNDAP-enabled resource over the web. Checking the variable standard name and units does graphic styling of data. OGC Web Coverage Services (WCSs) are available and can be used for data reprojection, subsetting, and conversion to other formats. Access to OPeNDAP services is done efficiently; multiple requests are aggregated into one and only the domain of interest is requested. This allows easy, quick, and interactive visualization of OPeNDAP-enabled datasets.

ADAGUC has a number of data converters and data postprocessors to support various data conventions. Supported file formats are "true color NetCDF" for satellite imagery, structured grids, curvilinear grids, satellite swaths, point observations, point time series, and polygons stored in GeoJSON. Datasets consisting of several NetCDF files can be aggregated into a single dataset and are offered over WMS, WCS, and OPeNDAP. ADAGUC can be used as a component for WPS to subset data and convert GeoJSON to grids.

Results and lessons learned are presented.

Community Data Management System

Denis Nadeau (DOE/LLNL), nadeau1@llnl.gov; Charles Doutriaux (DOE/LLNL), doutriaux1@llnl.gov; Dean N. Williams (DOE/LLNL), williams13@llnl.gov

The Analytics and Informatics Management Systems (AIMS) team will completely redesign and transform the Climate Data Management System Version 2 (CDMS2) into the Community Data Management System (CDMS). Designed in the mid to late 1990s, CDMS's original intent was to automatically locate and extract metadata (e.g., variables, dimensions, and grids) from collections of simulation runs and analysis files. Since then, it has grown to include multiple regridders, time components, masked arrays, and more. However, with the rapid changes in technology, it is time for an upgrade to broaden its scope and design to include

newer "community" data file formats (such as HDF5 and IDX), the latest "community" Numerical Python packages (such as NumPy 3.0 and Numba), and the latest "community" of regridders that combine the manipulation of simulation, observation, reanalysis, and point datasets. CDMS aims to incorporate 21st century technologies and integrate additional geoscience domains. In addition to conforming to the latest community standards and protocols, it will include the new CDAT ingest package.

Community Diagnostics Package

Zeshawn Shaheen (DOE/LLNL), shaheen2@llnl.gov; Charles Doutriaux (DOE/LLNL), doutriaux1@llnl.gov; Samuel Fries (DOE/LLNL), fries2@llnl.gov

Scientific code is often created for a single, narrowly focused goal. Such code is inflexible and over time may cause progress on a project to reach an impasse. The AIMS group and LLNL are developing the Community Diagnostics Package (CDP), a framework for creating new climate diagnostic packages in a generalized manner. Designed in an object-oriented method, CDP allows for a modular implementation of the components required for running diagnostics. CDP's design consists of modules to handle the user-defined parameters, metrics, provenance, file input/output, and output of results and algorithms for calculating the diagnostics.

Earth System Model Development and Analysis Using FRE-Curator and Live Access Servers: On-Demand Analysis of Climate Model Output with Data Provenance

Aparna Radhakrishnan (NOAA/GFDL), aparna.radhakrishnan@noaa.gov; V. Balaji (NOAA/GFDL), balaji@princeton.edu; Roland Schweitzer (NOAA/GFDL), roland.schweitzer@noaa.gov; Serguei Nikonov (NOAA/GFDL), serguei.nikonov@noaa.gov; Kevin O'Brien, (NOAA/PMEL), kevin.m.o'brien@noaa.gov; Hans Vahlenkamp (NOAA/PMEL), hans.vahlenkamp@noaa.gov

There are distinct phases in the development cycle of an Earth system model. During the model development phase, scientists make changes to code and parameters and require rapid access to results for evaluation. During the production phase, scientists may make an ensemble of runs with different settings and produce large quantities of output that must be further analyzed and quality controlled for scientific papers and submission to international projects such as CMIP. During this phase, provenance is a key concern: being able to track back from outputs to inputs. We discuss one of the paths taken at GFDL in delivering tools across this life cycle, offering on-demand analysis of data by integrating the use of GFDL's in-house FRE-Curator, Unidata's THREDDS, and NOAA PMEL's LAS.

Experience over this life cycle suggests that a major difficulty in developing analysis capabilities is only partially the scientific content. It is often devoted to answering the questions "Where is the data?" and "How do I get to it?" FRE-Curator is a database-centric paradigm used at NOAA GFDL to ingest information about the model runs into an RDBMS (Curator database). The components of FRE-Curator are integrated into a Flexible Runtime Environment workflow and can be invoked during climate model simulation. The front end to FRE-Curator, known as the Model Development Database Interface (MDBI), provides an in-house, web-based access to GFDL experiments: metadata, analysis output, and more. To provide on-demand visualization, MDBI uses LAS, which is a highly configurable web server designed to provide flexible access to georeferenced scientific data that makes use of OPeNDAP. Model output saved in GFDL's tape archive, the size of the database and experiments, and continuous model development initiatives with more dynamic configurations add complexity and challenges in providing an on-demand visualization experience to our GFDL users.

