LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-TR-763939

7th Annual Earth System Grid Federation Face-to-Face Conference Report

H. H. Auten, A. S. Ames, D. N. Williams

December 11, 2018

Disclaimer

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

7th Annual Earth System Grid Federation

December 2017

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.

7th Annual Earth System Grid Federation Face-to-Face Conference Report

December 4–8, 2017
San Francisco, California, USA

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)
European Network for Earth System Modelling (ENES)
Australian National Computational Infrastructure (NCI)
Canadian Network for the Advancement of Research, Industry and Education (CANARIE)

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)
Tom Landry (CANARIE Centre de Recherche Information de Montréal)

Disclaimer

Conference and Report Organizers

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

Preface by ESGF Executive Committee Chair

The Seventh Annual Earth System Grid Federation (ESGF) Face-to-Face (F2F) Conference held December 4–8, 2017, in San Francisco, California, USA, assembled together a collection of independently funded national and international projects comprised of government agencies, institutions, and companies dedicated to the creation, management, analysis, and distribution of extreme-scale scientific data. The purpose of the conference was to discuss sustaining and enhancing the resilient ESGF data infrastructure with friendlier tools for the expanding global scientific community—this year's emphasis was placed on the preparedness of the Coupled Model Intercomparison Project, phase 6 (CMIP6). It also focused on new tools that fulfill important and strategic capability gaps in scientific data archiving, access, analysis, and knowledge discovery.

As the Executive Committee Chair of ESGF, I would like to personally thank each of conference attendees and those who could not attend but contributed to and/or supported the development of the ESGF software stack. It is an exciting time for the ESGF consortium as we continue to grow and adjust, remaining always adaptable, motivated, and responsive to our growing base of community projects. As we move forward, our ESGF organization is confronting and addressing many changes during a time of larger national and international commitment with fewer community resources. That said, our commitment to our sponsors and the community remains strong as we continue to meet the challenges before us and bring inspired developers and the scientific community together through forums like this conference, ensuring our ESGF organization remains robust and at the cutting edge of technology.

Figure 1 summarizes the ESGF's development stages, which rest on several specific strategic research and development (R&D) areas listed center. Over the years, the capability in many of these areas has progressed from an R&D activity to a prototype to a feature that is widely used by the entire ESGF community. The conference examined existing capabilities and prototyped capabilities (e.g., user notification, long-tail publication, data citation, and provenance capture), and introduced several new strategic R&D areas (e.g., machine learning) that are critical for the ESGF's continued success.

More than 60 professionals from 10 countries gathered together to share their knowledge and experiences gained over the past several years. The goals of the conference were to improve the usefulness of the intelligent interagency infrastructure software, explore ideas for new spin-off projects for enhancing the federation, prepare and execute operations for supportive geoscience data archives, and learn from one another in ways that can only happen face-to-face. Conference presentations covered the state of ESGF, discussed development and implementation plans, focused on synergistic community activities, and outlined project and task deadlines. 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 in attendance were researchers interested in incorporating interagency federated service approaches into their science domains such as biology and hydrology.

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), the European Network for Earth System Modelling (ENES), the Australian National Computational Infrastructure (NCI), and the Canadian Network for the Advancement of Research, Industry and Education (CANARIE) (**Figure 1**). Support also comes from other national and international agencies and private industry.

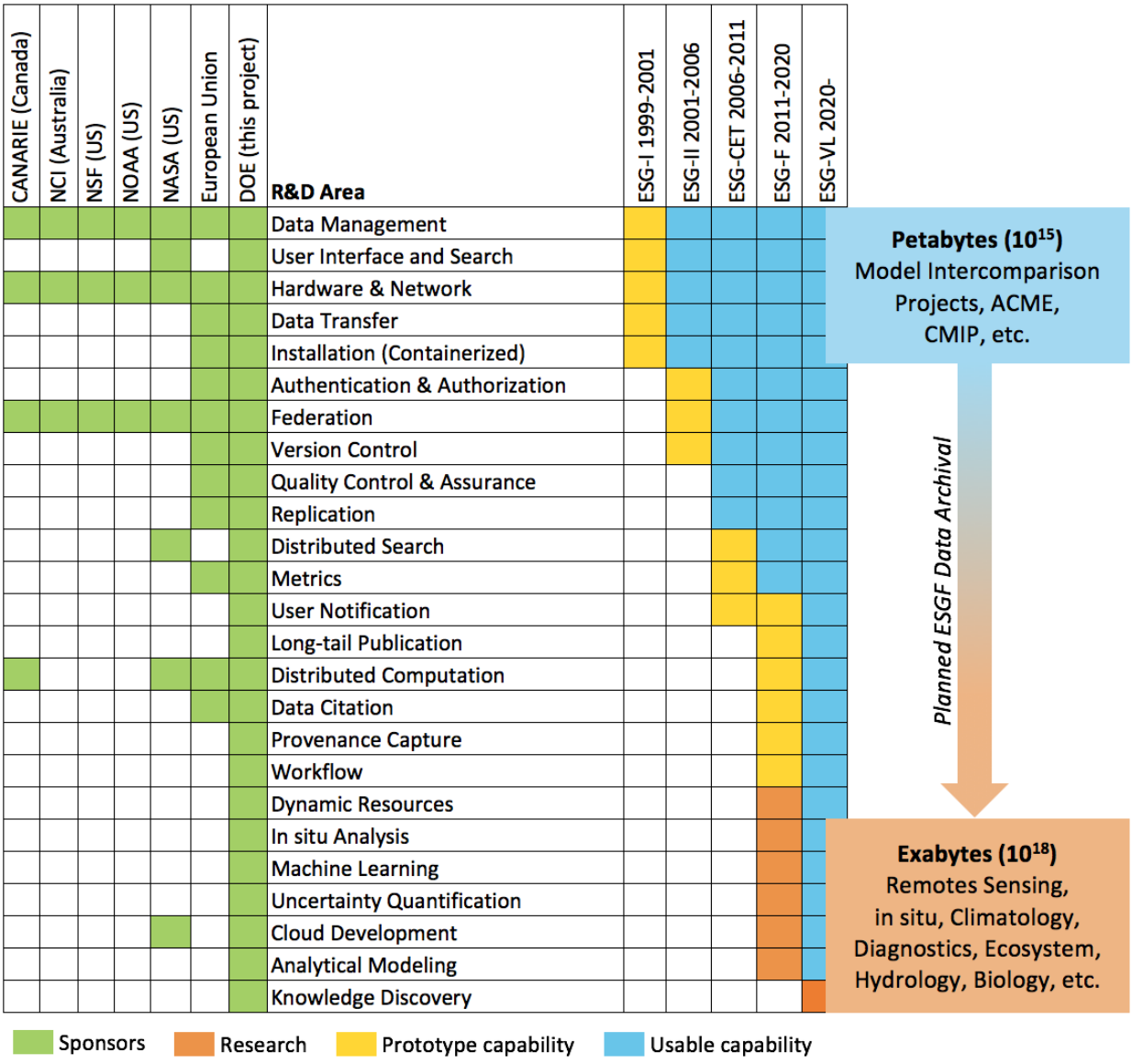
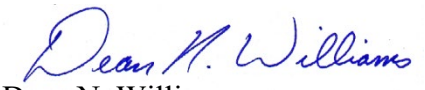


Figure 1. ESGF strategic roadmap. *Columns at left show how DOE, the EU, NASA, NOAA, NSF, Australia's NCI, and Canada's CANARIE support ESGF's strategic R&D areas. Development of the ESGF rests on several specific R&D areas, listed at right. These efforts bring many prototyped capabilities into full community use, and introduce several new R&D areas that are critical for ESGF's success.*

On behalf of everyone involved in organizing the conference, I would like to again thank each of you for attending our conference and bringing your expertise to our gathering. You, as organization leaders, have the vision, the knowledge, the wherewithal, and the experience to help

us pave our way into the future. You are truly our greatest asset today and tomorrow, and we could not accomplish what we do without your support and leadership.

Best wishes to all,



Dean N. Williams
ESGF Executive Committee, Chair

Contents

1. Executive Summary

The Earth System Grid Federation (ESGF) is driven by a collection of independently funded national and international projects that develop, deploy, and maintain the necessary open-source software infrastructure to empower geoscience collaboration and the study of climate science. This successful international collaboration manages the first-ever decentralized database for handling climate science data, with multiple petabytes (PBs) of data at dozens of federated sites worldwide. ESGF's widespread adoption, federation capabilities, broad developer base, and focus on climate science data distinguish it from other collaborative knowledge systems. The ESGF distributed archive holds the premier collection of simulations, observations, and reanalysis data to support the analysis of climate research. It is the leading archive for today's climate model data holdings—including the most important and largest datasets of global climate model simulations. For this long-standing commitment to collaboration through innovative technology, the ESGF has been recognized by *R&D Magazine* with a 2017 R&D 100 Award.

The ESGF facilitates advancements in geoscience by providing:

1. Federated, web-based, application programming interface (API) software and data infrastructure that are easy to use and secure.
2. A flexible infrastructure that allows participating data projects to customize parameters to address their specific requirements.
3. High-performance search, analysis, and visualization tools that make data accessible and useful to the climate research community.
4. Access to a broad set of data and tools for comparative and exploratory analysis.
5. A virtual collaborative environment for diverse research and analysis tasks that demand large and varied data sets.

The ESGF's mission is to facilitate scientific research and discovery on a global scale and maintain a robust, international federated data grid for climate research. The ESGF architecture federates a geographically distributed network of climate modeling and data centers that are independently administered yet united by common protocols and APIs. The cornerstone of its interoperability is peer-to-peer messaging, which continuously exchanges information among all nodes through a shared, secure architecture for search and discovery. The ESGF integrates popular open-source application engines with custom components for data publishing, searching, user interface (UI), security, metrics, and messaging to provide PBs of geophysical data to roughly 25,000 users from over 1,400 sites on six continents. It contains output from the Coupled Model Intercomparison Project (CMIP), used by authors of the Intergovernmental Panel on Climate Change (IPCC) Third, Fifth, and Sixth Assessment Reports, and output from the U.S. Department of Energy's (DOE's) Energy Exascale Earth System Model (E3SM) and the European Union's (EU's) Coordinated Regional Climate Downscaling Experiment (CORDEX) projects, to name only a few.

These goals will support a data-sharing ecosystem and, ultimately, provide predictive understanding of couplings and feedbacks among natural-system and anthropogenic processes across a wide range of geophysical spatial scales. They will also help to will expand access to relevant data and information integrated with tools for analysis and visualization supported by the necessary hardware and network capabilities to make sense of peta-/exascale scientific data.

In the future, the ESGF intends to widen its scope to include other climate-related datasets such as downscaled model data, climate predictions from both operational and experimental systems, and other derived data sets. Over the next few years, we propose to:

- sustain and enhance a resilient data infrastructure with friendlier tools for the expanding global scientific community, and
- prototype new tools that fill important capability gaps in scientific data archiving, access, and analysis.

By supporting the ESGF's critical role in the community, ESGF funders have led a confederation of national and international organizations to advance geoscience by addressing climate's big data issues (**Figure 2**). The ESGF integrates advanced software for data discovery, management, visualization, and analysis.

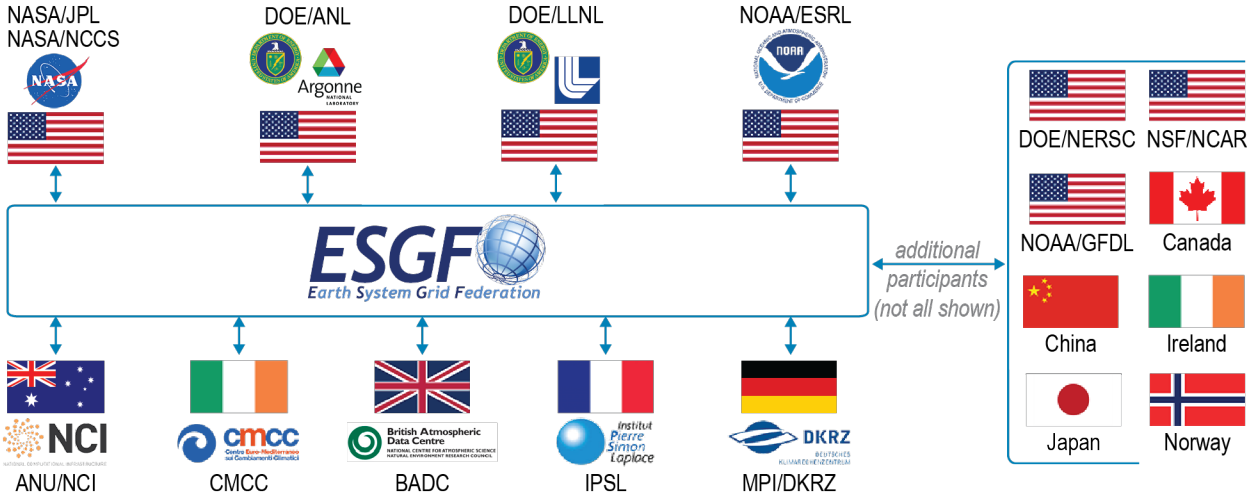


Figure 2. Federated sites. *The ESGF represents a close collaboration between interagency partners from disparate domains and is enabling the development of a data ecosystem that can support a broad variety of data and disciplines.*

The ESGF has transformed climate data into community resources by creating a virtual, collaborative environment that links climate centers and users around the world to models and data via a computing Grid environment that is based on the world's supercomputing resources and the Internet. The ESGF merges numerous independent software applications to:

- integrate the world's climate model and measurement archives;
- create infrastructure for national and international model/data intercomparison studies;
- analyze and visualize the world's climate simulation and observational datasets;
- share resources across multiple centers for high-performance computing (HPC) and storage; and
- move tens of PBs of data across national and international network infrastructure.

The information emerging from collaboration between interagency partner meetings influences requirements, development, and operations. In 2016, representatives from a significant fraction of projects utilizing ESGF to disseminate and analyze data attended the sixth annual ESGF Face-to-Face (F2F) Conference (DOI: 10.2172/1369382). 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. Focusing on federation resiliency and reaffirming 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, the ESGF Executive Committee decided to reorganize the working teams for better synergy and greater alignment. See Table 1 for the newly merged working team list and representatives.

Table 1. The current list of ESGF technologies, designated working team leads, and team descriptions.

Team	Team Leads and Funding Agencies/Institutions	Description
1. User Interface, Search, and Dashboard Working Team	Luca Cinquini (NASA/JPL), Guillaume Levavasseur (IPSL), and Alessandra Nuzzo (CMCC)	Improve ESGF search and data cart management and interface; ESGF search engine based on Solr5; discoverable search metadata; statistics related to ESGF user metrics
2. Compute and Data Analytics Working Team (CWT)	Charles Doutriaux (LLNL) and Daniel Duffy (NASA)	Develop the capability to enable data analytics within ESGF
3. Identity, Entitlement, and Access (IdEA) Working Team	Philip Kershaw (CEDA) and Lukasz Lacinski (ANL)	ESGF X.509 certificate-based authentication and improved interface
4. Installation Working Team and Software Security Working Team	William Hill (LLNL), Sasha Ames (LLNL), and Prashanth Dwarakanath (LiU)	Installation of the components of the ESGF software stack; security scans to identify vulnerabilities in ESGF software
5. Containers Working Team	Luca Cinquini (NASA/JPL) and Sebastien Gardoll (IPSL)	Design and implement a new ESGF architecture based on containerization technologies
6. International Climate Network Working Group and Replication/Versioning and Data Transfer Working Team (ICNWG)	Eli Dart (DOE/ESnet), Lukasz Lacinski (ANL), and Stephan Kindermann (DKRZ)	Increase data transfer rates between the ESGF climate data centers; Replication tool for moving data from one ESGF center to another; ESGF data transfer and enhancement of the web-based download
7. Node Manager Working Team and Tracking/Feedback Notification Working Team	Sasha Ames (LLNL) and Tobias Weigel (DKRZ)	Management of ESGF nodes and node communications

Team	Team Leads and Funding Agencies/Institutions	Description
8. Publication, Quality Control, Metadata, and Provenance Capture Working Team	Sasha Ames (LLNL), Katharina Berger (DKRZ), and Bibi Raju (PNNL)	Capability to publish datasets for CMIP and other projects to ESGF; integration of external information into the ESGF portal
9. User Support and Documentation Working Team	Matthew Harris (LLNL) (includes representatives from Tier 1 data centers, Tier 2 modeling centers, and the above working teams)	User frequently asked questions regarding ESGF and housed data; document the use of the ESGF software stack
10. Machine Learning Working Team	Sookyung Kim (LLNL/AIMS), Sandro Fiore (CMCC)	Research in the applicability of various ML techniques and development of tools/analysis capabilities for domain scientists
11. Diagnostics Working Team	Zeshawn Shaheen (LLNL/AIMS), Tom Landry (CRIM)	Diagnostics software compatible with the ESGF platform; data analysis and validation support

Long-Term Objectives

Although the ESGF has yielded major advances in climate modeling and the management and sharing of the distributed exascale data via Grid technology, climate researchers still have a daunting task of locating, acquiring, and integrating collections of simulation and observational peta-/exascale data that involve the multidisciplinary study of Earth systems.

Working directly with national and international geoscience analysis projects, the ESGF consortium has deployed data and computational resources useful to many stakeholders, including scientists, policymakers, and the public. Consortium members collaborate with other U.S. and international institutions, universities, and private industry entities to develop and integrate a software system for data discovery, management, visualization, workflow analysis, and provenance, which furthers climate researchers' understanding of disparate data from heterogeneous resources. These activities align with the overall goals of the ESGF:

- Work closely with scientific programs funded by national and international agencies to advance the development of state-of-the-art tools to meet geoscience requirements, which includes making public data archives of model output and associated observational data more useful to stakeholders—climate researchers, policymakers, and the public.
- Meet specific needs of national and international climate science projects by integrating and developing tools and techniques suitable for large datasets in a familiar, distributed, and federated infrastructure.
- Provide to national and international government-funded research institutions and international climate centers a wide range of climate data-analysis tools and diagnostic methods for ultra-large datasets for climate science research and applications.

In the long term, the ESGF will continue to operate a federated climate data infrastructure that spans multiple centers around the world, stores data from different multi-institutional partnership projects, and delivers a wide range of services that support knowledge discovery for a broad user base. Because we anticipate that medium-to-large projects will build downstream services and/or products based on ESGF services, our architecture and software stack infrastructure will be designed according to a general set of design principles that guarantee the highest level of service to projects and users while supporting their longevity and evolution. These principles include: federation of services, unified access control, individual administration of local resources, elimination of single points of failure, open-source software, conformance to standards, documented APIs, adhering to best software practices, software modularity, scalability, workflow and provenance, networks for high-speed data transfers, proactive engagement with stakeholders, and key performance indicators.