Toward a High-Performance Data Analysis Platform for Impact Analysis

Wim Som De Cerff (ENES/KNMI), wim.som.de.cerff@knmi.nl; Sandro Fiore (ENES/CMCC), sandro.fiore@cmcc.it; Maarten Plieger (ENES/KNMI), maarten.plieger@knmi.nl

The aim of Climate4Impact is to enhance the use of climate research data and 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 of use case owners.

In the data analytics landscape, Ophidia provides a big data framework for e-science, focusing on the analysis of large-scale n-dimensional datasets. Ophidia provides

data operators to manipulate data in the form of data cubes and array-based primitives to perform data analysis on large scientific data arrays (e.g., statistical analysis, predicate evaluation, subsetting, reduction, and compression).

KNMI and CMCC are working together toward a high-performance data analysis platform for impact analysis by integrating and properly extending and adapting Climate4Impact and Ophidia.

A key point to be addressed is the interoperability with ESGF, in terms of the security and access interface, which means working closely with the ESGF CWT and the ESGF IdEA working teams.

To support users for their analytics and scientific operations on large amounts of data, the Climate4Impact portal will interact with Ophidia for the back-end processing capabilities. The Ophidia WPS interface and Climate4Impact services will allow easy front-end controlling, visualizing, and tracking of remote Ophidia task submissions. Moreover, the Ophidia workflow engine in addition to the visualization capabilities of ADAGUC, will lead to near-real-time output of production and visualization of complex experiments.

Climate Model Output Rewriter Version 3.2 for CMIP6

Denis Nadeau (DOE/LLNL), nadeau1@llnl.gov; Karl Taylor (DOE/LLNL), taylor13@llnl.gov; Charles Doutriaux (DOE/LLNL), doutriaux1@llnl.gov; Dean N. Williams (DOE/LLNL), williams13@llnl.gov

Version 3.2 of the Climate Model Output Rewriter (CMOR) has been released to handle state-of-the-art MIPs. The Working Group Coupled Model Infrastructure Panel (WIP) has created an exhaustive Data Request database for CMIP6 that is used by CMOR to rewrite model output. The files created by CMOR also follow the CF-1 metadata conventions to promote the processing and sharing of CMIP6 data. The latest version of CMOR 3.2 incorporates a "Control Vocabulary" API to line up with continuously growing CMIP6 requirements from the WIP. This API also has been incorporated into the ESGF publisher to validate every published file for the CMIP6 project. Delineating new input table structure confines CMOR to strict CMIP6 requirements, which empower each model to maintain value delivery. CMOR is a robust program and can work with different types of grids, different projections, LIDAR tracks, or ship transects. Its high flexibility allows customization of global attributes to accommodate growth in capability needed by different MIPs, such as Obs4MIPs, Input4MIPs, or CREATEs.

QoS-Based Dynamic and Elastic Scenarios in the Cloud for Data Analytics in the BIGSEA Project

Donatello Elia (ENES/CMCC), donatello.elia@cmcc.it; Sandro Fiore (ENES/CMCC), sandro.fiore@cmcc.it; Giovanni Aloisio (ENES/CMCC), giovanni.aloisio@unisalento.it

EUrope-BRAzil Collaboration on BIG Data Scientific REsearch through Cloud-Centric Applications (EUBra-BIGSEA) is a project funded under the third coordinated call for Europe and Brazil. It targets the development of cloud services for big data applications to ensure quality of service (QoS), security, and data privacy. The integrated and fast big data ecosystem is the central component; it addresses the data management aspects of the EUBra-BIGSEA platform. Its key elements are the integration of different classes of big data technologies, such as the Ophidia framework or Spark; the dynamicity and elasticity of the environment based on QoS policies; and a secured-by-design architecture. The ecosystem joins these aspects in a cloud environment to tackle massive data processing scenarios like the ones proposed in the BIGSEA project. In particular, this poster shows how Ophidia has been integrated into a smart city context to deal with weather forecasting data in cloud QoS-based elastic and dynamic scenarios. The generality of the approach makes its adoption straightforward in the ESGF-based context, with special regard to the computing and analysis part, where different user needs and workloads could benefit from these new developments.

Appendix C. ESGF's Current Data Holdings

- Coupled Model Intercomparison Project Phase 6 (CMIP6) (coming soon)
- Coupled Model Intercomparison Project Phase 5 (CMIP5)
- Coupled Model Intercomparison Project Phase 3 (CMIP3)
- Empirical-Statistical Downscaling (ESD)
- Coordinated Regional Climate Downscaling Experiment (CORDEX)

Fig. 13. Major federated ESGF worldwide sites.