These long-term goals highlight the ESGF's special interest in geoscience. However, more recently, researchers in other science domains (e.g., hydrology and biology) have expressed interest in generalizing the ESGF for their use. They see the ESGF's potential to foster international cooperation across scientific domains through global scientific data exchange, improving globally federated scientific workflows, facilitating traceability and reproducible scientific results, and preserving data producers' visibility in data-intensive science applications. Such work will demand advances in distributed resources, Representational State Transfer (REST) APIs, and container-based architectures. This work is funded under the DOE Office of Advanced Scientific Computing Research Distributed Resource for the ESGF Advanced Management (DREAM) project and the EU-funded Copernicus Programme.

Roadmap

The ESGF's next challenge will be to establish better interagency integration amongst its federated institutions, which will resize the ESGF for the exascale realm and equip it to serve future scientific requirements of knowledge discovery. Driven by national and international scientific priorities over the next 5 to 10 years, our project goals are to address scientific needs of data management and analysis requirements of the future.

Figure 3 summarizes the ESGF's development, which rests on several specific research and development (R&D) areas listed at left. Over the years, the capability in many of these areas has progressed from an R&D activity to a prototype to a feature that is widely used by the entire ESGF community. The ESGF strengthens existing capabilities, brings many prototyped capabilities (e.g., user notification, long-tail publication, data citation, provenance capture) into full deployment for community use, and introduces several new R&D areas (e.g., in situ analysis, machine learning (ML), uncertainty quantification, analytical modeling) that are critical for interagency and virtual organization success.

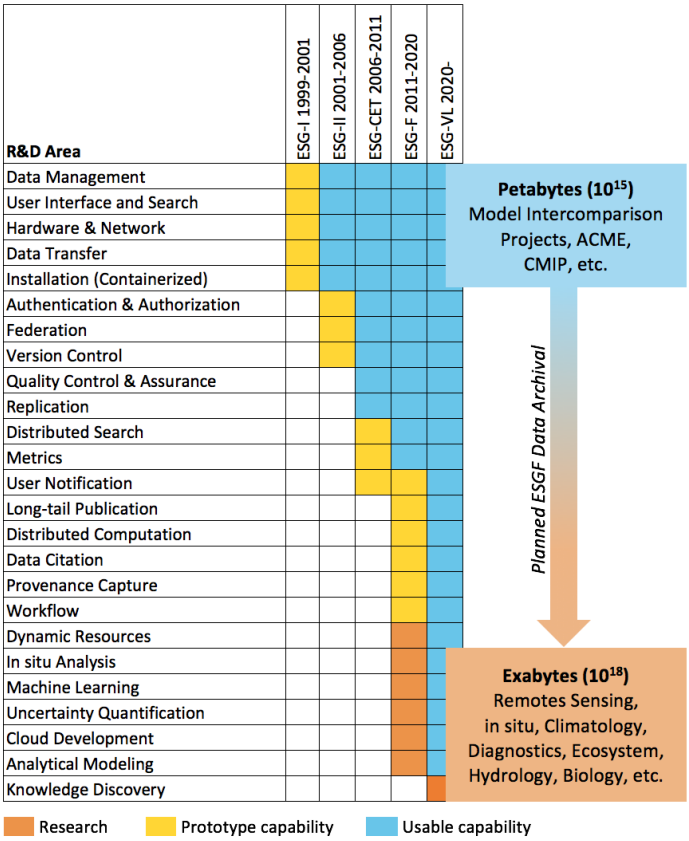


Figure 3. ESGF roadmap. *Development of the ESGF rests on several specific R&D areas, listed at left. The ESGF will strengthen several existing capabilities, bring many prototyped capabilities into full community use, and introduce several new R&D areas that are critical for the its continued success.*

The services sustained and developed by the ESGF will:

- Prepare for the current IPCC Sixth Assessment Report in 2020 (http://wg1.ipcc.ch/AR6/AR6.html), with the archive for CMIP Phase 6 (CMIP6) estimated to be over 40 PB of uncompressed data. This also includes archives for future assessment reports.
- Publish and process the massive data produced by the E3SM, CORDEX, Copernicus, and other supported national and international projects.
- Provide domain-specific tools for climate model evaluation activities under the community investments.
- Leverage the ESGF to help address the domain-specific data needs for other science application areas.

The expanded ESGF environment will provide researchers in climate science and related fields with access to, and deep analysis of, simulation and observational data from sources distributed throughout the world. This federated network will power the climate science community well into the future and strengthen the community by creating an integrated, collaborative data infrastructure. The project will continue to be staffed by climate and computer scientists and software engineers from agencies around the world.

Conference Findings

Progress has been made on many of the findings from the 2016 sixth annual ESGF F2F Conference Report[1] such as persistent identifiers (PIDs) within core data services; a modular Python framework to decrease the difficulty of ESGF installation; additional metrics aggregation functionality; and modularization of server-side computing processes. In discussing both short- and long-term development roadmaps, conference participants agreed that preparation for the release of new data from CMIP6 is the ESGF's top priority—i.e., the logistics and requirements for ensuring a smooth 2018 release of CMIP6 data on ESGF infrastructure—followed by development that utilizes new technologies to support future applications (e.g., in other scientific domains) and other strategic goals that reinforce the ESGF's mission.

A plenary session was held during the conference to determine the key findings from the 2017 presentations and discussions. These were reviewed by the ESGF Executive Committee, and are summarized here:

1. **Operational stability for CMIP6:** The executive committee discussed the importance of focusing on a few components that must work flawlessly across the whole service stack, instead of adding new functionality quickly and causing instability across the federation or for users. Key focus areas are publishing, search, download, and replication.
2. **Data and service challenges (DSCs):** In preparation for CMIP6, a series of tests and software updates will be conducted in early 2018. These challenges will ensure optimal performance and usability of core services (i.e., publication, replication, search, data download, and metadata documentation).
3. **Release management:** A new, integrated ESGF Release Working Team should be created to (1) establish and enforce release timelines and (2) thoroughly test release candidates.
4. **User experience and engagement:** Users need better guidance for understanding and using the services provided by the ESGF.
5. **ML/deep learning (DL):** ESGF working teams are committed to pushing technological frontiers in order to provide the community with consistently reliable, flexible functionality and an optimized user experience. ML/DL are promising technologies that can enhance the ESGF's capabilities for increasingly large amounts of data.
6. **Data access:** Consistent, reliable, and secure access and authentication efforts include access control for data downloads and upgrading to OAuth2 (migration from OpenID).
7. **Provenance:** As with the 2016 report, extending the ESGF's provenance capture capabilities is necessary for increasing usability and reproducibility.

[1] https://esgf.llnl.gov/media/pdf/2017-ESGF_F2F_Conference_Report.pdf

8. **Replication:** Automation of downloads, network tuning between data centers, and data transfer node (DTN) documentation are key facets of improving replication functionality for the federation.
9. **Containerization:** With encouraging progress toward implementing Docker microservices, the ESGF is poised to complete containerization of the architecture during 2018.
10. **Cloud:** Prototyping is under way for deploying ESGF node(s) onto commercial Cloud-based storage services (specifically Amazon). In 2018, more effort is needed to explore the business need, resource allocation, and additional node deployment, if warranted.
11. **Scalability:** Large data requirements will be significant factors in the ESGF's future development. The ESGF must ensure readiness and support ahead of major activities and releases.
12. **Opportunities for future engagement:** Future opportunities are numerous and include both continued development with existing programs and outreach to new global data centers.

2. User/Usage Demographics

Year after year, the ESGF collaboration continues to expand to include more data centers, store more data, and serve a larger scientific community. By the end of 2017, the federation included 31 data nodes and 11 index nodes, distributed across 5 continents (North and South America, Europe, Asia and Australia). Together, these nodes indexed and served data for 23 data collections, totaling approximately 650K datasets[2] and 8.7M files[3] (counting only latest version and non-replicas). At the same time the system of federated composable GUIs (CoG) portals included 147 scientific projects, and 15,571 users registered across all sites. We expect that in the next few years these numbers (data holdings and registered users) will increase ten-fold as ESGF starts serving output from CMIP6 and related model intercomparison projects (MIPs).

3. Scientific Challenges and Motivating Use Cases

ESGF provides science enablement services to support climate and other 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.

This coming year, the ESGF faces a key test of its infrastructure to support CMIP6. On the order of 20 PB of model output is expected to be ingested into the ESGF from 33 institutions, using 75 models while conducting 248 experiments sanctioned for CMIP6. Additionally, the ESGF will

[2] https://esgf-node.jpl.nasa.gov/esg-search/search/?offset=0&limit=0&type=Dataset&replica=false&latest=true

[3] https://esgf-node.jpl.nasa.gov/esg-search/search/?offset=0&limit=0&type=File&replica=false&latest=true

replicate key model output (on the order of 2–4 PB) on three continents (North America, Europe, and Australia) to provide better access to the global user base of the model output. This is an order of magnitude greater than the data volume supported for CMIP5, and it represents a major stressing use case for the ESGF. Development activities during the past year have been focused on this challenge. In preparation for a "go live" date of June 1, 2018, the ESGF will hold three data challenges prior to June 1 to verify the integrity and robustness of the infrastructure.

In looking to the future, the ESGF must face exascale data management challenges in a globally distributed data federation 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

Each of these will drive the ESGF requirements for scalability and analytics services for the end users. ESGF must also 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, ranging from general to more specific aspects.

Open Science Cloud Challenge

Funding agencies support diverse research programs in many scientific disciplines. Climate research is just one of them, and it even encompasses a range of different research projects and data types including numerical data (climate models), satellite data (Earth observation [EO]), 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 sectoral research infrastructure where specific "activities" curate data within the sector. Current examples are CMIP6 (climate modeling experiments), Obs4MIPs (satellite observations), CORDEX (downscaling experiments), and CREATE (Collaborative REAnalysis Technical Environment; reanalysis). In the longer term, the federation expects discipline-specific research infrastructures to form open science clouds. ESGF's data management, online analysis, and interoperability among disciplines will be foundational to the eventual existence of these science clouds. Exploratory work is beginning within the ESGF to enable public cloud computing directly upon the ESGF holdings.

Data Challenge

Scientific data are isolated within domain specific archives, specialized to a domain, growing exponentially, and difficult to analyze cross-discipline due to the following factors:

- Lack of common data structures and formats
- Incomplete and inconsistent metadata resulting in search and discovery difficulties
- The requirement to move more data than necessary from the repository to the analysis platform

Overcoming these challenges requires consistent metadata definitions across disciplines (or at least unambiguous translation of metadata among discipline ontologies) 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. The CMIP6 data challenges to be undertaken this year will be a key step towards addressing this data challenge.

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 enforcing and validating 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 being implemented this year to holdings in ESGF will be game changing for the research community in both attribution and provenance of research results.

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 as data volumes continue to grow at rates that exceed the growth in network speed and bandwidth. Visualization and intercomparison tools are already 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.

Current development activities within the ESGF on data analytics and visualization tools which are seamlessly integrated into the search and retrieval facilities will begin to be tested this year. These capabilities are expected to have a major impact on the ESGF user community for analyzing CMIP6 data.

4. Conference Findings

The annual ESGF F2F Conference offers development guidance and project prioritization to working teams within the ESGF developer community. The 2017 conference brought together a multidisciplinary, multinational group of experts from the computer science, climate science, and research communities to discuss the ESGF. 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 plenary discussions to address progress, component interoperability, and roadmaps. This year's conference focused on CMIP6 readiness and related projects, though significant consideration was also given to long-term roadmaps and the potential of new technology to enhance the ESGF's capabilities.

Operational Stability for CMIP6

The top priority for the whole ESGF collaboration in 2018 is to prepare and operate a very stable, reliable, and usable infrastructure in service of the CMIP6 community—both data providers and data users. Consensus during the conference was that the collaboration should

focus its efforts on a limited set of services that must operate flawlessly, rather than trying to add new functionality that might not be fully polished or stress-tested. This set of core services was identified to include publication, replication, search, download and metadata documentation (which encompasses PIDs, digital object identifiers [DOIs], errata, and model metadata). In turn, this requirement of operating a very stable federation has direct implications on several areas including easiness of installation, reliability of release and testing procedures, usability of services, and user support, as described in more details in the other findings that follow.

Data and Service Challenges

One major outcome of the conference was the recognition that the collaboration has much work to do in reassuring itself and the community about its readiness to manage and serve CMIP6 data. To address this issue, the collaboration has unanimously embraced the concept of establishing a series of DSCs that would progressively test and refine the reliability of ESGF services at all sites that will host CMIP6 data. We are currently planning to execute three DSCs, starting in January 2018. Each challenge will use the currently stable release of the ESGF software to test installation, operations, performance, and usability of the core services needed for CMIP6: publication, replication, search, data download and metadata documentation. Each challenge will take approximately 2 weeks to execute, followed by a 2-week period where necessary adjustments are made to the software to prepare a new stable release that addresses any issues that might have emerged. Each DSC will use an increasingly larger data volume, starting with the currently available test CMIP6 data, and multiplying that by a factor of 10 in each successive DSC.

Release Management

The ESGF service is a complex set of integrated services and software. As with all software, integrated ESGF releases must maintain a given version with major and minor release detail, such as clarifying which version is the current production service and updating appropriate documentation and communication to the user community. In addition, the release process must confirm uniformity of service of all major services at the Tier 1 nodes (rather than just released at one node). Accordingly, conference discussions revealed a need for an Integrated ESGF Release Working Group to (1) establish release policies, timelines, functionality checks, and rollback procedures; and (2) execute thorough testing and validation of release candidates.

User Experience and Engagement

Scientists using CMIP6 data for their research need uniform output with machine-interpreted metadata; easy access to model output via the data catalog, CoG search, and replication; documentation of models and simulations with standardized, searchable information (e.g., experiment conditions) for higher reproducibility and community review; and easy access to errata. Meeting these varied needs requires careful implementation of DOIs (high level) and PIDs (granular level). Citation and data-tracking services to handle DOI/PID assignments to datasets are in development, and ID-tracking functionality is a planned upgrade to the publication registration service. Essential server-side computation functionality includes subsetting (i.e., avoid downloading unnecessary data) and reducing volume prior to data transfer (e.g., climatology, zonal mean). Modeling groups expect to find a list of output fields; clearly defined specifications of CMIP6 model output, such as controlled vocabularies (CVs) and global attributes; and software that helps meet data and performance standards. Conference attendees discussed establishing a user group to review UI usability and make constructive suggestions, ideally in the short term and composed of external users (i.e., not ESGF developers).

Machine Learning/Deep Learning

ML is the field of programming computers to learn to complete new tasks without explicit instructions; DL is a method of ML in which computers learn based on data representations (e.g., image recognition). The use of these technologies in the climate community is highly innovative but in its infancy. As the ESGF looks for ways to incorporate new technology to improve performance, data accuracy, and user experience, ML/DL offer compelling possibilities. For example, the weather and climate modeling community can exploit ML/DL algorithms to identify patterns, perform quality control tasks, improve efficiency of physics code, and inform climate model development.

Scientists are starting to use pattern analysis in massive scaled climate data to can capture nonlinear, underlying patterns in massively scaled climate data, including climate object detection (e.g., eye of a hurricane) and time series analysis (e.g., forecasts). DL will be key for analyzing peta-scale ESGF data by saving human effort and time as well as computing power. Three ongoing ESGF research projects illustrate the potential of ML/DL to transform analysis of climate data:

1. **Detection and localization of extreme climate events:** A convolutional neural network (CNN) learns feature representation of extreme events, ultimately saving computing costs for numerical weather predictions. This three-stage model—collect and label data, detect, localize—produced almost 100% test accuracy in training iterations for hurricane detection (with approximately 4 degrees of regression error).
2. **Increase localization accuracy using pixel recursive super resolution:** A method developed by Google enables reconstruction of high-resolution images using low-resolution images. Then the computer's neural network (NN) is trained to detect image matches (e.g., detect eye, nose, and mouth in a human face). In climate data testing. The NN was trained to look for the centers of hurricanes.
3. **Tracking extreme climate events:** Long short-term memory (LSTM) units of a NN can be programmed to memorize time series data. This method was used to predict hurricane paths, with model predictions tracking closely to ground truth data.

When using ML/DL, it is important to keep in mind that the better the resolution, the better the predicted data. Every dataset is different, with varying consistency of labeling, resolution, and other characteristics. Formation of a new ML/DL-focused working group may be warranted to establish standards for training datasets and to identify priority use cases for future ESGF development. The ESGF would need an ML infrastructure and associated API, and additional research is needed to develop a DL emulator for physical parameterization as well as time series analysis and prediction for tracking climate events. Finally, a labeled dataset published through the ESGF would help promote DL research for climate science.

Data Access

Consistency and operability of access control are paramount in preparation for CMIP6. The ESGF must ensure that all CoG sites allow access to the same MIPs data. The UI must allow users to select MIP era and/or scenario. Additionally, access control should be removed for downloading data for specific projects (e.g., CMIP6, CMIP5, Obs4MIPS, CORDEX) while maintaining access control and registration for compute activities. This approach will require changes in the stack, such as updating wget script generation for index nodes (possibly through configuration file changes) and modifying access policies at all data nodes. The Executive

Committee recommends establishing a task force to evaluate data download methods and make sure that Globus downloads work at each site.

To increase security and ensure proper user permission, several authentication efforts are needed to support OpenID while migrating to OAuth2. Use cases to explore and resolve include accessing MIP data without authentication, using OAuth credentials with other platforms, handling different versions of OAuth, embedding OAuth certificate in wget scripts, and confirming all steps in the OAuth access token workflow.

Provenance

Recent work on the new HArvester Provenance Interface (HAPI) adds flexibility to the Provenance Environment (ProvEn) platform with a generic format based on community standards and wider applications. More investigation and development are needed to ensure CMIP6 provenance data are captured appropriately. For instance, the ProvEn platform can be enhanced with JavaScript Object Notation (JSON) support.

Replication

The conference brought focus to efforts needed to refine replication functionality. First, replication tests must be defined and automated. This includes automating download tests via Synda; work is already under way with selection files for each Tier 1 site. Second, the ESGF must use test replication performance data to tune the network between data centers. Third, the DTN setup procedure must be clearly documented—a process that will renew network architecture discussions, particularly with respect to network architecture diagrams from all Tier 1 sites. Finally, data must be published with DTN URLs. For CMIP6 data from Asian collaborators, questions arise about which data nodes and index nodes to use as well as how to replicate these data. Additional planning discussions are needed to accommodate these use cases.

Containerization

The 2017 F2F conference has reaffirmed and strengthened the need for ESGF to provide a "containerized" version of all its data and metadata services, an outcome which was already emerged from last year's conference findings. Over the past year, ESGF has made great progress in building Docker images for most of its software components, and in defining and prototyping a new container-based version of the full ESGF software stack. This kind of architecture, also known as "microservices" architecture, is becoming increasingly mainstream in the IT realm, since it provides several advantages over traditional software stacks: easiness of installation and testing, flexibility of deployment, modularization, separation of concerns, scalability, and many others. The ESGF/Docker architecture is currently using Docker Swarm as the underlying orchestration engine to manage deployment and interaction among containers, although work has already started in experimenting with Kubernetes as an alternative framework. The goal is to be able to offer a feature complete, totally containerized architecture as an installation option by the end of 2018.

Cloud

Following a general trend in IT for business and science, the ESGF collaboration has recognized the need to start experimenting with the Cloud as a possible platform for deployment of ESGF services. Reasons for using the commercial Cloud include high level of service, high availability, no hardware maintenance, automatic geographic replication, and scalability, just to name a few. This paradigm is particularly appealing for offering or executing data processing jobs on a scalable number of computing nodes, which might be allocated or terminated depending on user

demand. Building on its successful efforts of service containerization, in 2017 the ESGF collaboration has already prototyped deployment of a full ESGF node onto the Amazon ECS (EC2 Container Service) environment. In 2018 we plan to further pursue this task by refining the Amazon Web Services (AWS) deployment to a fully operational ESGF node, and by deploying the same architecture using the Google Compute Engine environment, perhaps leveraging OpenShift as enabling Platform as a Service (PaaS). At this time, it is still unclear what the correct business model for Cloud usage would be, since an ESGF institution could reasonably pay for allocation of Cloud storage and computing resources (if monitored) but would not likely want to assume responsibility for all data transfers initiated by the global user community.

Scalability

Although much effort is currently focused on CMIP6 readiness, the ESGF must always be forward-looking and anticipate the next generation of large data requirements. For example, submissions for the next IPCC Report are due in early 2020. As a major enabler of international climate science research, the ESGF must be ready to support and respond to users' needs ahead of IPCC deadlines with user-friendly data and functionality.

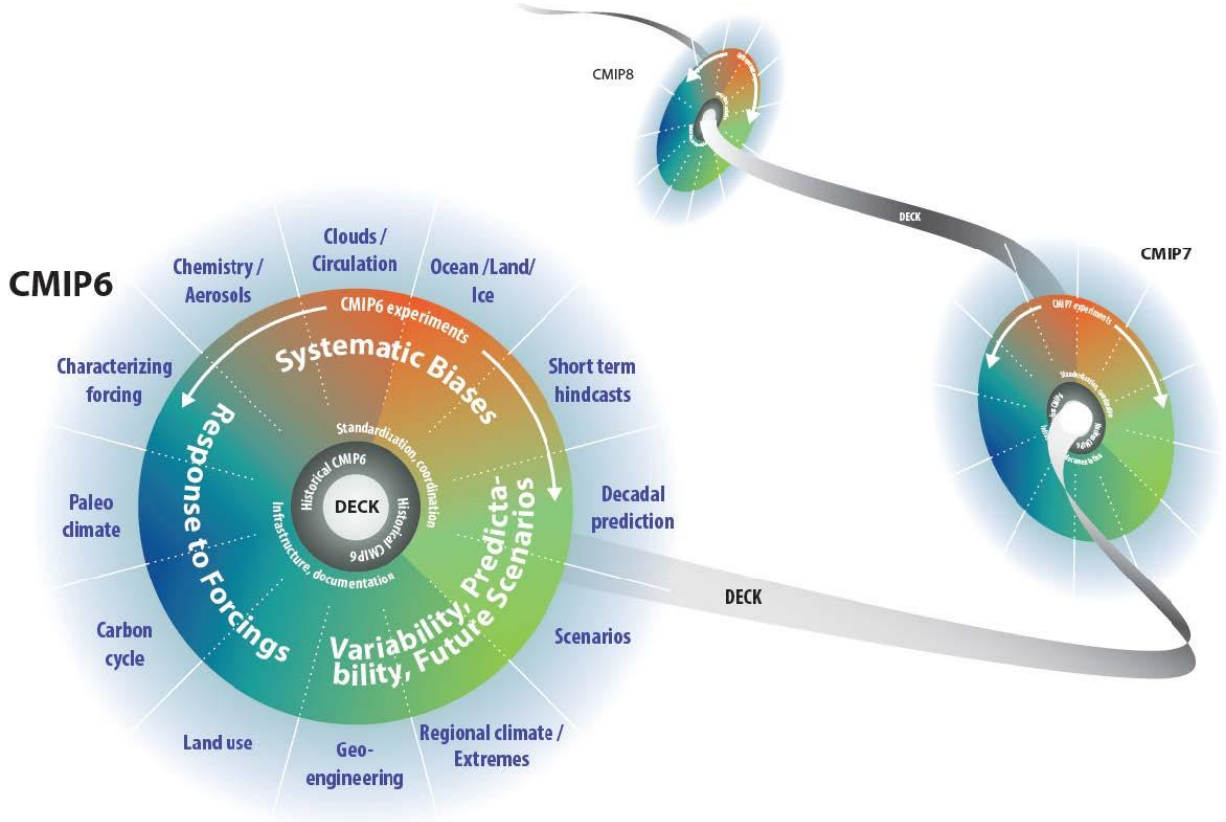


Figure 4. *CMIP provides continuity through DECK and an evolving suite of additional experiments addressing specific science questions. New experiments are proposed as new science questions arise. Planning for CMIP6 necessarily requires looking ahead to future CMIPx phases.*

Opportunities for Future Engagement

Other significant future-scaling efforts include integration with virtual labs and other APIs; continued development to support the Copernicus Programme; outreach to Asian programs and other global data centers not yet engaged; and additional integration of weather, climate, and

observational data. In addition, long-term activities must include ongoing support of CMIP6 beyond the initial launch as well as preparation for later CMIPx versions.

5. Technology Developments

The 2017 F2F conference provided an opportunity for all ESGF working groups to highlight their work over the past year and report on progress and future roadmaps. For more details, please see the working group reports posted online at https://esgf.llnl.gov/2017-F2F.html.

User Interface, Search, and Dashboard Working Team

This working group was formed in 2017 by consolidating three previously separate working groups, with the goal of enabling tighter collaboration in three areas that are closely related to the user experience.

Progress Report

This past year, development of the CoG UI has become a truly collaborative effort, with significant contributions from several institutions (National Aeronautics and Space Administration's [NASA's] Jet Propulsion Laboratory [JPL], Deutsches Klimarechenzentrum [German Climate Computing Centre, or DKRZ], ANL, Centre for Environmental Data Analysis [CEDA]). One major area of development was the integration of new functionality in support of prominent projects such as CMIP6 (data services, errata, PIDs) and Obs4MIPs (quality indicators and ancillary data), as well as interfacing with new capabilities such as the distributed computing services and Visualization Streams for Ultimate Scalability (ViSUS) data streaming. Additionally, development was almost completed to transition CoG to the new OAuth2 authentication framework. Finally, several improvements in usability were executed following last year response to the ESGF user survey, as well as bug fixes and security patches.

The search services have been stable for some time now, and most of the work in 2017 was about small bug fixes and improvements to solve some limiting data publishing use cases. Also, the underlying Solr engine was updated to the latest 5.x release to remedy a security vulnerability.

In 2017, the Dashboard team has worked on bringing the software to operational status and integrating it with the ESGF installer so that it is deployed automatically at all ESGF nodes. A federation testbed (TB) was setup that included JPL, Lawrence Livermore National Laboratory (LLNL), and CMCC with the goal of testing metrics collection and federation capabilities. A collector node at CMCC gathers all the metrics through the esgf-stats-api service and these are visualized through the dashboard-ui (http://esgf-ui.cmcc.it:8080/esgf-dashboard-ui/pages). The statistics include file downloads, such as total volume of downloads, total number of downloads, successful downloads and downloads related to a replicated file and average duration of the download time. At the moment, CMIP5 and Obs4MIPs projects are supported with project specific views, but other projects are also tracked at a coarse grain level in dedicated cross-projects views. New tables and views were designed and implemented to report metrics for the CORDEX project. Also, improvements to the Data Archive section of the dashboard-ui have been carried out with a cache system to make the output visualization highly responsive.

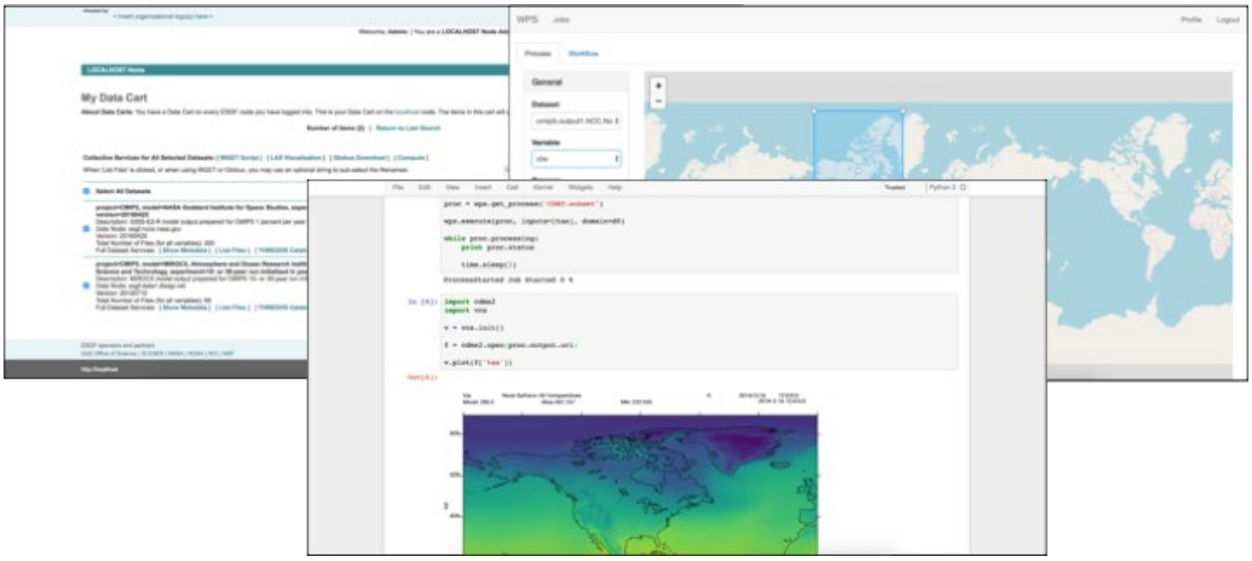


Figure 5. *The new CWT interface for remote distributed computing integrated with CoG.*

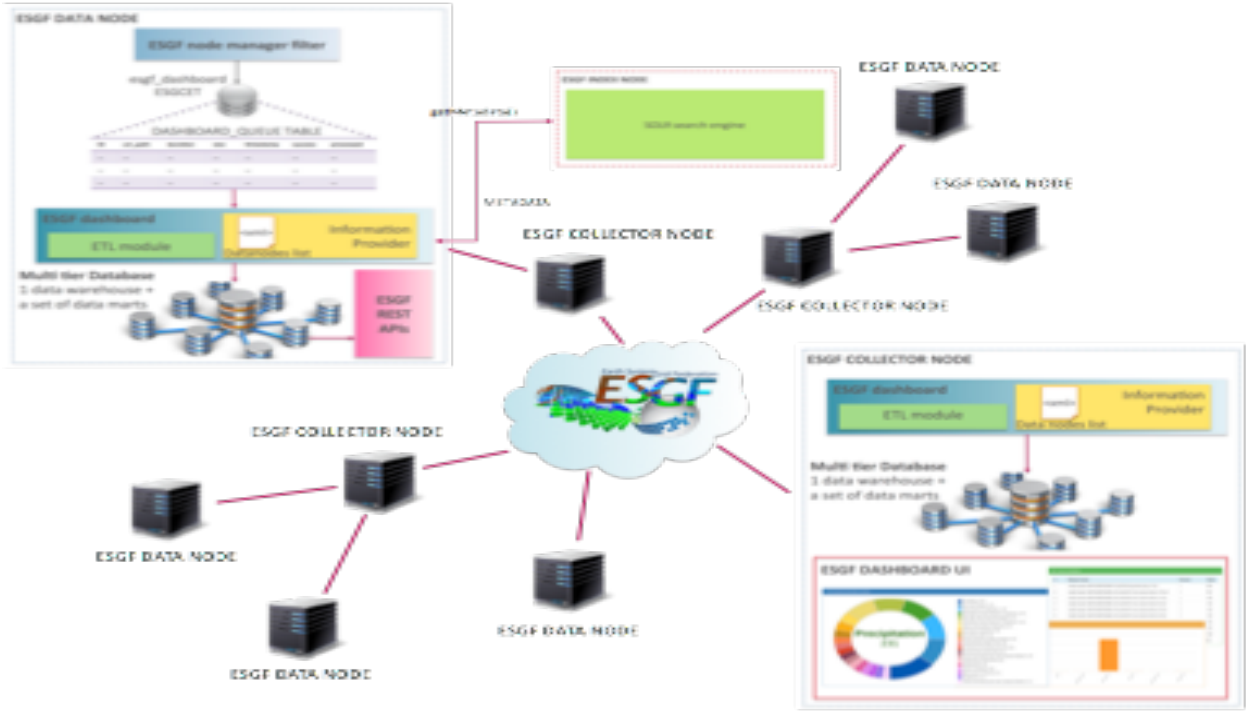


Figure 6. *The Dashboard architecture, deployed on all distributed data nodes, and a few collector nodes that provide metrics aggregation.*

Future Roadmap

For the next year, the topmost priority for CoG development will be to support search and access of CMIP6 data, which will entail changes to the search facets configuration, and testing and refinement of integration with external CMIP6 data services. Also, we plan to finalize the transition to OAuth2 authentication, eventually dropping support for the old ESGF identity

providers (IdPs) based on Openid 2.0. Finally, subject to funding availability, we intend to start working on a complete refactoring of the CoG source code, evolving it in a modular framework of CoGs. These are envisioned to be independent modules, each including a specific functionality (search, user management, wiki, project governance, etc.), that can be installed and configured separately at each ESGF node. As part of this work, the CoG's underlying JavaScript libraries will be upgraded to a more modern and secure framework, as well as its third-party wiki capabilities.

Similarly, the major goal for the search services in 2018 will be to support publishing and searching of CMIP6 data. Work will focus on developing and documenting a server-side API and client-side toolkit to update the Solr metadata "in-place," without the need to completely republish all data. This is necessary to enable backward compatibility with CMIP5 metadata, as some of the search facets will change (for example, from "project" to "activity id"). Also, the underlying Solr engine will need to be gradually updated from the current 5.x series, to the 6.x and eventually 7.x series, taking care to migrate the old catalogs with minimal interruption of service to the community. This process might be synergistic with another goal, which is to establish a new high performance and high-availability search infrastructure based on Solr Cloud, perhaps hosted on a commercial Cloud provider.

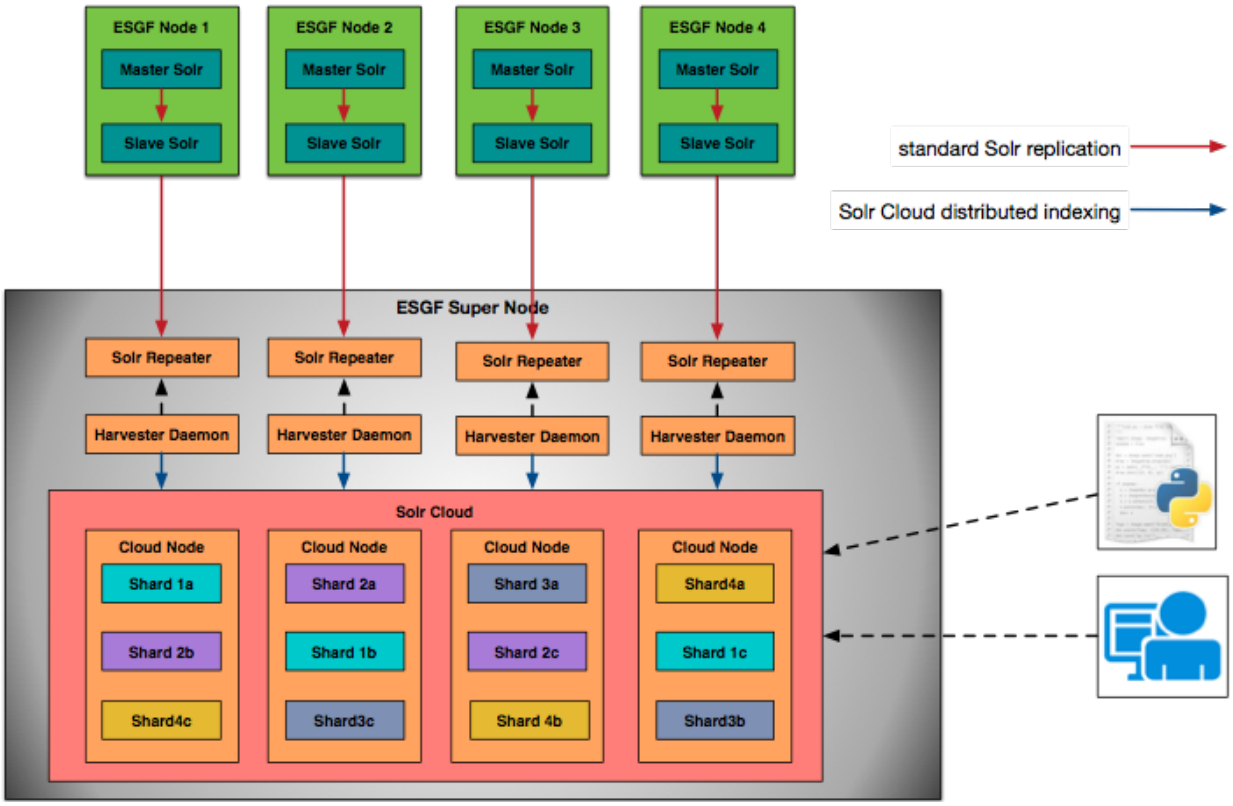


Figure 7. *Prototype architecture for an ESGF "search super node" based on Solr Cloud.*

In 2018, the Dashboard team plans on finalizing deployment and testing of the software at all ESGF nodes, including collection and validation of federation-level metrics. New functionalities are planned, including integration with perfSONAR (Performance Focused Service Oriented Network Monitoring Architecture) statistics, number of users from the new OAuth2 IdPs,

collection of metrics from other data services such as GridFTP servers, and possibly new views and tables for currently supported projects (CMIP5, CMIP6, Obs4MIPs, CORDEX). The team plans to deploy a RESTful service also on the federation-level collector node at the CMCC site to also provide the ESGF community with federated statistics in a programmatic manner.

Compute and Data Analytics Working Team (CWT)

The CWT has accomplished many tasks since the 2016 F2F conference, and much remains in progress going into 2018. With mature backend functionality and a compatible end-user API integrated in to CoG, the team is planning a four-phase roadmap consisting of several important objectives including full support for OAuth, integration into the ESGF release cycle, and more advanced caching.