- Accelerated Climate Modeling for Energy (ACME)
- Parallel Ocean Program (POP)
- North American Regional Climate Change Assessment Program (NARCCAP)
- Carbon Land Model Intercomparison Project (C-LAMP)
- Atmospheric InfraRed Sounder (AIRS)
- Microwave Limb Sounder (MLS)
- Cloudsat
- Observations for Model Intercomparison Projects (Obs4MIPs)
- Analysis for Model Intercomparison Projects (ana4MIPs)
- Cloud Feedback MIP (CFMIP)
- Input4MIPs
- European Space Agency Climate Change Initiative (ESA CCI) Earth Observation data
- Seasonal-to-decadal climate Prediction for the improvement of European Climate Services (SPECS)
- Inter-Sectoral Impact Model Intercomparison Project (ISI MIP)
- Computational Modeling Algorithms and Cyberinfrastructure (CMAC)
- Vertical Structure and Physical Processes of Weather and Climate (GASS and YoTC)
- Collaborative REAnalysis Technical Environment - Intercomparison Project (CREATE IP)
- NASA NEX Global Daily Downscaled Climate Projections (NEX GDDP)
- NASA NEX Downscaled Climate Projections (NEX-DCP30)
- High Impact Weather Prediction Project (HWPP)
- Coupled NEMS
- Climate Model Development Task Force (CMDTF)

Appendix D. Conference Participants and Report Contributors

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)/Jet propulsion laboratory (JPL)
- **Daniel Duffy** – ESGF Executive Committee, NASA
- **Sébastien Denvil** – ESGF Executive Committee, Institut Pierre-Simon Laplace (IPSL)
- **Robert Ferraro** – ESGF Executive Committee, NASA
- **Claire Trenham** – ESGF Executive Committee, National Computational Infrastructure (NCI) Australia

Fig. 14. Participants in the 2016 ESGF F2F Conference.

ESGF Program Managers in Attendance

- **Justin Hnilo** – Chair of the ESGF Steering Committee, DOE Office of Biological and Environmental Research (BER)
- **Sylvie Joussaume** – ESGF Steering Committee, ENES
- **Tsengdar Lee** – ESGF Steering Committee, NASA
- **Ben Evans** – ESGF Steering Committee, NCI

Attendees and Contributors		
Name		**Affiliation**
1.	Aloisio, Giovanni	European Network for Earth System Modeling (ENES), Euro-Mediterranean Centre on Climate Change (CMCC) Foundation, and University of Salento
2.	Ames, Alexander "Sasha"	U.S. Department of Energy (DOE) Lawrence Livermore National Laboratory (LLNL)
3.	Anantharaj, Valentine	DOE Oak Ridge National Laboratory (ORNL)
4.	Arms, Sean	University Corporation for Atmospheric Research (UCAR) Community Programs (UCP) and Unidata
5.	Ben Nasser, Atef	European Network for Earth System Modelling (ENES) and Pierre Simon Laplace Institute (IPSL)
6.	Berger, Katharina	ENES and German Climate Computing Center (DKRZ)
7.	Buurman, Merret	ENES and DKRZ
8.	Carriere, Laura	National Aeronautics and Space Administration (NASA) Goddard Space Flight Center (GSFC)
9.	Chaudhary, Aashish	Kitware, Inc.
10.	Chen, Kangjun	State Key Laboratory of Numerical Modeling for Atmospheric Sciences and Geophysical Fluid Dynamic

Attendees and Contributors		
Name		**Affiliation**
11.	Christensen, Cameron	Scientific Computing and Imaging (SCI) Institute, University of Utah
12.	Cinquini, Luca	NASA Jet Propulsion Laboratory (JPL)
13.	Dart, Eli	DOE Energy Sciences Network (ESnet)
14.	Denvil, Sébastien	French National Centre for Scientific Research (CNRS) – IPSL
15.	Devarakonda, Ranjeet	DOE ORNL
16.	Doutriaux, Charles	DOE LLNL
17.	Duffy, Daniel	NASA GSFC
18.	Durack, Paul	DOE LLNL Program for Climate Model Diagnosis and Intercomparison (PCMDI)
19.	Dwarakanath, Prashanth	National Supercomputer Centre
20.	Evans, Ben	Australia's National Computational Infrastructure (NCI) and Australian National University (ANU)
21.	Ferraro, Robert	NASA JPL
22.	Fiore, Sandro	ENES, CMCC Foundation, and University of Salento
23.	Fries, Samuel	DOE LLNL
24.	Gauvin St-Denis, Blaise	Ouranos
25.	Geernaert, Gary	DOE Office of Biological and Environmental Research (BER) Climate and Environmental Sciences Division (CESD) Director
26.	Gleckler, Peter	DOE LLNL PCMDI
27.	Hansen, Rose	DOE LLNL
28.	Harr, Cameron	DOE LLNL Livermore Computing (LC)
29.	Hill, William	DOE LLNL
30.	Hnilo, Jay	DOE BER CESD Program Manager
31.	Hollweg, Heinz-Dieter	ENES and DKRZ
32.	Huard, David	Ouranos
33.	Inoue, Takahiro	Research Organization for Information Science and Technology (RIST), Japan
34.	Jefferson, Angela	DOE LLNL
35.	Joseph, Renu	DOE BER CESD
36.	Joussaume, Sylvie	ENES, CNRS, ENES Coordinator
37.	Kindermann, Stephan	ENES and DKRZ
38.	Kolax, Michael	Swedish Meteorological and Hydrological Institute (SMHI)
39.	Lacinski, Lukasz	University of Chicago
40.	Landry, Tom	Computer Research Institute of Montréal (CRIM, Canada)
41.	Lautenschlager, Michael	ENES and DKRZ