Progress Report

The web processing service (WPS) API provides a uniform interface to all compliant analytic servers. Several analytic servers have implemented the CWT API, maturing over the past year:

1. Community Data Analysis Tools' (CDAT's) (LLNL) supported operations include serial subset, aggregate, regrid, min, max. Supports curvilinear grids.
2. EDAS's (NASA/NCCS) supported operations include parallelized (Spark) subset, aggregate, regrid, ensemble (mul, diff, min, max, ave, sum), and reduction (mul, diff, min, max, ave, sum, rms). EDAS supports composition of canonical operations into workflows to support additional (composite) operations such as anomalies.
3. Ophidia's (CMCC) supported operations include subset and reduction (max,min).

The CWT WPS API has been integrated into CoG and is now mature enough to be considered as part of the ESGF installation. In addition, the team began work on a common test suite, accessible through GitHub (https://github.com/Ouranosinc/CWT-API-TestSuite).

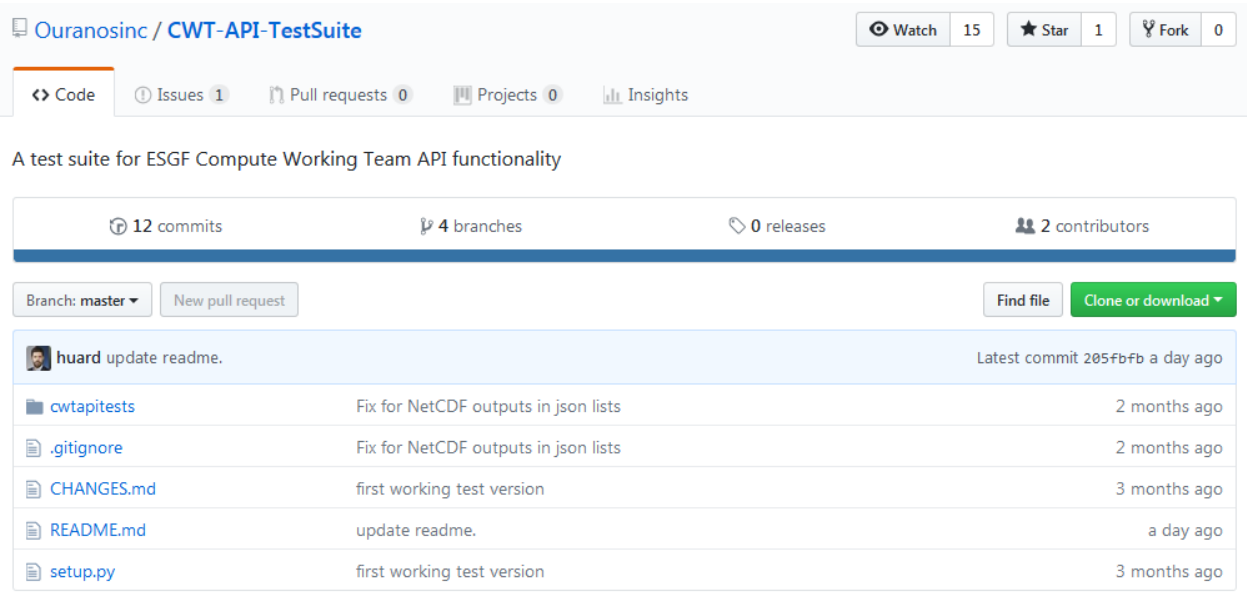


Figure 8. *GitHub interface for the CWT's common test suite. The API allows the same test logic to be applied to multiple implementations.*

We have also developed two versions of the abstract workflow description language: (1) CWT API workflow description, and (2) JSON schema defined to describe workflows based on WPS requests (https://ouranosinc.github.io/pavics-sdi/en/workflows/vocabulary.html).

Future Roadmap

The CWT's 2018 roadmap is organized into four phases. In Phase 1, we intend to standardize operation definitions and provider namespaces (e.g., nasa.averager), which includes defining use specifications and defining discoverability specifications. We will also define the ESGF operation certification process and designate a standard dataset/results for verification testing. This phase includes completing the specification of the common test suite as well as defining and developing services to expose a server's test results to users. For this latter objective, we must make enough information (input files, outputs) available for users to be able to compare their results with those of the CWT implementations.

During Phase 2, the CWT plans to integrate full support for OAuth. In addition to supporting OpenID groups, this means users log onto one site and can then run many. This phase will see integration of the CWT API into the ESGF release cycle, which requires use to (1) define a vetting system for official stack, (2) document methods for adding analytics to local nodes, and (3) develop the v. 3.0 installer. Server interoperability features will be defined and developed:

- Enable compliant analytic servers to call operations on other (remote) servers.
- Ensure scalability of compliant server analytics.
- Define methods for discoverability of accessible (remote) analytics services.

The team will also standardize CWT API exception handling and support other teams' efforts in achieving server compliance (e.g., Ouranos/PAVICS) during Phase 2.

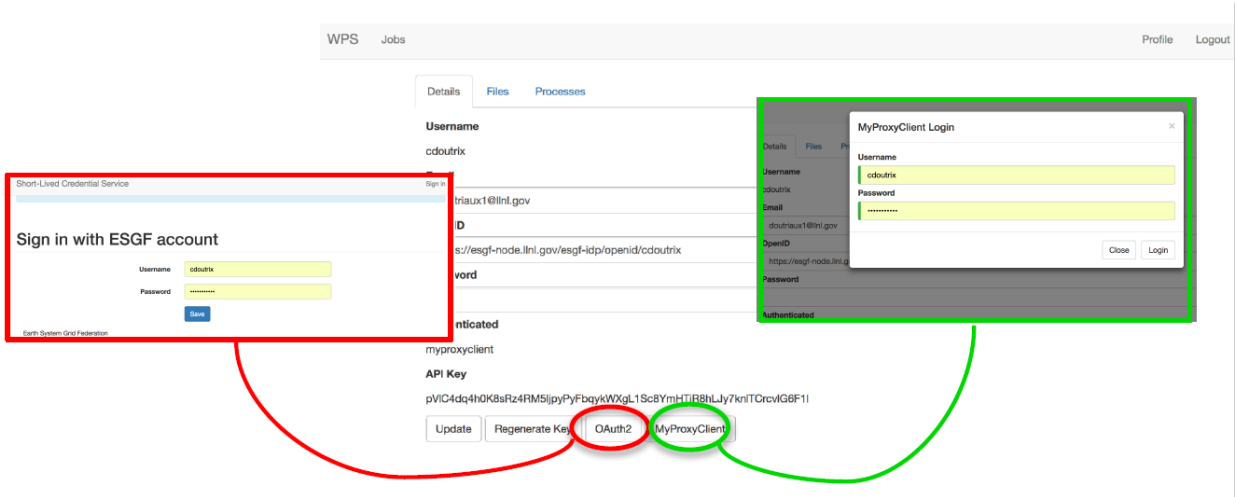


Figure 9. *Interface showing dual certificate support for OAuth2 and MyProxyClient.*

In Phase 3, main objectives include generalizing the workflow grammar and develop an execution engine to support meta (cross-server) workflow execution. Specifically:

- Develop the syntax to specify a generalized operation (as opposed to a server-specific implementation).
- Optimally map operations to existing services and frameworks.
- Develop an introspection API to query available services/operators.

We also plan to implement additional operators for binning operations, statistics and correlations, handling of irregular grids (possibly CMIP6), and ML. Additional work includes performance logging and analytics (i.e., record data size and compute time; assist in workflow optimization) and advanced produce caching (i.e., define cached product expiration times). Last but not least, we expect to integrate provenance handling to record the data, process, and server for each execution.

Finally, in Phase 4 the CWT will define a resource management operation that considers user permission, storage, and CPU use. We plan to conduct a dry run of basic checks for data availability and permissions.

Identity, Entitlement, and Access (IdEA) Working Team

Progress Report

Significant progress has been made this year with the integration of new identity management services into ESGF: OAuth2 and the Contrail short-lived credential service (SLCS). OAuth2, enables support for new usage scenarios in which third-party software agents (e.g., compute or visualization services) may act on behalf of a user to access secured resources. The Contrail SLCS provides equivalent capability to MyProxyCA already in use with ESGF but with the key difference that it supports a web service interface enabling it to be used in conjunction with OAuth to support the new delegation usage scenarios.

OAuth server-side components and SLCS have been integrated into the ESGF Installer, deployed in production at the Program for Climate Model Diagnosis and Intercomparison (PCMDI) and CEDA and test at other nodes in the federation. A key challenge has been to re-engineer dependent services to support the new OAuth client interface and to do so in such a way as to enable a smooth migration from existing legacy identity services. In summary, the new components are:

- A standalone Django OAuth Client which together with changes to TDS (THREDDS data server) authentication filter functionality, enables OAuth2-based sign in for Data Node. This is deployed in Alpha at Argonne National Laboratory (ANL).
- OAuth2-based sign-in for CoG (code fork, in development).
- Service discovery functionality to enable dual support for OAuth and OpenID 2.0 in the production federation. This will allow smooth transition between old and new technologies.

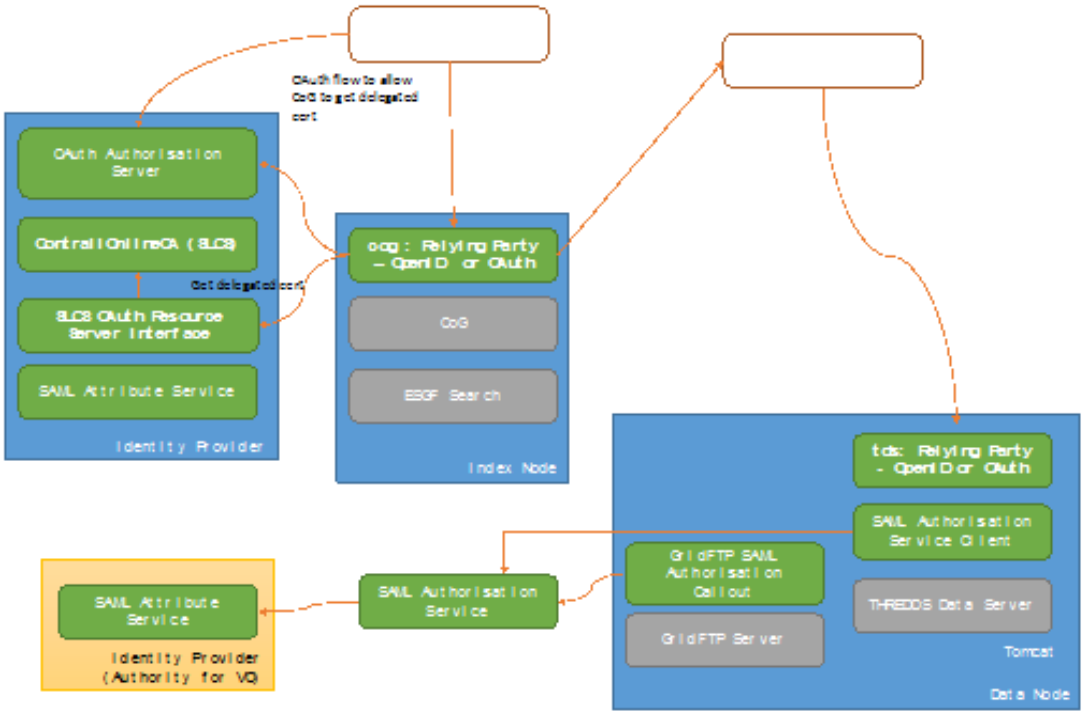


Figure 10. *ESGF IdEA architecture showing new and enhanced components in bold.*

Future Roadmap

The new year will focus on the development and integration of remaining client components:

- Full operational support for OAuth: Integration of OAuth Client and new TDS authentication filter into the Installer and the Container-based distribution under development with the Container WT.
- Simplified scripted HTTP data download mechanism. Using OAuth, it is possible for a server-side component such as CoG to obtain a delegated user certificate on behalf of the user and embed this in a data download script. With the credential included in the script, the user no longer needs to re-authenticate when they invoke the script.
- Over the course of the year we would like to retire legacy services: OpenID 2.0 and MyProxyCA.
- Investigate extension of OAuth capabilities for other use cases in ESGF:
    - Support for OpenID Connect. OpenID Connect builds on OAuth2. OpenID Connect support would facilitate use with Globus Connect Server v5.
    - Work alongside development efforts with the CWT and other usage scenarios in ESGF which require access control and user delegation.
    - Tooling to simply the registration of OAuth clients services with OAuth Authorization Services hosted by IdPs (e.g., use of the Dynamic Client Registration Protocol [https://tools.ietf.org/html/rfc7591]).
    - Data download using OAuth access token. This would provide a simpler mechanism than the existing SSL-based method using user certificates.

- Replace SAML (Security Assertion Markup Language) Attribute Service with OAuth equivalent. This would amalgamate the functionality of the Attribute Service into the OAuth Authorization Server reducing the code base and number of interfaces that need to be supported in ESGF.

Installation and Software Security Working Team

The ESGF Installation Working Team (IWT) made progress in 2017 that provided enhancements to the current installation stack as well as progress towards a new, revamped version of the installation code. Two minor versions of ESGF were released this year, version 2.4 and version 2.5. Each version also had numerous patch releases to improve security and stability.

Progress Report

A number of releases were completed in 2017. Version 2.4 of ESGF was released in January 2017. Twenty-four additional patch releases for version 2.4 were pushed into production. Version 2.4 featured updates to the ESG Publisher, ESG-Search, and COG. A Certificate Policy and Certification Practices Statement (CP/CPS) was delivered with an aim to regulate certificate authority practices on ESGF. Additionally, an application was developed to display the certificate authority's CP/CPS and lists of signed and revoked certificates. Version 2.5.9 of ESGF was released in June 2017. Several patch releases have occurred since the initial production version, with the current version being 2.5.17. Version 2.5 featured significant changes to several components in preparation for publishing CMIP6 data. The auto-installer script was also refactored and streamlined for easier use.

The ESGF build process has been condensed from ten scripts to one. The build scripts were ported from Bash to Python for increased readability and error handling. A menu interface was added to make it easier to build individual components. GitHub Issues is now the definitive source of reporting issues and tracking new features. A set of guidelines were created for reporting new issues when they are discovered. We have also integrated the GitHub issue tracking with Slack, the team communication tool.

For the ESGF 3.0 effort, a full refactor of the ESGF installation stack has been in development and has nearly reached an Alpha release. The entire installation script is being rewritten in Python 2.7. In addition to being ported to Python, the code base is being refactored to be more modular and more testable.

Future Roadmap

The IWT plans several major tasks in 2018. The new Python-based script for ESGF installation will be completed. A continuous integration pipeline will be implemented to facilitate automated testing. Upgrades to the Installer will allow support of Redhat/CentOS7 and Python 3. Documentation for contributing and installing ESGF will be revised and updated with a cleaner format. In the first half of 2018, we anticipate at a least an additional 2.x release (starting with 2.6). Such a release will not feature any major component overhaul, but primarily focus on minor changes needed for CMIP6 readiness.

Containers Working Team

This working group was constituted in 2017 to provide a unified strategy in designing a new ESGF architecture based on containerization technology, which is a fast-growing trend in computing applications and infrastructure. This effort builds on previous work in Docker

containers funded by the DREAM project, and it will now be co-funded by the European Copernicus program.

Progress Report

In 2017, the working group has made great progress towards the ultimate goal of being able to install and operate a fully functional ESGF node where all services are containerized and deployed on a scalable cluster of nodes. Work is currently focusing on using Docker containers and Docker Swarm for orchestration, with the intent of later evaluating Kubernetes as alternative orchestration option. ESGF/Docker version 1.4 was released in December, which is *almost* a feature-complete ESGF architecture, since it includes containerized services for user registration, authentication, authorization, data publishing and searching, the CoG UI, the dashboard, and the new prototype OAuth2 security components. The most prominent ESGF components not yet included in this release are GridFTP/Globus and the Live Access Server. This software stack was successfully deployed on developers' laptops, an in-premise Linux cluster at Institut Pierre-Simon Laplace (IPSL), and on the Amazon Cloud.

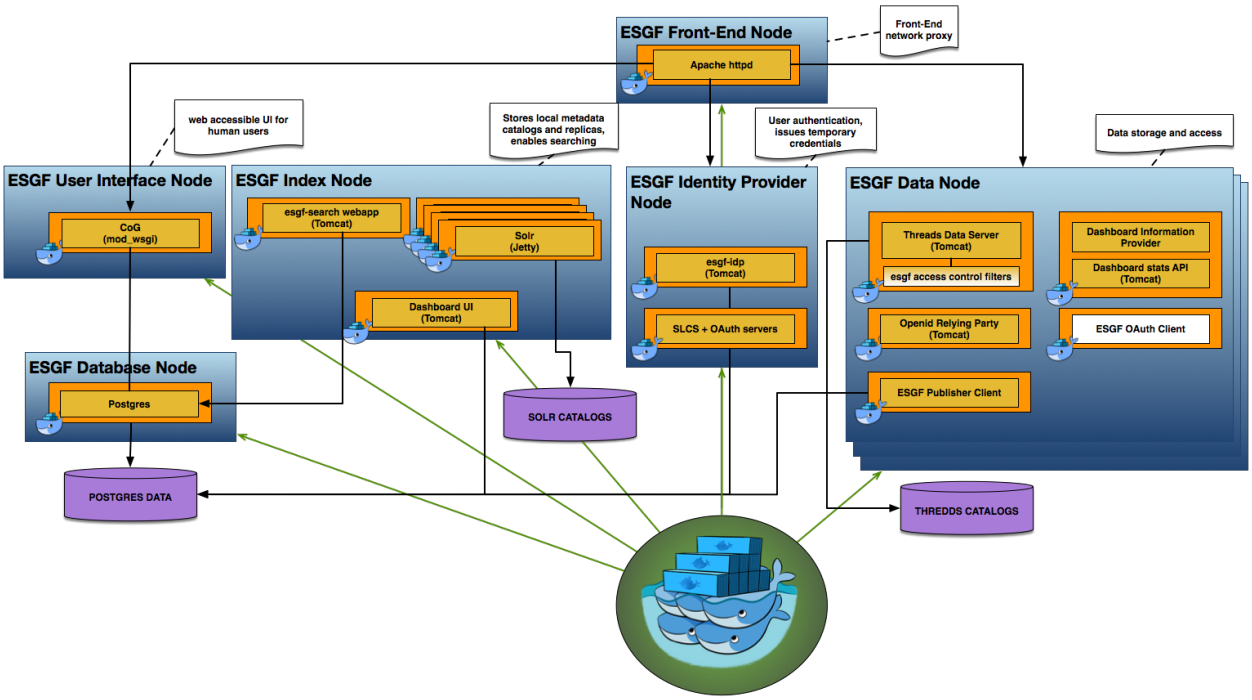


Figure 11. *ESGF Docker architecture v1.4 as deployed with Docker Swarm on a 6-node cluster.*

Future Roadmap

For next year, the group intends to bring all current work to fruition, by making available a fully featured, container-based ESGF Node ready for operations. JPL plans to switch its ESGF node to the Docker based architecture sometimes in 2018, as well as establishing a reference node on the commercial Cloud, and other sites will probably follow suit.