Attendees and Contributors		
Name		**Affiliation**
42.	Lee, Tsengdar	NASA Headquarters
43.	Levavasseur, Guillaume	ENES and IPSL
44.	Lu, Kai	ENES, National Supercomputer Centre (NSC) at Linkoping University, Sweden
45.	Maxwell, Thomas	NASA GSFC
46.	McFarland, Sally	DOE BER CESD
47.	Nadeau, Denis	DOE LLNL
48.	Nienhouse, Eric	National Science Foundation (NSF) National Center for Atmospheric Research (NCAR)
49.	Nikonov, Serguei	National Oceanic and Atmospheric Administration (NOAA) Geophysical Fluid Dynamics Laboratory (GFDL) and Princeton University
50.	Nuzzo, Alessandra	ENES, CMCC Foundation, and University of Salento
51.	Ogochi, Koji	Japan Agency for Marine-Earth Science and Technology (JAMSTEC)
52.	Pagé, Christian	Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique (CERFACS)
53.	Peterschmitt, Jean-Yves	IPSL Laboratory for Sciences of Climate and Environment (LSCE)
54.	Plieger, Maarten	ENES and Royal Netherlands Meteorological Institute (KNMI)
55.	Pobre, Alakom-Zed	NASA GSFC National Center for Climate Simulation (NCCS)
56.	Potter, Gerald	NASA GSFC
57.	Prakash, Giri	DOE ORNL
58.	Prichard, Matt	Science and Technology Council (STFC), United Kingdom
59.	Qin, Peihua	Institute of Atmospheric Physics, Chinese Academy of Sciences
60.	Rathmann, Torsten	ENES and DKRZ
61.	Rumney, George	NASA GSFC
62.	Santhana Vannan, Suresh Kumar	DOE ORNL
63.	Stephan, Eric	DOE Pacific Northwest National Laboratory (PNNL)
64.	Stephens, Ag	ENES, STFC Centre for Environmental Data Analysis (CEDA)
65.	Stockhause, Martina	ENES and DKRZ
66.	Story, Matthew	DOE LLNL
67.	Tamkin, Glenn	NASA
68.	Taylor, Karl	DOE LLNL PCMDI
69.	Trenham, Claire	NCI and ANU
70.	Tucker, William	STFC
71.	Vahlenkamp, Hans	NOAA GFDL
72.	Wagner, Rick	DOE and University of Chicago's Globus
73.	Williams, Dean	DOE LLNL
74.	Xu, Hao	Tsinghua University

Online Attendees and Contributors		
Name		**Affiliation**
1.	Berkley, Mike	Canadian Centre for Climate Modeling for Analysis (CCCma)
2.	Bretonniere, Pierre-Antoine	Barcelona Supercomputing Center
3.	Calo, Giuseppe	CMCC
4.	Chunpir, Hashim	ENES DKRZ
5.	Cofino, Antonio	ENES and University of Cantabria, Spain
6.	Harris, Matt	DOE LLNL
7.	Hertz, Judy	NASA GSFC NCCS
8.	McCoy, Renata	DOE LLNL
9.	McEnerney, Jim	DOE LLNL
10.	Osvaldo, Marra	ENES and CMCC
11.	Petrie, Ruth	ENES and STFC, U.K.
12.	Shaheen, Zeshawn	DOE LLNL
13.	Wang, Dali	DOE ORNL
14.	Wu, Qizhong	Beijing Normal University, China

Appendix E. Awards

Every year, the climate software engineering community gathers to determine who has performed exceptional or outstanding work developing community tools for the acceleration of climate science in the Earth System Grid Federation (ESGF) data science domain. This year, the ESGF Executive Committee determined the winners of the ESGF Achievement Awards. 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 for 2016 include:

- **Katharina Berger** [European Network for Earth System Modelling (ENES) and the German Climate Computing Centre (DKRZ)] won an award for her development of the Climate Model Intercomparison Project version 6 (CMIP6) early citation services and CMIP6 Persistent Identifier (PID) services, which will be used to reference data collections prior to long-term archiving in the Intergovernmental Panel on Climate Change (IPCC) Data Distribution Centre (DDC), hosted at the German Climate Computing Centre (DKRZ). Her involvement in developing the Data Citation concept, with emphasis on the technical requirements and the dependencies on other teams' developments, is instrumental to the success of CMIP6. Additionally, Katharina leads ESGF's Quality Control Working Team (QCWT) and plays a critical role in the installation, testing, and operation of the ESGF worldwide federation of nodes.
- **Philip Kershaw** [ENES and Center for Environmental Data Analysis (CEDA), U.K.] won an award for his years of ESGF security leadership and coordination across the federation. This work includes leading the Identity Entitlement Access Management (IdEA) Working Team, which this year focused on establishing a roadmap for integration of OAuth 2.0 for user delegation. This work demonstrates prototype security connections needed for tight integration for future server-side computing and visualization services carried out by ESGF's Compute Working Team (CWT).
- **Lukasz Lacinski** [U.S. Department of Energy's Argonne National Laboratory (ANL)] won an award for working with multiple ESGF teams both to improve data transfer performance and to simplify transferring of datasets to and from an ESGF data node. Working with the CoG User Interface Working Team and the Publication Working Team, he has led the integration of the ESGF services with

Fig. 15. ESGF Achievement Award recipients for 2016. From left to right: Lukasz Lacinski, Katharina Berger, Sébastien Denvil, Angela Jefferson, and Dean N. Williams (holding awards for Philip Kershaw and Jérôme Raciazek).

Globus Online. More recently, Lukasz has been working at replacing OpenID authentication within CoG with OAuth2, and later OpendID-Connect as part of a much-needed upgrade of the ESGF security infrastructure.

- **Jérôme Raciazek** [ENES and the Institut Pierre-Simon Laplace (IPSL)] won an award for his ongoing commitment to and support for Synda software. Since 2010, he has closely followed software changes relevant to the ESGF search service and has supported enhancements to the authentication mechanism and to the specific project requirement for optimizing and enhancing the user experience. In 2016, Jérômeworked with the Replication and Versioning Working Team to ensure that Synda can take advantage of Globus capabilities when available. This sustained effort allows ESGF to offer a robust tool for efficient and agile replication (download and replica publication) for any current and future ESGF-hosted projects.
- **Angela Jefferson** [DOE's Lawrence Livermore National Laboratory (LLNL)] won an award for serving as the primary administrator for the ESGF F2F conferences. Working with the international community, she has helped to design, manage, develop, and arrange the six international ESGF conferences. Angela's work on conferences and administration has had a great impact on ESGF and many other LLNL and community climate projects.
- **Sébastien Denvil** [ENES IPSL] won an award for his years of strong support of ESGF. He has been critical in developing tight and proficient collaboration between climate efforts in Europe and other countries. Recently, he has started leading the CMIP Data Node Operations Team (CDNOT), which will guarantee a high level of service for delivering CMIP6 data to the scientific community. Additionally, Sébastienis either leading or highly involved with the development of several ESGF services within the European Community, such as Synda and ES-DOC.

Appendix F. Acknowledgments

The 2016 Earth System Grid Federation (ESGF) Face-to-Face (F2F) Conference organizers wish to thank national and international agencies for providing travel funding for attendees to join the conference in person, the U.S. Department of Energy's (DOE) 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 conference organization, 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 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., Coupled Model Intercomparison Project (CMIP), Accelerated Climate Modeling for Energy (ACME), Coordinated Regional Climate Downscaling Experiment (CORDEX), Model Intercomparison Projects (MIPs) in general, 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; German Climate Computing Centre; 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 (lead institution); and the University of Utah. Many other organizations and institutions have contributed to the efforts of ESGF, and we apologize to any whose names we have unintentionally omitted.

DOE, 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, and the Australian National Computational Infrastructure provide major funding for the ESGF community hardware, software, and network infrastructure efforts.

Additional thanks go to Betty Mansfield, Kris Christen, Holly Haun, Stacey McCray, Marissa Mills, and Judy Wyrick, of Oak Ridge National Laboratory's Biological and Environmental Research Information System for editing and preparing this report for publication.

Appendix G. Acronyms

Acronym	Definition
2D, 3D	Two dimensional, three dimensional
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/).
ADAGUC	Atmospheric Data Access for the Geospatial User Community—Open-source GIS system.
AIMS	Analytics and Informatics Management Systems—Team led by LLNL.
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 (airs.jpl.nasa.gov).
AMIP	Atmospheric Model Intercomparison Project—An experimental global atmospheric circulation model that provides a community-based infrastructure for climate model diagnosis, validation, intercomparison, and data access.
Ana4MIPs	Analysis for Model Intercomparison Projects
ANL	Argonne National Laboratory—Science and engineering research national laboratory near Lemont, Illinois, operated by the University of Chicago for DOE (anl.gov).
ANU	Australian National University (anu.edu.au)
API	Application Programming Interface(en.wikipedia.org/wiki/Application_programming_interface/).
AR	Assessment Report
ASCR	DOE Office of Advanced Scientific Computing Research—Supports the discovery, development, and deployment of computational and networking capabilities for the analysis, modeling, simulation, and prediction of complex phenomena important to DOE's advancement of science (science.energy.gov/ascr/).
BADC	British Atmospheric Data Centre—The Natural Environment Research Council's designated data center for atmospheric sciences (badc.nerc.ac.uk/home/index.html).
BASEJumper	Named after the Berkeley Archive Storage Encapsulation library. System that allows archived data to be requested and retrieved via ESGF.
BASH	Bourne-Again Shell—Command language and Unix shell.
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/).
C4I	Climate4Impact—Web portal that enables visualization of climate model datasets targeted to the climate change impact assessment and adaptation communities (climate4impact.eu/impactportal/general/).

Acronym	Definition
CA	Certificate Authority
CAFE	Framework for Collaborative Analysis of Distributed Environmental Data—A Java-based distributed data management and analysis framework.
CCI	ESA Climate Change Initiative
CDAS	Climate Data Analytics Services—NASA framework that brings together the tools, data storage, and HPC required for timely analysis of large-scale climate datasets where they reside.
CDAT	Community Data Analysis Tools
CDMS	Climate Data Management System—Object-oriented data management system specialized for organizing multidimensional gridded data used in climate analyses for data observation and simulation; to be redesigned and transformed into the Community Data Management System.
CDNOT	Coupled Model Intercomparison Project Data Node Operations Team
CDO	Climate Data Operators software is a collection of many operators for standard processing of climate and forecast model data.
CDP	Community Diagnostics Package
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 (ceda.ac.uk).
CESD	DOE BER Climate and Environmental Sciences Division
CF	Climate Forecast conventions and metadata (cfconventions.org).
CLIPC	Climate Information Platform for Copernicus
CMCC	Centro Euro-Mediterraneo sui Cambiamenti Climatici (Euro-Mediterranean Center on Climate Change)—This Italian scientific organization enhances collaboration and integration among climate science disciplines (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).
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).

Acronym	Definition
CNRS	Centre Nationale de la Recherche Scientifique (French National Centre for Scientific Research)—Largest fundamental science agency in Europe (cnrs.fr).
CoG	Collaborative software enabling projects to create dedicated workspaces, network with other projects, and share and consolidate information within those networks (earthsystemcog.org/projects/cog/).
CORDEX	Coordinated Regional Climate Downscaling Experiment—Provides global coordination of regional climate downscaling for improved regional climate change adaptation and impact assessment (cordex.org).
CP	Certificate Policy
CPS	Certificate Practices Statement
CRCM	Canadian Regional Climate Model
CREATE	Collaborative REAnalysis Technical Environment—NASA project that centralizes numerous global reanalysis datasets into a single advanced data analytics platform.
CREATE-IP	Collaborative REAnalysis Technical Environment Intercomparison Project—Data collection, standardization, and ESGF distribution component of CREATE (earthsystemcog.org/projects/create-ip/).
CRIM	Computer Research Institute of Montréal (Canada)
CSW	Catalogue Service for the Web
CV	Controlled Vocabulary
CWT	ESGF Compute Working Team
DAC	Distributed Area Computing
DDC	Data Distribution Centre—The IPCC's DDC provides climate, socioeconomic, and environmental data, both from the past and in scenarios projected into the future, for use in climate impact assessments (ipcc-data.org).
DKRZ	Deutsches Klimarechenzentrum (German Climate Computing Centre)—Provides HPC platforms and sophisticated, high-capacity data management and services for climate science (dkrz.de).
DOE	U.S. Department of Energy—Government agency chiefly responsible for implementing energy policy (energy.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).
DMZ	Demilitarized zone, or perimeter network, where an organization's external-facing services are exposed to untrusted networks.

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 (dream.llnl.gov).
DRS	Data Reference Syntax—Naming system used within files, directories, metadata, and URLs to identify datasets 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
EDSL	Embedded Domain Specific Language
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 (verc.enes.org).
ES-DOC	Earth System Documentation
ESA	European Space Agency—International organization coordinating the development of Europe's space capability, with programs to develop satellite-based technologies and services and to learn more about Earth, its immediate space environment, the solar system, and universe (esa.int/ESA/).
ESGF	Earth System Grid Federation—Led by LLNL, a worldwide federation of climate and computer scientists deploying a distributed multi-PB archive for climate science (esgf.llnl.gov).
ESMF	Earth System Modeling Framework
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 (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 (esrl.noaa.gov).
EuBra-BIGSEA	Europe-Brazil Collaboration of Big Data Scientific Research Through Cloud-Centric Applications (eubra-bigsea.eu).
F2F	Face to Face
GB	Gigabyte
GeoMIP	Geoengineering Model Intercomparison Project (climate.envsci.rutgers.edu/GeoMIP/)
GFDL	Geophysical Fluid Dynamics Laboratory—Scientists at NOAA's GFDL develop and use mathematical models and computer simulations to improve our understanding and prediction of the behavior of the atmosphere, the oceans, and climate (gfdl.noaa.gov).