To achieve this goal, the group needs to work on providing a container-based version of the Globus Connect Server, as well as on thoroughly testing and refining existing containers such as the Dashboard, Node Manager, the OAuth2 services, the WPS compute engine, and ViSUS engine. Additionally, work will start on providing a complete testing suite, continuous

integration, migration tools from current ESGF nodes, automatic security updates and scalability. Finally, we hope to support both Docker Swarm and Kubernetes as alternative deployment infrastructure for the ESGF containers.

International Climate Network Working Group and Replication/Versioning and Data Transfer Working Team (ICNWG)

Progress Report

An efficient replication infrastructure is of key importance to support users in accessing and analyzing the coming CMIP6 high-volume datasets. A core component of this infrastructure is a well-integrated "backbone" of Tier 1 sites with optimized data replication paths offering large replica pools for end users and downstream (Tier 2) sites as well dedicated service providers (e.g., supporting the climate impact community).

In 2017, the work of the ICNWG concentrated on the establishment of a test-federation based replication TB involving Tier 1 sites at CEDA, DKRZ, IPSL, LLNL, and Australia's National Computational Infrastructure (NCI). This testbed was used for replication tests based on Synda. Efforts to deploy replication in production are ongoing.

Future Roadmap

A TB was established by partners to test the end-to-end replication workflow. Work initially concentrated on the publication of replica data collection at the sites. As different sites follow different policies, the goal to establish a complete end-to-end replication workflow across sites was abandoned. Some partners integrated and tested automatic republication of replicas (based on the Synda post-processing module), yet strong requirements exist at other sites to separate the replication by the re-publication by manual control steps of data managers. Whereas the establishment of TB-based routine replication exercise was successful between partners, performance is still low. The main reason for this is the different state of readiness for production deployment at sites. Many sites are still preparing and adapting their infrastructure (e.g., by deploying dedicated DTNs) to the needs of CMIP6 replication starting in mid-2018.

The future roadmap is separated into two streams: short-term activities to support CMIP6 data replication in 2018/2019, and long-term planning to efficiently support future replication efforts (CMIP6+).

Short-term roadmap:

- Engage in the CMIP6 data challenge effort to test CMIP6 data replication functionality as well as scalability based on real CMIP6 test datasets. A set of replication related tasks were defined to ensure core replication infrastructure is operational as soon as the first CMIP6 experiments are published in the ESGF (staring in July 2018).
- Based on data challenge experiences, finalize network- and DTN-related infrastructure setup at Tier 1 sites.
- Start integration work of new Globus connect server v5 into ESGF infrastructure (as part of data node installation) as well as DTN deployments.
- Start coordination discussions with respect to CMIP6 replication priorities at sites.

Long-term roadmap:

- Integrate and exploit Globus transfer in the replication pipeline.
- Expand DTN deployments to match future infrastructure requirements as well as data scale.
- Integrate replication speeds and available bandwidth information into the ESGF monitoring dashboard.
- Beyond Tier 1 replication infrastructure, begin looking into optimizing Tier 2 (Tier 1 replication pipeline).
- Work on DTN endpoint installation instructions as well as installation support (e.g., providing "ESGF-ready" installation packages).

Node Manager and Tracking/Feedback Notification Working Team

In mid-2017, Tobias Weigel was selected to co-lead this working team to include the node manager, user notification, and messaging/PID services.

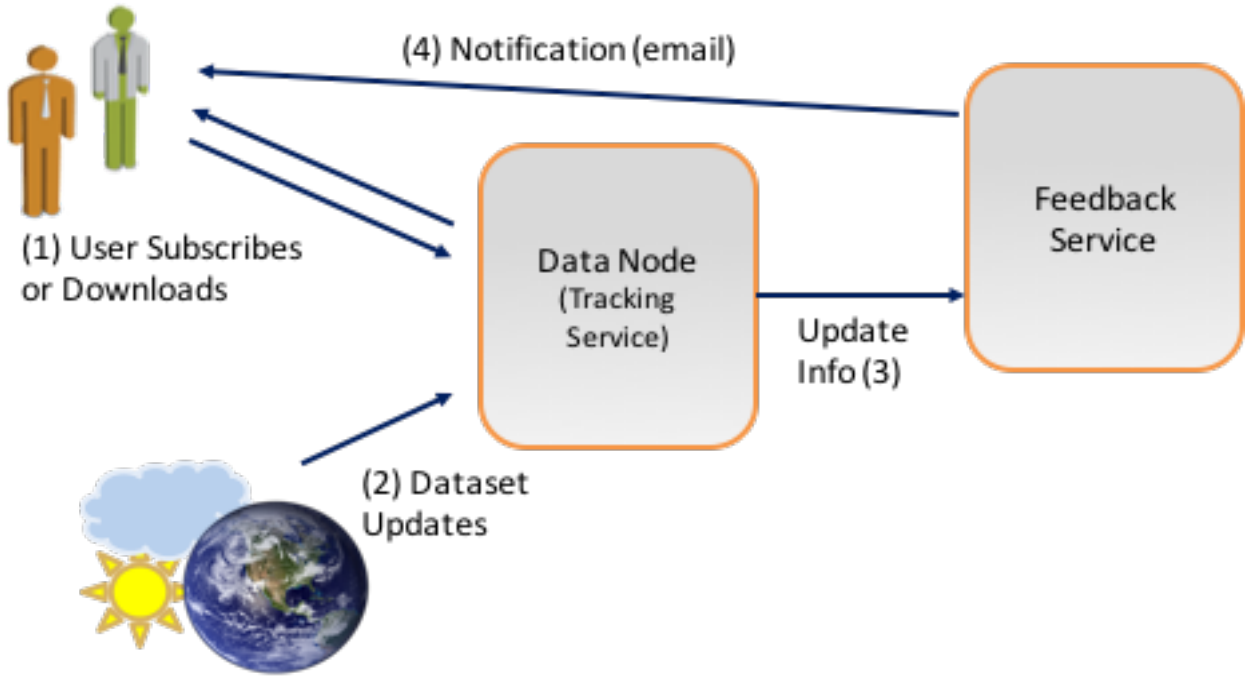


Figure 12. *Notification workflow from (1) user subscription or download to (2) dataset updates, after which the data node tracking service (3) sends updated information to the feedback service prior to (4) user notification via email.*

Progress Report

Heading into 2017, the node manager was slated for introduction in the 2.5.x versions of the ESGF software stack. The node manager successfully federated the registry among the nodes that deployed that version into production. To facilitate the need of the PID services to distribute information about the RabbitMQ servers and their respective credentials, a secure distribution API was implemented as part of the node manager endpoint service. This API has been tested by several sites.

A result of the node manager discussion at the ESGF F2F conference was that a node manager "daemon" that actively communicates with other sites is not needed to maintain a unified registry. Previously, the esgf-dashboard module made use of the registration.xml file, but no longer is this the case.

A feedback service to create user email notifications was created. For subscription management, we have devised the necessary database and had conducted feasibility studies into the use of the esg-search API to track changes to the published content within the federation.

The ESGF PID services have further evolved in 2017 to ensure the operational stability required to serve CMIP6 in particular.

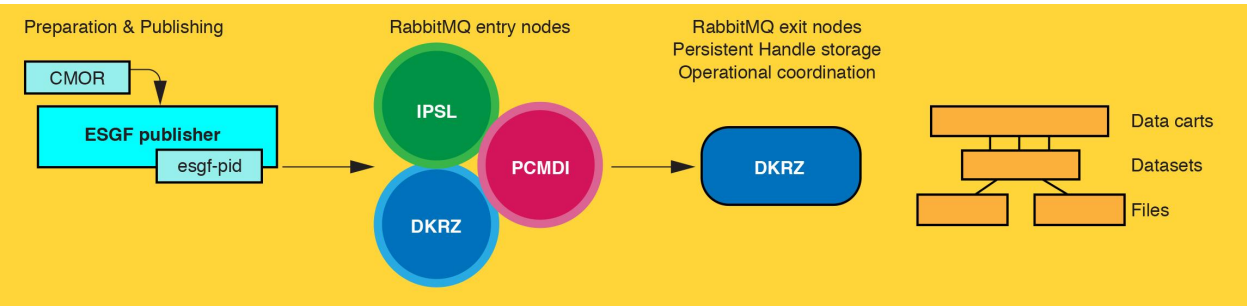


Figure 13. *Architecture of PID services showing workflow from publisher to RabbitMQ to consumer.*

The first step is actually covered by the Climate Model Output Rewriter (CMOR), or tools that follow the specifications, as it writes the tracking_id in the new formal "hdl:21.14100/<UUID>" into netCDF headers. These files then undergo the publishing process; as part of that, the publisher uses the esgf-pid Python library, which turns actions on files (e.g., publication, un-publication, replica location change) into RabbitMQ messages. The publisher also generates dataset PIDs based on hashes of data reference syntax (DRS) and version. RabbitMQ messages are received by currently one of three entry nodes at PCMDI, IPSL, and DKRZ. If a node is down, an alternative node is chosen. The nodes form a RabbitMQ federated exchange. Messages in this exchange are finally consumed by a continuous process at the federation exit node at DKRZ, where they are interpreted into concrete actions on Handles, such as registering Handles or changing values in Handle records.

The software components involved in this service are now stable and designed with multiple fallback options in mind. DKRZ has also set up monitoring for the exit nodes, including monitoring the number of messages on hold to detect scalability issues. During 2017, all channels in the RabbitMQ federation were also secured by transport layer security. The configuration of the federation was also changed to be able to accommodate projects other than CMIP6 in the future. In particular, Handle prefixes have been registered to serve CORDEX and Obs4MIPs needs. In addition to the PID management process, the PID services also include a web interface to mint PIDs for user-generated data carts, integrated into the ESGF portals.

Future Roadmap

Given the reformulation, the working team should choose a new name that encapsulates the various functions succinctly. During Q1 2018, the existing node manager API will be refactored into a "registry" API. The API should support client credential checking in the same manner as the secure credentials API for sensitive information such as software component versions.

Looking ahead to Q3, that information will be helpful for installation management federation-wide. Additionally, we propose that monitoring services be integrated into the auspices of this working team. Therefore, registry information should be included in monitoring. It will be advantageous to deploy monitoring at several sites for redundancy.

Implementation efforts on the subscription-based user notification services will continue during 2018. The major integration effort for this will be to connect a user's preference database with the search results indicating updates of published data. We believe it is well within our reach to test and deploy these services before the year's close.

For the PID services, the prefixes currently in operation are planned to be organizationally transferred to the European Data Infrastructure (EUDAT)/European Open Science Cloud (EOSC)–hub B2HANDLE service umbrella. This will improve options for Handle mirroring (backup and scalability) and open up further development opportunities. As part of the latter, Handle mass management tools may become available that may be designed within and for EOSC-hub Handle services, which can then also be used in the ESGF context to support testing and make manual error recovery easier.

Publication, Quality Control, Metadata, and Provenance Capture Working Team

Progress Report

The publisher migrated its Python deployment from the RPM-based Ultrascale Visualization–Climate Data Analysis Tools (UV-CDAT) environment to a condo-based deployment to be co-located with CDAT/CDMS (Climate Data Management System) and CMOR for integration with PrePARE. These changes have been deployed into production with the ESGF 2.5.x versions of the software stack. Much progress with esgprep has been made over the year. Fetch ini was simplified to allow for convenient download of all inis, the drs tool was released, and a refactored mapfile function greatly improved performance. The PrePARE and PID integrations have been tested and appear to work on test federation deployments (in test mode). For CMIP6 testing, IPSL and Geophysical Fluid Dynamics Laboratory (GFDL) test data were published to the test federation. CDF2CIM trigger was implemented into the publisher.

Additionally, the automated publishing test improved the ability to confirm that all services required for publisher operation on an ESGF deployment are operational. NCI re-published their entire data catalog, making use of esgprep following a directory structure reorganization. Much activity occurred in Input4MIPs publishing, including non–network common data form (netCDF) data, PSIPPS, and E3SM published new collections.

Pacific Northwest National Laboratory (PNNL) extended ProvEn with a HAPI. This library extracts existing file-based information produced by applications. HAPI's generic format can be used to harvest provenance from relational database tables as well as other scientific applications that log provenance-related information. HAPI uses community standards like W3C PROV to enrich the scruffy provenance for traceability and cross-comparison. One HAPI use case is the E3SM project: HAPI was used to reproduce a tiny E3SM simulation by recovering enough information from an original E3SM simulation.

Future Roadmap

As 2018 gets under way, the team will play an important role in the execution of the key steps in the Data Challenges for CMIP6. A major publication task will be the republication of Input4MIPs, scheduled for January. Plans are underway to refresh the egg-publisher software.

These include an update to Python 3.0, switch to the "requests" module for https connections, addition of parallelization of some processes, improvement of the logs, and complete the switch to the REST API (testing).

High-performance storage system (HPSS) publishing is an important action item, and a team at Lawrence Berkeley National Laboratory (LBNL) is aiding in that effort. It will be important to revisit the ingest service being developed at ANL. Missing CMIP5 replicas will continue to be republished at LLNL, and the issue of many replicas incorrectly being listed as latest will be addressed.

Revamping the CV integration into the publisher requires attention, and we plan to address this in three phases. (1) In the short term, we will simply project ini file generation. (2) In the mid-term, we will migrate to a CV tool that centralizes information, likely around existing the Working Group on Coupled Modelling (WGCM) CV repository and format. (3) In the long term, we will converge to a single CV service.

For provenance, the following activities are planned in 2018: capture provenance from ESGF components and the first component will be ESGF publication; develop the capability to compare simulations or application runs; integrate with other components within ESGF that use PROV standards; extend the ProvEn platform to support raw JSON data format and to annotate the JSON to store in triple store and time series store within ProvEn; and extend ProvEn with more interfaces for data transfer.

User Support and Documentation Working Team

Progress Report

While the team was able to maintain responsiveness to user support needs in 2017—thanks to many people pitching in—the consensus is that a dedicated support agent and a managerial task force would be beneficial for ensuring a higher level of support. We need to be fully prepared for CMIP6, when usage, and questions, will increase. Weekly team meetings are bringing these concerns to the fore, and progress is being made to explore solutions to a number of identified problems.

The past year saw fewer user emails to the support mailing list, prompting speculation that users are either not asking questions, the site is more usable, there are fewer users, or some combination of factors. For instance, when CMIP5 launched, initially we saw 20 to 25 questions per day; the volume has dropped to 1 to 2 questions per day, on average. Most questions from users are related to user permissions, downloading, or account issues.

Future Roadmap

A task force should be formed to identify and prioritize key action items for CMIP6 preparedness and, in the longer term, overall improvements in user support. Many potential solutions and ideas were proposed at the ESGF 2017 F2F conference, including establishing group mailing lists to assist with triage and identifying key personnel (subject matter experts) who can answer specific data questions. Other ideas include better error messaging or user prompts may reduce the number of questions; educating users to troubleshoot common problems themselves; and making FAQs more robust.

Another issue to resolve is management of user questions—specifically, how to collect questions and communicate resolution. Various applications such as Jira, Stack Overflow, Ask Bot, and

GitHub Issues were discussed, along with a potential homegrown solution of creating a "chat with a representative" window on CoG. The team needs to evaluate these options and consider ways to determine whether a user support strategy is effective.

Finally, documentation tends to be team specific and decentralized, lacking overarching guidelines or requirements. Commercial off-the-shelf solutions like Confluence/Jira are underutilized. The team plans to work on solutions that achieve (1) consistent, centralized location for documentation; (2) documentation tailored to audience (developers versus end users); (3) a reasonable maintenance/updating schedule; and (4) better organization/structure within documents.

Machine Learning Working Team

The Machine Learning Working Group was formed in 2017 to start a new activity in the ESGF context related to ML applied to climate change. A first session on ML was held at the ESGF 2017 F2F conference.

Progress Report

The activities reported here are mainly those presented at the ESGF 2017 F2F conference. In particular, we showed successful cases in ML application in analyzing extreme climate event including detection, classification, and tracking target events. In addition, a super-resolution technique has been applied to improve resolution of climate data. Three successful results suggest the potential application of DL for massive climate data and can provide tremendous benefit to climate community including saving human effort and computing the cost required for conventional data analysis methods.

1. For detection and classification, fully connected CNNs are applied. We showed more than 90% detection and classification accuracy for a tropical cyclone use case.
2. For tracking an extreme climate event, a multi-layered LSTM model has been applied for CAM5 reanalysis data. We showed successful tracking result for our use case of extratropical cyclone (ETC) given the initial position of cyclone location.
3. A pixel-recursive super resolution model, which is an auto-regressive model, has been applied to improve resolution of climate data. Given the condition that a tropical cyclone is present in an image, we succeeded in improving the resolution of the image by learning prior distribution of each pixel in the image. Improving resolution given pre-existing data has potential to compute the overhead required for a conventional simulation scheme heavily dependent on physical simulation principles.

Future Roadmap

During the next year, the working group will continue to pursue extreme event analysis using state-of-the-art DL techniques. Specifically, we will focus on time series climate data and potentially predict long-term climate patterns in massively scaled past data using recursive NNs. Additionally, the link with the CWT will be investigated to make compute engines also capable of integrating ML algorithms, thus extending the set of features from simple processing to more advanced knowledge discovery.

Diagnostics Working Team

This working group provides diagnostics software that works with data obtained by the ESGF platform. The goal of diagnostics software is to allow scientists to validate their model data.

Another objective is to bridge the gap between technology and groups of international scientists by providing new ways for data to be analyzed.

Progress Report

In 2017, we expanded on the Community Diagnostics Package (CDP). CDP is a framework for creating diagnostics packages and allowing such packages to interoperate with each other. Work began on a CDP-based diagnostics package, the E3SM Diagnostics. The goal of the E3SM Diagnostics is to provide diagnostics for the E3SM model, which uses the observational data from remote sensing, reanalysis, and in situ datasets. The initial version of this diagnostics package supports five different diagnostics plots and can compare model versus observations, model versus model, and observations versus observations.

Additionally, the PCMDI metrics package (PMP) was refactored to use CDP. This will allow for new features of CDP to be automatically added into PMP and allow PMP to work with other diagnostics packages. Support for portrait plots, and both diurnal and monsoon diagnostics, were added as well. The Atmospheric Radiation Measurement (ARM) Diagnostics package is used to facilitate the use of long-term, high-frequency measurements from the ARM program to evaluate regional climate simulations of clouds, radiation, and precipitation. This package was ported to use CDP with plans to interoperate with the other CDP-based packages.

Future Roadmap

For the upcoming year, common input argument conventions are planned for incorporation into CDP, so packages can interoperate with more ease. Regarding the E3SM Diagnostics package, we plan on incorporating more plot sets such as pressure-longitude maps, Taylor diagrams, and more. New datasets and variables will also be added. One of the first diagnostics package we plan to implement the aforementioned common input arguments into is PMP, work on which is scheduled to begin shortly. The ARM Diagnostics plans to provide some interoperation functionality with the E3SM Diagnostics package.

6. Implementation Roadmap

Following the presentations and discussions of the 2017 F2F meeting, the ESGF-XC and ESGF developers' community have converged on the following roadmap to guide future development and operations in the short, medium- and long-term future.

Short Term (1–2 Years)

The next 1–2 years will be critical for ESGF to prove to the international climate community that it remains the premier infrastructure for serving current and upcoming data collections, and for scaling into the future. Therefore, focus will be based on strengthening and improving current capabilities to publish, search, download and analyze PB-scale distributed data holdings, most prominently the upcoming CMIP6 model output. The following priorities have been identified—in roughly decreasing order of importance:

1. **Data and Services Challenges.** The ESGF top-most priority this year is to conduct the 2017 Data and Services Challenges, leading to a successful official opening of the ESGF system for publishing and access of CMIP6 data by June 1, 2018. Along the way, we intend to fix any problems, or upgrade existing software components, that are critical to the smooth operation of the system. This includes testing and certification of the new CMIP6 services such as PIDs, errata, model documentation, and DOIs.

2. **Simplify Data Access.** We are also committed to improving the overall user experience for the scientific community. Specific areas that should be addressed in the short term include removing access control restrictions for all possible data collections (starting with CMIP6, CMIP5 and Obs4MIPs), and enable easier tools for downloading large amounts of data.
3. **Node Installation.** Two efforts are currently underway to make it easier for administrators to install, upgrade and manage an ESGF Node, and they should both be completed by the end of 2018:
    a. Re-writing the standard ESGF installer as a Python library, known as the ESGF 2.0 installer. The new installer will setup all the ESGF modules in an architecture that is very close to the current one, but promises to be extremely more usable and maintainable, not to mention less error prone.
    b. Designing and implementing a next-generation ESFG architecture based on containerization, called ESGF/Docker. This architecture will leverage leading tools and frameworks such as Docker, Docker Swarm, Kubernetes, and OpenShift to enable flexible and scalable deployment of ESGF services onto clusters of resources that can be provisioned either in-premise or on the Cloud. On the other hand, this architecture will be a considerable departure from the way that ESGF services are configured and operated right now, so it will require a considerable learning curve on part of ESGF administrators and developers.
4. **Security Infrastructure Upgrade.** ESGF is in the process of upgrading its security infrastructure to replace the current, outdated OpenID 2.0 security protocol with the more current OAuth2 standard, which is commonly used throughout the commercial enterprise. This will involve replacing the current authentication servers (IdP and MyProxy) with the new REST-based SLCS server to issue X509 certificates and OAuth2 tokens, upgrade some authentication clients (CoG), and replace other clients (the ORP [OpenID relying party]) with new applications (the new ESGF-Auth2 client). Also, the plan includes simplifying the generation of wget scripts to improve usability.
5. **Remote Computation.** By the end of this year, we intend to finalize the release and integration of the ESGF Computing engine alongside the rest of the ESGF software stack, so that the scientific community can start executing some basic data operations (subsetting, re-gridding, averaging, and others) on the server side. The goal is to enable server-side computations at least for the major data centers that will be hosting CMIP6 data, in time to be useful for the community. ESGF will also define and execute a "certification" process for computing Nodes, to guarantee that the supported algorithms will be able to run accurately and efficiently on the Nodes that advertise this capability.
6. **User Support and Documentation.** We recognize that overall the documentation for using the ESGF system is not complete, sometimes out of date, and spread across multiple locations. In the next few years we must dedicate enough resources to remedy this situation, and commit ourselves to write, review and maintain top-class documentation for end users and Node administrators. We plan to convene one or more meetings to address this issue and to draft an action plan, then to allocate adequate resources for implementing this plan.

Medium Term (3–5 Years)

Looking beyond the most pressing needs of being able to successfully serve the current and upcoming data collections to the scientific community, ESGF needs to plan its longer-term strategy to be able to retain and expand its position as the premier global infrastructure for serving climate data. Following is a list of the major focus area to direct future ESGF efforts in the next 3–5 years:

1. **Scalability.** One of the most critical challenges for ESGF in the future will be the ability to scale its services to ever increasing data volumes, sparing the users from experiencing any degradation of service. Scalability will likely have to be achieved by a combination of two factors:
    a. *Scaling out the underlying computing resources.* ESGF services will need to be deployed on clusters of computing resources, either on premise or on the Cloud, and possibly configured for auto-scaling. Searching, data download, computation could all be scaled by load-balancing client requests across multiple servers.
    b. *Application refactoring.* The performance of many ESGF applications could be improved by re-designing them to be able to run multiple threads in a dynamic, distributed environment. For example, we already plan to base the next generation ESGF search on Solr Cloud. Data publishing could also be executed with a multi-threaded process, writing catalogs in parallel and acquiring locks for top-level catalogs when needed. Server-side computations could also be orchestrated over multiple nodes.
2. **Cloud.** As demand for computing and data resources increases, at least some ESGF sites are going to look at the commercial Cloud to host all or part of their services, even if perhaps only to burst some temporary computations. Using the Cloud will be easier, more scalable and efficient if ESGF services are deployed as Docker containers, orchestrated through one of the widely adopted frameworks such as Docker Swarm or Kubernetes. It will also be very important to work with Cloud providers to define a cost-effective model such that ESGF sites do not get charged for data transfer or computation that are executed by the users of the system.
3. **ML and DL.** We will develop limited supervised and unsupervised feature learning by implementing existing HPC-enabled deep feature learning algorithms for ESGF. These algorithms will allow us to automatically learn feature representations from massive amounts of unlabeled data and use these features as the basis for performing high-quality classification, clustering, prediction, and anomaly detection. At first, we will feature an API design based on the CWT WPS API that will allow users to run existing trained models on datasets of their choosing. Later, we will provide a capability to allow users to train new models on their own or through ESGF site-provided computation facilities and publish those models to be used by others. Finally, we will provide a framework, based on the CDP for users to incorporate additional ML or NN techniques for creating models.
4. **Interoperability.** Recognizing that effective scientific research needs to analyze data assets that are managed by multiple agencies and institutions, ESGF is fully committed to achieve an increasingly higher level of interoperability with all other major climate infrastructures around the world. As always, our strategy will be to foster interoperability based on common data and metadata standards, industry-wide protocols, and science-

specific APIs. We will start to hold discussions, gather requirements, design architectures and implement prototype services with some of the agencies that we are closer to – namely, NASA and the National Oceanic and Atmospheric Administration (NOAA). Our collaborative efforts should inspire other parties around the world to join in and evolve their infrastructure to conform to the same specifications, and possibly adopt the same implementations.

5. **Server-Side Computing.** Although much progress has already been achieved by the ESGF Computing Working Team, much remains to be done to enable a fully operational distributed computing environment where scientific algorithms are executed at multiple sites around the world and results are combined to spawn additional computations. Building on the existing server and client-side APIs, ESGF will continue to work to (1) enable full orchestration of workflows across the federation; (2) expand the set of basic algorithms that are supported and certified at all sites; and (3) possibly enable users to supply their own scientific algorithms to be run server-side across the world.
6. **UI Upgrade.** Our current UI (CoG) is more than 5 years old now, which in the IT world is quite a long time. We feel the need to upgrade this component to newer standards and look and feel, to make it more usable, efficient and scalable. To do so, we will work in strict collaboration with all the stakeholders involved (end users, Node administrators and program sponsors) and solicit continuous feedback from the community. As part of this plan, we will strip out, or make optional, some of the current functionality such as project governance, which is not required to administer data-oriented services.
7. **Extension to Other Domains.** Much of the ESGF infrastructure and services are not specific to climate science and could easily be translated to serve data from other scientific domains, such as biology, hydrology, and others. In particular, our effort to containerize all ESGF services and design a microservices architecture for science provides a blueprint that is widely applicable across the board by replacing climate-specific services with other domain-specific implementations. Some of the ESGF participants are already engaged, and will continue to pursue, efforts to generalize the ESGF infrastructure in this direction.

Long Term (5–10 Years)

In the long term, the vision for the ESGF system is completely aligned with the concept of "Virtual Laboratory" that DOE has been promoting over the past few years. That is, the goal is to evolve ESGF towards a web-based environment which fully integrates data from distributed geographic locations and science domains, making access to data and resources completely transparent to the user. This environment should enable users to visually explore the data, apply their own analysis, refine hypothesis, test them and export results in multiple formats, maintaining complete provenance information throughout the process. Human aspects of this future platform include facilitating collaboration between scientists at different institutions, targeting education, and supporting direct application to the commercial enterprise. To realize this vision, ESGF will have to retain and improve some of its key architectural characteristics such as adaptability, extensibility and scalability.

Appendices

A. Conference Agenda

<table>
<tr><th>Time</th><th>Topic</th></tr>
<tr><td colspan="2">Monday, December 4, 2017</td></tr>
<tr><td>2:00 p.m. – 4:00 p.m.</td><td>Pre-conference registration: Sheraton; Presidio Ballroom</td></tr>
<tr><td>5:00 p.m. – 6:00 p.m.</td><td>Social Activity: Meet and Greet (NO HOST)
Sheraton-Fisherman’s Wharf – Restaurant/Bar</td></tr>
<tr><td colspan="2">Tuesday, December 5, 2017</td></tr>
<tr><td>7:30 a.m. – 8:30 a.m.</td><td>Registration: Sheraton; Presidio Foyer</td></tr>
<tr><td>8:00 a.m. – 8:30 a.m.</td><td>Coffee/tea reception and meet & greet: Sheraton, Presidio Foyer</td></tr>
<tr><td>8:30 a.m. – 8:35 a.m.</td><td>DOE opening comments—Justin Hnilo, U.S. DOE’s Office of Biological and Environmental Research (BER) Program Manager for Data Management
● Includes welcome, safety, introduction, conference charge, and agenda overview</td></tr>
<tr><td>8:35 a.m. – 9:00 a.m.</td><td>State of the Earth System Grid Federation (ESGF)—Luca Cinquini (NASA/JPL)
● How conference attendees contribute to the conference’s final report (hand out last year’s 2016 6th Annual ESGF F2F Conference Report)
● Framing of the 2017 7th Annual ESGF F2F Conference</td></tr>
<tr><td colspan="2">Science Drivers: Project Requirements and Feedback</td></tr>
<tr><td>9:00 a.m. – 12:00 noon (3 hours)</td><td>Science Drivers
Session Discussion Lead — V. Balaji
<table><tr><td>9:00 a.m. – 9:30 a.m.</td><td>Karl Taylor and V. Balaji—Coupled Model Intercomparison Project, phase 6 (CMIP6) and the Working Group on Coupled Modeling Infrastructure Panel (WIP)</td></tr></table></td></tr>
</table>

<table>
<tr><th>Time</th><th>Topic</th></tr>
<tr><td></td><td>
<table>
<tr><td>9:35 a.m. – 10:05 a.m.</td><td>Peter Gleckler, Duane Waliser, Denis Nadeau, Robert Ferraro, Karl Taylor, Luca Cinquini, Paul Durack—Observations for Model Intercomparison Project (Obs4MIPs) from an ESGF Perspective: Progress, Plans, and Challenges</td></tr>
<tr><td>10:10 a.m. – 10:40 a.m.</td><td>Sébastien Denvil, Michael Lautenschlager, Sandro Fiore, Francesca Guglielmo, Martin Juckes, Stephan Kindermann, Michael Kolax Wim Som de Cerff—Copernicus and H2020 Programme</td></tr>
<tr><td>10:40 a.m. – 10:55 a.m.</td><td>Break</td></tr>
<tr><td>10:55 a.m. – 11:25 a.m.</td><td>Jerry Potter, Laura Carriere, Judy Hertz—Collaborative REAnalysis Technical Environment Intercomparison Project (CREATE-IP)</td></tr>
<tr><td>11:30 a.m. – 12:00 noon</td><td>Dean N. Williams, Dave Bader, Renata McCoy—Energy Exascale Earth System Model (E3SM) Workflow</td></tr>
</table>
Questions for presenters to answer during their presentations
● What are the key things that are difficult to do today and are impeding scientific progress or productivity and the sharing of data?
● What are key development effort that you see are needed for the future success of your projects?
● What is your timeline for data production and distribution from climate model and observations, HPC, 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 your project (e.g., metrics collection and reporting, persistent and DOIs, deriving data, user publication [i.e., long-tail publication])?
● What are your expected strategic roadmaps for the ESGF's short-term (1–3 years), mid-term (3–5 years), and long-term (5–10 years) development efforts?
● What are known use cases and workflows to help describe your ESGF future needs?
Homework assignment
● Before the conference adjourns, convert all known science drivers to use cases for ESGF development.
</td></tr>
<tr><td>12:00 noon – 1:30 p.m.</td><td>Lunch</td></tr>
</table>

<table>
<tr><th>Time</th><th>Topic</th></tr>
<tr><td>1:30 p.m. – 3:30 p.m. (2 hours)</td><td>Science Driver Town Hall Discussion
Session Discussion Lead — Ben Evans
Town Hall Panel: (Karl Taylor, V. Balaji, Peter Gleckler, Robert Ferraro, Sébastien Denvil, Michael Lautenschlager, Jerry Potter, Renata McCoy)

Questions to prepare for science driver presentation and discussion
What is working, and what is not working?
What are the key challenges to your programs concerning big data challenges?
What data services would address the identified challenges?
What exists already today?
What do we still need from ESGF?
What are the key characteristics that these services need to have to be successful (i.e., integrated, easy to customize)?
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?</td></tr>
<tr><td>3:30 p.m. – 3:45 p.m.</td><td>Awards Ceremony</td></tr>
<tr><td>3:45 p.m. – 4:00 p.m.</td><td>Break</td></tr>
<tr><td>4:00 p.m. – 6:00 p.m. (2 hours)</td><td>Poster and Live Demonstration Session
Session Discussion Lead — Ben Evans

<table>
<tr><th>No.</th><th>Title</th><th>Name</th><th>Poster</th><th>Demo</th></tr>
<tr><td>1</td><td>The Earth Data Analytics Services (EDAS) Framework</td><td>Thomas Maxwell
Dan Duffy</td><td>Yes</td><td>Yes</td></tr>
<tr><td>2</td><td>PAVICS: A platform for the Analysis and Visualization of Climate Science – toward inter-operable multidisciplinary workflows</td><td>D. Huard
T. Landry
D. Byrns
B. Gauvin-St-Denis</td><td>Yes</td><td>Yes</td></tr>
<tr><td>3</td><td>OGC Testbed-13 Earth Observation Clouds</td><td>T. Landry
D. Byrns</td><td>Yes</td><td>No</td></tr>
</table></td></tr>
</table>

Time	Topic				
	4	**Using the ESGF CWT API in the context of the EUDAT-EGI e-infrastructure and the ENES Climate4Impact platform**	Christian Pagé Xavier Pivan Asela Rajapakse Wim Som de Cerff Maarten Plieger Ernst de Vreede Alessandro Spinuso Lars Barring Antonio Cofino Alessandro d'Anca Sandro Fiore	Yes	Yes
	5	**Managing growth and complexity – technologies to meet the challenges of operating data, services and infrastructure at scale**	Phil Kershaw Jonathan Churchill Alan Iwi Bryan Lawrence Neil Massey Sam Pepler Matt Pritchard Matt Pryor Ag Stephens	Yes	No
	6	**Ophidia: an interoperable 'big data' framework for climate change analytics experiments**	Sandro Fiore Charles Doutriaux Cosimo Palazzo Alessandro d'Anca Zeshawn Shaheen Donatello Elia Jason Boutte Valentine Anantharaj Dean N. Williams Giovanni Aloisio	Yes	Yes

Time	Topic				
	7	**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	Yes	Yes
	8	**WPS based processing services for the Copernicus Climate Change Service (C3S)**	Stephan Kindermann Carsten Ehbrecht Ag Stephens Björn Brötz Wim Som de Cerff Maarten Plieger Sébastien Denvil	Yes	Yes
	9	**Diagnostics Package for the E3SM Model**	Chengzhu Zhang Zeshawn Shaheen Chris Golaz Jerry Potter	Yes	Yes
	10	**ESGF Errata Service**	Guillaume Levavasseur Atef Ben-nasser Mark A. Greenslade	No	Yes
	11	**DREAM Data Services for Biological Data and Beyond**	Sasha Ames Luca Cinquini Dean N. Williams	Yes	Yes

Time	Topic				
	12	**Community Data Analysis Tools**	Charles Doutriaux Denis Nadeau Dan Lipsa Dean N. Williams Aashish Chaudhary	Yes	Yes
	13	**Visual Community Data Analysis Tools (vCDAT)**	Matthew Harris Dan Lipsa James Crean Matthew Ma Charles Doutriaux Dean N. Williams Aashish Chaudhary	Yes	Yes
	14	**Integrating ES-DOC (Earth System Documentation) with the ESG Publisher**	Alan Iwi David Hassell Mark A. Greenslade Ag Stephens	Yes	Yes
	15	**Compute Working Team End-User Application Programming Interface**	Jason Boutte Charles Doutriaux	Yes	Yes
	16	**A compliance-checking framework for CMIP7**	Ag Stephens Antony Wilson Guillaume Levavasseur	No	Yes
	17	**Google Earth Engine and Project Jupyter**	Tyler Erickson	No	Yes
	18	**New Approach to Evaluate Large-scale Variability in CMIP models**	Ji-Woo Lee Kenneth R. Sperber Peter J. Gleckler	Yes	Yes

<table>
<tr><th>Time</th><th>Topic</th></tr>
<tr><td></td><td>
<table>
<tr><td></td><td></td><td>Celine W. Bonfils
Karl E. Taylor
Charles Doutriaux</td><td></td><td></td></tr>
</table>

Questions to address in your presentation and/or demonstration

What is working and what is not working?
What are the key challenges to your application concerning big data challenges within the ESGF infrastructure?
How does your application/services integrate into ESGF?
What do you still need from ESGF for software integration?
What are the key impediments in delivering your application/services in ESGF (i.e., installation, customization)?
What are the key characteristics or functionalities that your application/services offer the community within the ESGF infrastructure?
Which services or functions are your application’s highest development priorities, and what would be their measurable impact on science/programs (i.e., what is in store for the future)?