Acronym	Definition
GIS	Geographical Information System
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—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 the Earth, sun, solar system, and universe (nasa.gov/centers/goddard/home/).
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 (hdfgroup.org/HDF5/).
HPC	High-Performance Computing
HPSS	High-Performance Storage System—Modern, flexible, performance-oriented mass storage system (hpss-collaboration.org).
HTTP	Hypertext Transfer Protocol
ICCLIM	Indice Calculation CLIMate—Python library that calculates various climate indices.
ICNWG	International Climate Network Working Group—Formed under the ESGF to help set up and optimize network infrastructure for climate data sites around the world (icnwg.llnl.gov).
IdEA	ESGF Identity Entitlement Access Management Working Team
IdP	identity provider
IDX	A type of multiresolution file format
INDIGO	Integrating Distributed data Infrastructures for Global ExplOitation—European Horizon 2020 project to develop a cloud platform for big data and computing ESGF use involves extending and exploiting Ophidia (https://www.indigo-datacloud.eu/the_project).
Input4MIPs	Input Datasets for Model Intercomparison Projects—A database used for preparing forcing datasets and boundary conditions for CMIP6 (pcmdi.llnl.gov/projects/input4mips/).
IPCC	Intergovernmental Panel on Climate Change—Scientific body of the United Nations that periodically issues assessment reports on climate change (ipcc.ch).
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 (ipsl.fr/en/).
IS-ENES	Infrastructure for the European Network for Earth System Modeling—Distributed e-infrastructure of ENES models, model data, and metadata (is.enes.org). IS-ENES2 refers to phase two of this project.
JPL	Jet Propulsion Laboratory—A federally funded research and development laboratory and NASA field center in Pasadena, California (jpl.nasa.gov).

Acronym	Definition
JSON	JavaScript Object Notation—An open, text-based, and standardized data exchange format better suited for Ajax-style web applications (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 (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. Berkeley Lab uses interdisciplinary teams and advanced new tools for scientific discovery (lbl.gov).
LiU	Linköping University's National Supercomputer Centre in Sweden—Houses an ESGF data node, test node, ESGF code sprint, user support, and Bi and Frost clusters (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 (llnl.gov).
LLNL/AIMS	Analytics and Informatics Management Systems—Program at LLNL enabling data discovery and knowledge integration across the scientific climate community (aims.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 (gisclimat.fr/en/laboratory/lsce-climate-and-environment-sciences-laboratory/).
MDBI	Model Development Database Interface
MIP	Model Intercomparison Project
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).
MPI	Message Passing Interface—Standardized, portable message-passing system designed to function on a variety of parallel computers.
NASA	National Aeronautics and Space Administration—U.S. government agency responsible for the civilian space program as well as aeronautics and aerospace research (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).
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 (nccs.nasa.gov).

Acronym	Definition
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 (ncdc.noaa.gov).
NCI	National Computational Infrastructure—Australia's high-performance supercomputer, cloud, and data repository (nci.org.au).
NCO	NetCDF Operators—A suite of programs using NetCDF files (nco.sourceforge.net).
ncWMS	Web Map Service for geospatial data stored in CF-compliant NetCDF files (reading-escience-centre.gitbooks.io/ncwms-user-guide/content/).
NERDIP	National Environmental Research Data Interoperability Platform—NCI's *in situ* petascale computational environment enabling both HPC 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 (nersc.gov).
NetCDF	Network Common Data Form—Machine-independent, self-describing binary data format (unidata.ucar.edu/software/netcdf/).
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 (noaa.gov).
NSC	National Supercomputer Centre—Provides leading-edge, HPC resources and support to users throughout Sweden. NSC is an independent center within Linköping University funded by the Swedish Research Council via the Swedish National Infrastructure for Computing (nsc.liu.se).
NSF	National Science Foundation—Federal agency that supports fundamental research and education in all the nonmedical fields of science and engineering (nsf.gov).
OAuth	Open standard for authorization (oauth.net)
Obs4MIPs	Observations for Model Intercomparisons—Database used by the CMIP modeling community for comparing satellite observations with climate model predictions (earthsystemcog.org/projects/obs4mips/).
OCGIS	Open Climate GIS—Set of geoprocessing and calculation tools for CF-compliant climate datasets.
OGC	Open Geospatial Consortium—International nonprofit organization that develops quality open standards to improve sharing of the world's geospatial data (opengeospatial.org).
OLAP	Online Analytical Processing
OMIP	Ocean Model Intercomparison Project
OPeNDAP	Open-Source Project for a Network Data Access Protocol—Architecture for data transport including standards for encapsulating structured data and describing data attributes (opendap.org).