</td></tr>
<tr><td>6:00 p.m.</td><td>Adjourn Day 1</td></tr>
<tr><td colspan="2">Wednesday, December 6, 2017</td></tr>
<tr><td>8:00 a.m. – 8:30 a.m.</td><td>Coffee/tea reception and meet & greet: Sheraton; Presidio Foyer</td></tr>
<tr><td>8:30 a.m. – 9:45 a.m.
(1 hour & 15 min)</td><td>
Plenary Discussion: International Climate Network Working Group, Replication/Versioning and Data Transfer Working Team
Session Discussion Leads and Presenters — Eli Dart, Lukasz Lacinski, Stephan Kindermann
<table>
<tr><td>8:30 a.m. – 8:45 a.m.</td><td>Eli Dart, Lukasz Lacinski, Stephan Kindermann— Presentations on data transfers and replication progress (ESNet, Globus, DKRZ)</td></tr>
<tr><td>8:45 a.m. – 9:30 a.m.</td><td>Group discussion</td></tr>
<tr><td>9:30 a.m. – 9:45 a.m.</td><td>Conclusion recap</td></tr>
</table>
Questions for the ICNWG (i.e., network) plenary discussion

ICNWG network software and hardware integration requirements for Tier 1 and Tier 2 sites

</td></tr>
</table>

<table>
<tr><th>Time</th><th>Topic</th></tr>
<tr><td></td><td>

ICNWG network preparation services and tools (e.g., perfSONAR, Globus)
Automated replication network requirements for ESGF (i.e., CMIP6 and other projects)
ICNWG network security requirements
ICNWG dashboard integration into ESGF dashboard
Resource discovery and allocation services
Identify key gaps, identify benefiting communities, and prioritize ICNWG future work

</td></tr>
<tr><td>9:45 a.m. – 10:30 a.m.</td><td>Guest Presentation and Open Discussion: Google Cloud for Scientific Infrastructure
Session Presenter — Karan Bhatia (Google)</td></tr>
<tr><td>10:30 a.m. – 10:45 a.m.</td><td>Break</td></tr>
<tr><td>10:45 a.m. – 12:00 noon (1 hour & 15 minutes)</td><td>
Plenary Discussion: Compute and Data Analytics Working Team
Session Discussion Lead and Presenters — Charles Doutriaux and Daniel Duffy
<table>
<tr><td>10:45 a.m. – 11:00 a.m.</td><td>Charles Doutriaux and Daniel Duffy—Compute Working Team Update</td></tr>
<tr><td>11:00 a.m. – 11:15 a.m.</td><td>Cameron Christensen, Giorgio Scorzelli, Peer-Timo Bremer, Shusen Liu, Ji-Woo Lee, Brian Summa, Valerio Pascucci—ViSUS Streaming Analysis/Visualization</td></tr>
<tr><td>11:15 a.m. – 12:00 noon</td><td>Group Discussion and Conclusion recap</td></tr>
</table>
Questions for server-side computing

Define a scalable compute resource (clusters and HPCs) for ESGF data analysis
Data analytical and visualization capabilities and services
Performance of model execution
Advanced networks as easy-to-use community resources (i.e., resource management)
Provenance and workflow
Automation of steps for the computational work environment
Resource management, installation and customer support
Identify key gaps, identify benefiting communities, and prioritize next steps
Analysis services when multiple datasets are not co-located (future work)

</td></tr>
</table>

<table>
<tr><th>Time</th><th>Topic</th></tr>
<tr><td>12:00 noon – 1:30 p.m.</td><td>Lunch</td></tr>
<tr><td>1:30 p.m. – 2:40 p.m. (1 hour & 10 minutes)</td><td>Plenary Discussion: Identity, Entitlement, and Access (IdEA) Working Team
Session Discussion Lead — Philip Kershaw and Lukasz Lacinski

<table>
<tr><td>1:30 p.m. – 1:45 p.m.</td><td>Philip Kershaw and Lukasz Lacinski—Presentation on authentication and authorization and IdEA progress</td></tr>
<tr><td>1:45 p.m. – 2:30 p.m.</td><td>Group discussion</td></tr>
<tr><td>2:30 p.m. – 2:40 p.m.</td><td>Conclusion recap</td></tr>
</table>

Questions for authentication and authorization
● What tools have been identified for authentication and authorization (i.e., OAuth2) and how well will they integrate with other projects (i.e., Copernicus, NASA distributed active archive centers)?
● What is needed for authentication and authorization integration with the ESGF software stack installation (i.e., address key needs)?
● What services must be made available today and in the future for authentication and authorization?
● What level of support would be expected from the science community?
● How do we want to assess the maturity and capability of authentication and authorization (e.g., benchmarks or crowdsourcing)?
● What are the future efforts to be expected from ESGF-IdEA?</td></tr>
<tr><td>2:40 p.m. – 3:55 p.m. (1 hour & 15 minutes)</td><td>Plenary Discussion: User Interface, Search, and Dashboard Working Teams
Session Discussion Lead — Luca Cinquini, Guillaume Levavasseur, and Alessandra Nuzzo

<table>
<tr><td>2:40 p.m. – 2:55 p.m.</td><td>Luca Cinquini, Guillaume Levavasseur, and Alessandra Nuzzo—UI, Search, and Dashboard Working Group</td></tr>
<tr><td>2:55 p.m. – 3:30 p.m.</td><td>Group discussion</td></tr>
<tr><td>3:30 p.m. – 3:45 p.m.</td><td>Break</td></tr>
<tr><td>3:45 p.m. – 3:55 p.m.</td><td>Conclusion recap</td></tr>
</table></td></tr>
</table>

<table>
<tr><th>Time</th><th>Topic</th></tr>
<tr><td></td><td>Questions
Do you have any plan for engaging the user community to provide ongoing feedback for the UI?
How do you expect the search services to scale to support new data holdings in the next 5 years?
Do you have any plans for federating the search services with other sites/agencies/institutions?
How do you validate the metrics obtained from the dashboard, both for a single node, and across the whole federation?</td></tr>
<tr><td>3:55 p.m. – 5:30 p.m. (1 hour & 15 minutes)</td><td>Plenary Discussion: Installation and Software Security Working Team
Session Discussion Leads — William Hill, Prashanth Dwarakanath, Luca Cinquini, and George Rumney<table><tr><td>3:55 p.m. – 4:05 p.m.</td><td>William Hill and Prashanth Dwarakanath—Software Installation Working Team (IWT)</td></tr><tr><td>4:05 p.m. – 4:15 p.m.</td><td>Luca Cinquini—Software Container Architecture (i.e., Docker)</td></tr><tr><td>4:15 p.m. – 4:25 p.m.</td><td>George Rumney—Software Security Working Team (SSWT)</td></tr><tr><td>4:25 p.m. – 5:20 p.m.</td><td>Group discussion</td></tr><tr><td>5:20 p.m. – 5:30 p.m.</td><td>Conclusion recap</td></tr></table>Questions
How close are you to have an operation version of the Docker/Cloud ESGF?
Which services or functionality are still missing from this architecture?
How do you plan to address security risks with this architecture?
Is there a plan for migrating an operational system from the current shell-based installer to Docker/Cloud?</td></tr>
<tr><td>5:30 p.m. – 6:00 p.m. (30 minutes)</td><td>Open Discussion: Long-Term Future of ESGF</td></tr>
<tr><td>6:00 p.m.</td><td>Adjourn Day 2</td></tr>
<tr><th colspan="2">Thursday, December 7, 2017</th></tr>
<tr><td>8:00 a.m. – 8:30 a.m.</td><td>Coffee/tea reception and meet & greet: Sheraton; Presidio Foyer</td></tr>
</table>

<table>
<tr><th>Time</th><th>Topic</th></tr>
<tr><td>8:30 a.m. – 9:45 a.m. (1 hour & 15 minutes)</td><td>
Plenary Discussion: Publication, Quality Control, Metadata, and Provenance Capture Working Team
Session Discussion Leads — Sasha Ames and Heinz-Dieter Hollweg
<table>
<tr><td>8:30 a.m. – 8:40 a.m.</td><td>Sasha Ames—Publication Working Team</td></tr>
<tr><td>8:40 a.m. – 8:50 a.m.</td><td>Heinz-Dieter Hollweg—Quality Control Progress</td></tr>
<tr><td>8:50 a.m. – 9:00 a.m.</td><td>Bibi Raju—Provenance Data Harvest and Scientific Results Reproducibility</td></tr>
<tr><td>9:00 a.m. – 9:10 a.m.</td><td>Stephan Kindermann—Data Citation Service</td></tr>
<tr><td>9:10 a.m. – 9:45 a.m.</td><td>Group discussion and conclusion recap</td></tr>
</table>
Questions for publications, quality control, metadata, and provenance capture plenary discussion

Data integration and advanced metadata capabilities
Data and metadata collection and sharing capabilities for possible provenance
Data Quality and ancillary information
Data preparation services and tools
Authentication and security
Local and remote publication services
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 ESGF node 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?

</td></tr>
<tr><td>9:45 a.m. – 10:45 a.m. (1 hour)</td><td>
Plenary Discussion: Machine Learning
Session Discussion Lead — Sookyung Kim
<table>
<tr><td>9:45 a.m. – 9:55 a.m.</td><td>Sookyung Kim—Deep Learning Application for Community Machine Learning</td></tr>
<tr><td>9:55 a.m. – 10:05 a.m.</td><td>Sébastien Denvil, Sandro Fiore, Philip Kershaw—Copernicus and H2020 Programme Machine Learning and Big Data Needs</td></tr>
</table>
</td></tr>
</table>

<table>
<tr><th>Time</th><th>Topic</th></tr>
<tr><td></td><td>
<table>
<tr><td>10:05 a.m. – 10:15 a.m.</td><td>Tom Landry—Imagery, Text, and Geospatial Machine Learning
Applications in Montreal’s Booming ML landscape</td></tr>
<tr><td>10:15 a.m. – 10:45 a.m.</td><td>Group Discussion</td></tr>
</table>
Questions for the ML plenary discussion

What problems can ML and DL methodologies solve in climate domain?
What can these technologies not solve?
What is the recent metrics in DL which can applied to climate data?
What exist already in climate community using artificial intelligence?
What is the highest priority problem using ML in climate community?
What are the key challenges to ESGF implementing ML algorithms?
How can we solve these challenges with respect to programs?
How can we solve data labeling and scalability issue?

</td></tr>
<tr><td>10:45 a.m. – 11:00 a.m.</td><td>Break</td></tr>
<tr><td>11:00 a.m. – 12:00 noon (1 hour)</td><td>
Plenary Discussion: Diagnostics
Session Discussion Lead — Zeshawn Shaheen, Tom Landry, others
<table>
<tr><td>11:00 a.m. – 11:10 a.m.</td><td>Zeshawn Shaheen—Community Diagnostics Package (CDP)</td></tr>
<tr><td>11:10 a.m. – 11:20 a.m.</td><td>Sébastien Denvil—Copernicus and H2020 Programme Diagnostic Needs and Overview</td></tr>
<tr><td>11:20 a.m. – 11:30 a.m.</td><td>Tom Landry—Canada Diagnostics</td></tr>
<tr><td>11:30 a.m. – 11:50 a.m.</td><td>Group discussion</td></tr>
<tr><td>11:50 a.m. – 12:00 noon</td><td>Conclusion recap</td></tr>
</table>
Questions for the diagnostics plenary discussion

What are the key diagnostics challenges that scientists encounter?
What diagnostics capabilities would address the identified challenges?
What diagnostics exists already today?
What diagnostics are still need?
What are the diagnostics impediments for resource providers (i.e., hardware) and software developers to provide these missing capabilities?

</td></tr>
</table>

<table>
<tr><th>Time</th><th>Topic</th></tr>
<tr><td></td><td>

- Which diagnostics requirements need to be addressed with the highest priority and what would be their measurable impact on science?

</td></tr>
<tr><td>12:00 noon – 1:30 p.m.</td><td>Lunch</td></tr>
<tr><td>1:30 p.m. – 3:00 p.m. (1 hour & 30 minutes)</td><td>

Plenary Discussion: CMIP6 Data Node Operations Team (CDNOT)
Session Discussion Lead — Sébastien Denvil

<table>
<tr><td>1:30 p.m. – 1:50 p.m.</td><td>Sébastien Denvil—CDNOT Overview and Goals</td></tr>
<tr><td>1:50 p.m. – 2:10 p.m.</td><td>Katharina Berger—CDNOT Developer Perspective</td></tr>
<tr><td>2:10 p.m. – 2:30 p.m.</td><td>Sergei Nikonov—Case Study: CMIP6 and ESGF Reciprocation</td></tr>
<tr><td>2:30 p.m. – 2:50 p.m.</td><td>Kim Serradell—Case Study: EC-Earth and ESGF</td></tr>
<tr><td>2:50 p.m. – 3:00 p.m.</td><td>Group discussion and conclusion recap</td></tr>
</table>

Questions for the CDNOT plenary discussion

- What are the ESGF services and tools that are needed for CDNOT to be successful?
- Should CDNOT’s mode of operation be made more widely accessible to other projects and the community?
- What is the distinction between CDNOT and ESGF?

</td></tr>
<tr><td>3:00 p.m. – 4:30 p.m. (1 hour & 30 minutes)</td><td>

Plenary Discussion: Node Manager and Tracking/Feedback Notification
Session Discussion Lead — Sasha Ames and Tobias Weigel

<table>
<tr><td>3:00 p.m. – 3:10 p.m.</td><td>Sasha Ames—Node Manager and Services Working Team (NWT)</td></tr>
<tr><td>3:10 p.m. – 3:20 p.m.</td><td>Tobias Weigel— PID Services and Tracking/Feedback</td></tr>
<tr><td>3:20 p.m. – 4:05 p.m.</td><td>Group discussion</td></tr>
<tr><td>4:05 p.m. – 4:20 p.m.</td><td>Break</td></tr>
<tr><td>4:20 p.m. – 4:30 p.m.</td><td>Conclusion recap</td></tr>
</table>

</td></tr>
</table>

<table>
<tr><th>Time</th><th>Topic</th></tr>
<tr><td></td><td>Questions for the node manager and notification plenary discussion
What are the key challenges for the node manager and notification?
What 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)?
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?</td></tr>
<tr><td>4:30 p.m. – 5:30 p.m. (1 hour)</td><td>Open Discussion: User Support and Documentation
Session Discussion Lead — Matthew Harris

<table><tr><td>4:30 p.m. – 4:50 p.m.</td><td>User support group discussion</td></tr><tr><td>4:50 p.m. – 5:20 p.m.</td><td>Documentation group discussion</td></tr><tr><td>5:20 p.m. – 5:30 p.m.</td><td>Conclusion recap</td></tr></table>Questions
What level of support and documentation are needed for ESGF services, tools and the community?
What support and documentation do data provider and users want to see from ESGF?
What type of support and documentation is there for ESGF (i.e., FAQs, Jupyter Notebook, online tutorials, presentations)?
Where are the support tools and documentation located?
What can we expect in the future in terms for user support and documentation?</td></tr>
<tr><td>5:30 p.m. – 6:00 p.m. (30 minutes)</td><td>Open Discussion: Meeting Outcomes
Session Discussion Lead — Ben Evans</td></tr>
<tr><td>6:00 p.m. – 6:30 p.m. (30 minutes)</td><td>Breakout Session: Conference Report Planning</td></tr>
<tr><td>6:30 p.m.</td><td>Adjourn Day 3</td></tr>
<tr><th colspan="2">Friday, December 8, 2017</th></tr>
<tr><td>8:00 a.m. – 8:30 a.m.</td><td>Coffee/tea reception and meet & greet: Sheraton; Presidio Foyer</td></tr>
</table>

Time	Topic
8:30 a.m. – 10:00 a.m.	ESGF Executive Committee Breakout Meeting: Sheraton; Lombard Room ● Discuss of the construction of the annual conference report ● Discuss meeting location and time of the next ESGF F2F conference ● Discuss strategic and implementation documents Working Teams Meeting ● All working teams discuss conference findings for their area of annual reporting
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 — Tom Landry* ● ESGF Team Leads findings on conference feedback ● Prioritize the feedback ● Open discussion
12:00 noon	**Adjourn Day 4**
12:00 noon – 1:30 p.m.	**Lunch**
1:30 p.m. – 5:30 p.m.	General Code Sprint (optional): Sheraton; Lombard Room ● Working Teams and Leads
5:30 p.m.	**Conference Adjourn Day 4**
Concludes the 7th Annual ESGF F2F Conference	

B. Conference Abstracts

Day 1: Tuesday, December 5, 2017

ESGF Steering Committee and Executive Committee

Department of Energy Office of Biological and Environmental Research Data Management
Justin Hnilo (DOE/BER), Justin.Hnilo@science.doe.gov

The ESGF multi-agency, international software infrastructure has become critical to understanding climate change. Effectively managing the vast volumes of resulting simulation and observation data has become a major challenge for the climate and computational scientists who support climate projections. To manage the massively distributed data volumes, the ESGF connects diverse federated archives from over 21 countries for knowledge discovery. These distributed data archives have aided many DOE researchers in producing significant articles and reports, such as those contributing to the IPCC's Third and Fifth Assessment Reports. Today, the

ESGF infrastructure houses PBs of data generated by DOE projects, such as the E3SM and the international CMIP, and securely makes these data available to scientists and nonscientists.

In addition, the infrastructure provides data access services for DOE's broad community by conforming to DOE data standards. Data to be ingested, stored, maintained, and served by the infrastructure include DOE observational, experimental, and model-generated information and associated metadata, plus the tools and models directly associated with data generation, value-added products, simple analysis, display, and data serving. Thus, the access the ESGF provides has translated into an impressive volume of new DOE research. Over the next three years, it is estimated that the ESGF distributed archive will grow to tens of PBs of data storage and bridge the critical gaps between many DOE projects concerning big data issues.

The State of the Earth System Grid Federation
Luca Cinquini (NASA/JPL), luca.cinquini@jpl.nasa.gov

The ESGF is a multi-institutional, international software infrastructure and development collaboration led by scientists and software engineers worldwide. The ESGF's mission is to facilitate scientific research and discovery on a global scale. The ESGF architecture federates a geographically distributed network of climate modeling and data centers that are independently administered yet united by common protocols and APIs. The cornerstone of its interoperability is peer-to-peer messaging, which continuously exchanges information among all nodes through a shared, secure architecture for search and discovery. The ESGF integrates popular open-source application engines with custom components for data publishing, searching, UI, security, metrics, and messaging to provide PBs of geophysical data to roughly 25,000 users from over 1,400 sites on six continents. It contains output from the CMIP, used by authors of the IPCC's Third, Fifth, and Sixth Assessment Reports, and output from DOE's E3SM project and the EU's Copernicus Programme.

Over the next three years, we propose to:

1. sustain and enhance a resilient data infrastructure with friendlier tools for the expanding global scientific community; and
2. prototype new tools that fill important capability gaps in scientific data archiving, access, and analysis.

These goals will support a data-sharing ecosystem and, ultimately, provide predictive understanding of couplings and feedbacks among natural-system and anthropogenic processes across a wide range of geophysical spatial scales.

Science Drivers: Project Requirements and Feedback

Coupled Model Intercomparison Project, Phase 6 (CMIP6) and the Working Group on Coupled Modeling Infrastructure Panel (WIP)
Karl Taylor (DOE/LLNL/PCMDI), taylor13@llnl.gov
V. Balaji (NOAA/GFDL & Princeton University), balaji@princeton.edu

The World Climate Research Programme (WCRP) WGCM Infrastructure Panel, referred to as the "WIP," was established to provide clear guidance to ESGF and other projects supporting CMIP6 as to infrastructure needs from the perspective of climate modelling centers and the end users. The WIP is responsible for oversight of the CMIP6 "data request" and establishing metadata requirements and CVs that make it possible to automate management, access, and

interaction with the data archive. The WIP also considers the dependencies among various services built to support CMIP6 and guides their development so that they interact smoothly. It also attempts to encourage development of data standards and metadata specifications for closely related projects (e.g., Input4MIPs, Obs4mips) so that ESGF can provide a more uniform interface to the data produced by them. Following a summary of the current status of CMIP6 and the infrastructure supporting it, we shall identify high-priority needs or concerns regarding ESGF's critical contributions to WCRP activities.