Acronym	Definition
OpenID	An open standard and decentralized authentication protocol. (CoG uses an ESGF OpenID as its authentication mechanism.)
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 (ornl.gov).
ORP	OpenID Relying Party
OS	Operating System
OWL	Web Ontology Language
PAVICS	Power Analytics and Visualization for Climate Science—A platform designed by Ouranos for the analysis and visualization of climate science data (ouranos.ca/publication-scientifique/PAVICS2016_ENG.pdf)
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).
perfSONAR	Performance Focused Service Oriented Network monitoring Architecture—Open-source software for running network tests (perfsonar.net/).
PID	Persistent IDentifier—A long-lasting reference to a digital object, a single file, or set of files (en.wikipedia.org/wiki/Persistent_identifier).
PKI	Public Key Infrastructure
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 (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).
PMP	PCMDI Metrics Package
PNNL	Pacific Northwest National Laboratory—DOE national laboratory in Richland, Washington, where multidisciplinary scientific teams address problems in four areas: science, energy, the Earth, and national security (pnnl.gov).
POSIX®	Portable Operating System Interface—Family of standards specified by the IEEE Computer Society for maintaining compatibility between OSs.
PROV	World Wide Web Consortium's provenance representation standard
ProvEn	Provenance Environment (Elsethagen, T. O., et al. 2016. ieeexplore.ieee.org/document/7747819/)
PyWPS	Python Web Processing Service

Acronym	Definition
QA	Quality Assurance
QC	Quality Control
QCWT	ESGF Quality Control Working Team
QGIS	Quantum GIS—Open-source geographical information system.
QoS	Quality of Service—Elastic and dynamic Cloud environment for data analytics.
RabbitMQ	Open source message broker software (www.rabbitmq.com/).
RDBMS	Relational Database Management System
React	JavaScript library, which was created by a collaboration of Facebook and Instagram. It is used to build a framework of reusable components for user interfaces.
REST	Representational State Transfer—Computing architectural style consisting of a coordinated set of constraints applied to components, connectors, and data elements within a distributed hypermedia system such as the World Wide Web.
RFMIP	Radiative Forcing Model Intercomparison Project
RVWT	ESGF Replication and Versioning Working Team
SAML	Radiative Forcing Model Intercomparison Project (rfmip.leeds.ac.uk/)
SDLC	System/Software Development Lifecycle
SKOS	Simple Knowledge Organization System
SMHI	Swedish Meteorological and Hydrological Institute
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/).
SPECS	Seasonal-to-decadal climate Prediction for the improvement of European Climate Services— Project aimed at delivering a new generation of European climate forecast systems on seasonal to decadal time scales to provide actionable climate information for a wide range of users (specs-fp7.eu).
SSWT	ESGF Software Security Working Team
STFC	Science and Technology Facilities Council—CEDA's multidisciplinary science organization, whose goal is to deliver economic, societal, scientific, and international benefits to the United Kingdom and, more broadly, the world (stfc.ac.uk).
TB	Terabyte
TDS	THREDDS Data Server

Acronym	Definition
THREDDS	Thematic Real-time Environmental Distributed Data Services—Web server that provides metadata and data access for scientific datasets using a variety of remote data access protocols (dataone.org/software-tools/thematic-realtime-environmental-distributed-data-services-thredds/).
UI	User Interface
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).
URL	Uniform Resource Locator
UV-CDAT	Ultrascale Visualization–Climate Data Analysis Tools—Provides access to large-scale data analysis and visualization tools for the climate modeling and observational communities (uvcdat.llnl.gov).
VDI	Virtual Desktop Infrastructure (nci.org.au/systems-services/national-facility/vdi/)
ViSUS	Visualization Streams for Ultimate Scalability
W3C	World Wide Web Consortium—An international community that develops web standards.
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 (wcrp-climate.org).
WCS	Web Coverage Service Interface Standard
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 (wcrp-climate.org/WDAC.shtml).
WFS	Web Feature Service
WG	IPCC Working Group
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 (wcrp-climate.org/index.php/unifying-themes/unifying-themes-modelling/modelling-wgcm).
Wget	Web get—A command line web browser and downloader.
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 (earthsystemcog.org/projects/wip/).
WMS	Web Map Service—Standard protocol for serving (over the Internet) geo-referenced map images that a map server generates using data from a geographic information system database.

Acronym	Definition
WPS	Web Processing Service—Provides rules for standardizing inputs and outputs (requests and responses) for geospatial processing services (opengeospatial.org/standards/wps/).
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/).
ZeroMQ	A high-performance asynchronous messaging library.