Observations for Model Intercomparison Project (Obs4MIPs) from an ESGF Perspective: Progress, Plans, and Challenges

Peter Gleckler (DOE/LLNL/PCMDI), gleckler1@llnl.gov
Duane Waliser (NASA/JPL), duane.waliser@jpl.nasa.gov
Denis Nadeau (DOE/LLNL/AIMS), nadeau1@llnl.gov
Robert Ferraro (NASA/JPL), robert.d.ferraro@jpl.nasa.gov
Karl Taylor (DOE/LLNL/PCMDI), taylor13@llnl.gov
Luca Cinquini (NASA/JPL), Luca.Cinquini@jpl.nasa.gov
Paul Durack (DOE/LLNL/PCMDI), durack1@llnl.gov

During the last year, substantial effort has been devoted to coordinating the use of CMOR3 in CMIP6 withOobs4MIPs. This has included further alignment of the Obs4MIPs data specifications with CMIP6. Recently, these metadata specifications have largely been finalized, opening up the potential to include a next generation of Obs4MIPs datasets with more enhanced searching capabilities available via the ESGF. Two other recent ESGF-related advancements will be discussed: (1) the inclusion of dataset specific information in the form of a "suitability matrix," and (2) the ability for data providers to include supplemental data and metadata along with their best-estimate contribution to Obs4MIPs. After summarizing this progress, this presentation will be describing how Obs4MIPs can be further advanced via new ESGF capabilities.

Copernicus and H2020 Programme

Sébastien Denvil (ENES/IPSL), sebastien.denvil@ipsl.jussieu.fr
Michael Lautenschlager (DKRZ), lautenschlager@dkrz.de
Sandro Fiore (CMCC), sandro.fiore@cmcc.it
Francesca Guglielmo (ENES/IPSL), francesca.guglielmo@lsce.ipsl.fr
Martin Juckes (CEDA), martin.juckes@stfc.ac.uk
Stephan Kindermann (DKRZ), kindermann@dkrz.de
Michael Kolax (SMHI), Michael.Kolax@smhi.se
Wim Som de Cerff (KNMI), wim.som.de.cerff@knmi.nl

European Network for Earth System Modelling (ENES) partners are involved in numbers of projects funded by either the Horizon 2020 (H2020) Programme or the Copernicus Programme. Some of those projects will contribute to pieces of the development of the ESGF or will use ESGF results. This talk will introduce both H2020 and the Copernicus Programme and will highlight the major contribution to ESGF activities that are expected by the ENES Data Task Force from currently running projects.

Horizon 2020 is the biggest-ever EU research and innovation Programme. Almost €80 billion of funding is available over seven years (2014–2020) in addition to the private and national public investment that this money will attract. The goal is to ensure Europe produces world-class

science and technology, removes barriers to innovation, and makes it easier for the public and private sectors to work together in delivering solutions to big challenges facing our society.

Within H2020, three types of activities will be supported to make world-class research infrastructures accessible to all researchers in Europe and to fully exploit these resources' potential for scientific advancement and innovation:

- The first activities are targeted to the development of new world-class research infrastructures. Support will be provided for the implementation and operation of the research infrastructures listed on the European Strategy Forum on Research Infrastructures (ESFRI) roadmap. Support will cover the preparatory phase of new ESFRI projects and the implementation and operation phases of prioritized ESFRI projects. Further world-class facilities will also be part of this action.
- The second set of activities aims at optimizing the use of national facilities by integrating them into networks and opening their doors to all European researchers. This is a continuity of the so-called Integrating Activities under the Seventh Framework Programme.
- The third action will support further deployment and development of information and communication technologies–based e-infrastructures which are essential to enable access to distant resources, remote collaboration, and massive data processing in all scientific fields.

Copernicus has been specifically designed to meet user requirements. Through satellite-based, in situ observations and simulations, the services deliver data at a global level which can also be used for local and regional needs. This is essential to help us better understand our planet as well as sustainably manage the environment we live in. Copernicus is served by a set of dedicated satellites (the Sentinel families) and contributing missions (existing commercial and public satellites). The Sentinel satellites are specifically designed to meet the needs of the Copernicus services and their users. Since the launch of Sentinel-1A in 2014, the EU plans to place a constellation of almost 20 more satellites in orbit before 2030.

The main users of Copernicus services are policymakers and public authorities who need the information to develop environmental legislation and policies or to make critical decisions in the event of an emergency, such as a natural disaster or a humanitarian crisis. Based on the Copernicus services and on the data collected through the Sentinels and the contributing missions, many value-added services can be tailored to specific public or commercial needs, resulting in new business opportunities.

These value-adding activities are streamlined through six thematic streams of Copernicus services. One of them is the Copernicus Climate Change Service (C3S).

Collaborative REAnalysis Technical Environment Intercomparison Project (CREATE-IP)
Jerry Potter (NASA/GSFC), gerald.potter@nasa.gov
Laura Carriere (NASA/GSFC), laura.carriere@nasa.gov
Judy Hertz (NASA/GSFC) judy.hertz@nasa.gov

In light of recent extreme weather events, it has become increasingly evident that quick and easy access to multiple up-to-date modern reanalysis products is useful to a variety of researchers. The CREATE service offers reanalysis data repackaged in a form that is easily accessible

through the ESGF. The data adhere to the standard CMIP metadata requirements, and the datasets are extended every two months through automation. In addition to monthly and six-hour data for selected variables, CREATE provides five of the major reanalyses regridded to a standard grid, along with the ensemble average and standard deviation. These data were processed using the Earth Data Analytics Services (EDAS), a high-performance big data analytics framework developed at the NASA Center for Climate Simulation (NCCS) for the ESGF, and the data are provided at six-hour and monthly intervals with a common set of vertical levels.

CREATE data are also made available through CREATE-V, a web tool that leverages the common data format to provide visualization and comparison features and utilizes EDAS to display plots of monthly anomalies and yearly cycles at a user-selected location.

Using multiple reanalyses, NASA will use the CREATE service and products to explore examples of the recent droughts, floods, and hurricanes and study longer term climate trends.

Energy Exascale Earth System Model (E3SM) Workflow

Dean N. Williams (LLNL/AIMS/E3SM), williams13@llnl.gov
Valentine Anantharaj (ORNL/E3SM), anantharajvg@ornl.gov
Dave Bader (LLNL/E3SM), bader2@llnl.gov
Renata McCoy (LLNL/AIMS/E3SM), mccoy20@llnl.gov

The advanced model development, testing, and execution infrastructure has been designed to strongly accelerate the model development and testing cycle for the new DOE E3SM model by automating labor-intensive tasks, providing intelligent support for complex tasks, and reducing duplication of effort through collaborative Workflow Group support. The Workflow Group has two important assignments: (1) advance model development by developing, testing, and executing an end-to-end infrastructure that automates labor-intensive tasks; and (2) provide intelligent support for complex tasks in model development through scientific model component (i.e., atmosphere, land, ocean, sea ice) collaboration.

To achieve our primary objectives, the team was split into several epic subtasks: (1) E3SM Workbench and Process Flow, (2) Data Management, (3) Analysis and Visualization, (4) Diagnostics, (5) Provenance Capture, and (6) Hardware Infrastructure. These open-source projects have grown in scope as requirements have shifted or completely changed over the course of the project. The tools and experience resulting from their development provides the foundation on which the end-to-end model TB infrastructure will be based. As the global view of the E3SM project expands across the component model space, the usefulness and urgency of the workflow software becomes more apparent. The end goal of every quarter for the Workflow Group is to advance a step closer to reducing the level of effort to successfully run the E3SM model, archive output, generate diagnostics, and share the results of both the model output and diagnostics results with E3SM colleagues.

Poster and Live Demonstration Session

The Earth Data Analytics Services (EDAS) Framework

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

Faced with unprecedented growth in Earth data volume and demand, NASA has developed the EDAS framework, a high-performance big data analytics framework built on Apache Spark. This

framework enables scientists to execute data processing workflows combining common analysis operations close to the massive data stores at NASA. The data are accessed in standard formats (e.g., netCDF, hierarchical data format [HDF]) in a Portable Operating System Interface (POSIX) file system and processed using vetted Earth data analysis tools (e.g., Earth System Modeling Framework, CDAT, netCDF operators [NCO]). EDAS utilizes a dynamic caching architecture, a custom distributed array framework, and a streaming parallel in-memory workflow for efficiently processing huge datasets within limited memory spaces with interactive response times.

EDAS services are accessed via a WPS API being developed in collaboration with the CWT to support server-side analytics for 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. New analytic operations can be developed in Python, Java, or Scala (with support for other languages planned). Client packages in Python, Java/Scala, or JavaScript contain everything needed to build and submit EDAS requests.

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

PAVICS: A platform for the Analysis and Visualization of Climate Science – Toward Inter-Operable Multidisciplinary Workflows

D. Huard (Ouranos), huard.david@ouranos.ca
T. Landry (CRIM), tom.landry@crim.ca
D. Byrns (CRIM), David.Byrns@crim.ca
B. Gauvin-St-Denis (Ouranos), Gauvin-St-Denis.Blaise@ouranos.ca

Climate services comprise the necessary data and expertise to describe current and future climate conditions and their potential impact on human and environmental systems. Climate services are by their nature interdisciplinary, and an important bottleneck in delivering relevant and timely climate services lies at the interface between disciplines; differences in jargon, conventions, data formats, and programming languages act as barriers to effective collaborations. Here we describe how a scientific model—one in which researchers publish not only papers and data but also "expertise" in the form of online interoperable services—has the potential to drastically reduce the friction across disciplines. This scientific model could be exploited by scientific gateways such as the Power Analytics and Visualization for Climate Science (PAVICS) to widen the scope and relevance of climate services.

The PAVICS platform is built from modular components that target the various requirements of climate data analysis. The data components host and catalog netCDF data as well as geographical information layers. The analysis and processing components are made available as atomic operations through WPS, which can be combined into workflows and executed on a distributed network of heterogeneous computing resources. The visualization components range from Open Geospatial Consortium (OGC) standards (e.g., web map service [WMS], web coverage service [WCS], web feature service) to a complete front-end for searching the data, launching workflows, and interacting with maps of the results. Each component can easily be deployed and executed as an independent service using Docker. Permissions on data and

processes are managed via a RESTful interface and are enforced systematically with a token-based service. PAVICS includes various components from Birdhouse, a collection of WPS developed by the DKRZ and IPSL. Further connectivity is made with the ESGF nodes, and local results are made searchable using the same API terminology.

PAVICS: A Platform for the Analysis and Visualization of Climate Science

D. Huard (Ouranos), huard.david@ouranos.ca
D. Byrns (CRIM), David.Byrns@crim.ca
T. Landry (CRIM), tom.landry@crim.ca

The PAVICS project aims to provide climate scientists and climate service providers with tools that simplify the creation of value-added products from raw climate datasets. It includes a TDS, a search engine that connects with the ESGF database, a GeoServer instance to supply geographical layers, a diverse set of analytical services and a workflow engine to chain them together, along with a graphical workspace interface that can overlay geographic information system (GIS) layers and netCDF gridded datasets. The project relies heavily on projects orbiting the ESGF landscape, namely Birdhouse, OpenClimate GIS, and Indice Calculation CLIMate.

This demonstration shows how users create a project, define a data ensemble, subset the data over multiple geographical regions, compute climate indices, and create graphics displaying the results without any knowledge of netCDF. The demonstration will also cover how administrators manage services and permissions, as well as upload additional netCDF data and GIS layers via OGC standards. We will show JavaScript-based UI components pertaining to experience management, search, thematic layer management, web mapping, scientific workflows, and user workspaces. We will also discuss how the services that are planned by the ESGF CWT could be integrated in the platform.

OGC Testbed-13 Earth Observation Clouds

T. Landry (CRIM), tom.landry@crim.ca
D. Byrns (CRIM), David.Byrns@crim.ca

On a yearly basis, EO satellites already generate PBs of raw data. Resources required to process and store that data are quickly increasing due to higher resolutions, larger number of bands, and growing satellites and constellation count. The cloud computing landscape is well suited to cover most requirements of EO data and its applications. OGC's thirteenth testbed initiative (TB-13) aims to clarify cloud API interoperability and application portability as key elements in cloud computing research. The goal of participants of the EO Clouds (EOC) thread of TB-13 is to develop an integrated solution compatible with the ESA's Thematic Exploitation Platforms and the Canadian Forestry Service's operation, which is part of Natural Resources Canada (NRCan).

Centre de Recherche Informatique de Montréal (CRIM) is mandated by OGC to deliver a cloud-enabled application that extracts forest features or biophysical parameters from Radarsat-2 Synthetic Aperture Radar (SAR) to estimate forest biomass in Canada, in accordance with NRCan's specifications. CRIM implemented the SAR decompositions using ESA's Sentinel Application Platform graph processing tool and packaged the application using Docker, in an OpenStack environment. In that implementation, WCS and WMS capabilities were provided by GeoServer 2.11, while WPS requests were served by PyWPS 4. Participants in TB-13 are required to participate in the elaboration of an Engineering Report to be presented to designated OGC Working Groups. To address sponsors' requirements for the EOC thread, CRIM provided

engineering content and prototypes on cloud computing, remote sensing, authentication and security, asynchronous processing, workflow execution, and job management.

For TB-13, U.S. DOE provided OGC with specifications related to ESGF. Helped by ESGF CWT, CRIM explored cloud implementations for climate processes and initiated integration in both PAVICS and Birdhouse. An implementation goal was the development of an integrated solution compatible for both NRCan and ESGF. Both TB-13 deliverables will be made available via managed services at CRIM to other TB participants and their authorized affiliates for a period of one calendar year. Ideas on the upcoming TB at OGC proposes collaborative TB experiments in federated environments. Initial findings indicate that ESGF would be an appropriate case study. OGC demonstration event for TB-13 is planned for the second week of December, in Reston, Virginia.

Using the ESGF CWT API in the Context of the EUDAT-EGI e-Infrastructure and the ENES Climate4Impact Platform

Christian Pagé (CERFACS), christian.page@cerfacs.fr
Xavier Pivan (CERFACS), xavier.pivan@cerfacs.fr
Asela Rajapakse (MPI-M), asela.rajapakse@mpimet.mpg.de
Wim Som de Cerff (KNMI), wim.som.de.cerff@knmi.nl
Maarten Plieger (KNMI), maarten.plieger@knmi.nl
Ernst de Vreede (KNMI), ernst.de.vreede@knmi.nl
Alessandro Spinuso (KNMI), alessandro.spinuso@knmi.nl
Lars Barring (SMHI), Lars.Barring@smhi.se
Antonio Cofino (University of Cantabria), antonio.cofino@unican.es
Alessandro d'Anca (CMCC), alessandro.danca@cmcc.it
Sandro Fiore (CMCC), sandro.fiore@cmcc.it

Supporting data analytics in climate research with respect to data access is a challenge due to increasing data volumes. Several international and European initiatives have emerged and provide standalone solutions that offer potential for interoperability. In Europe, the IS-ENES (https://is.enes.org) consortium has developed a platform to ease access to climate data for the climate impact community (https://climate4impact.eu). It exposes data from ESGF data nodes as well as any OPeNDAP (Open-Source Project for a Network Data Access Protocol) server. It provides UIs, wizards, and services for search and discovery, visualization, processing, and downloading. Also in Europe, an emerging e-infrastructure is being designed and built for several scientific domains, led by EUDAT (https://eudat.eu) and EGI (https://egi.eu), which will form the basis of the future EOSC to support scientific researchers. This e-infrastructure provides services within the EUDAT Collaborative Data Infrastructure (CDI). The ENES climate community is participating in the EUDAT CDI.

Within the EUDAT project, work has been done to integrate these existing e-infrastructures. The goal is to develop interoperable interfaces:

1. A first-level prototype has been completed that deploys the Generic Execution Framework (GEF) Docker backend onto the EGI FedCloud to perform computations and feeds the results into the EUDAT CDI.
2. The second-level prototype involves integrating the GEF backend and the ESGF CWT API. The GEF backend pulls data from the ESGF infrastructure through the CWT API so that data reduction is achieved through on-demand calculations. Furthermore, complex

calculations are then executed on the EGI FedCloud, and the results are fed back into EUDAT B2 services. This raises the authentication and authorization integration between the ESGF and EUDAT/EGI. A first solution would be to use the token-based approach of Climate4Impact.

3. The third-level prototype is the same as the second one, except that the GEF is executed by the Climate4Impact platform on demand by the user, and the final results are fed back into the Climate4Impact user space for visualization, storage, or download. A variant of this third-level prototype that could be implemented is getting the data directly from one or several ESGF data nodes, using the Climate4Impact Search WPS to locate the data files.

A demo of the third-level prototype will be presented as well.

Managing Growth and Complexity – Technologies to Meet the Challenges of Operating Data, Services, and Infrastructure at Scale

Phil Kershaw (ENES/CEDA), philip.kershaw@stfc.ac.uk
Jonathan Churchill (ENES/CEDA), jonathan.churchill@stfc.ac.uk
Alan Iwi (ENES/CEDA,) alan.iwi@stfc.ac.uk
Bryan Lawrence (University of Reading), bryan.lawrence@ncas.ac.uk
Neil Massey (ENES/CEDA), neil.massey@stfc.ac.uk
Sam Pepler (ENES/CEDA), sam.pepler@stfc.ac.uk
Matt Pritchard (ENES/CEDA), matt.pritchard@stfc.ac.uk
Matt Pryor (ENES/CEDA), Matt.Pryor@stfc.ac.uk
Ag Stephens (ENES/CEDA), ag.stephens@stfc.ac.uk

CEDA hosts data and services for a wide range of communities and with international collaboration efforts such as the ESGF. Since 2012, the underlying computing infrastructure has been provided by JASMIN, a shared computing platform for the environmental sciences community consisting of a high-performance, high-volume storage system to host key datasets co-located with computing resources for processing and analysis.

Operational experiences to date and lessons learnt are informing decisions about JASMIN's and ESGF's future technical direction. With the success of the system, both the quantity of data stored and the number of supported users have grown. As part of its technical evolution, a program of work is underway to address the challenges associated with this growth. Here we highlight two specific technologies being used as part of that program of work: object storage and containers. These will bring fundamental changes to how we operate JASMIN and their consequent impact on our existing service infrastructure like the ESGF.

A number of factors are driving the adoption of object storage, but there are issues with adopting it in the JASMIN environment. We will briefly discuss these factors and issues before introducing our plans to migrate to object storage for JASMIN. These plans include the development of domain-specific software customized to exploit HDF version 5/netCDF4 data held in object stores, providing the user community efficient access consistent with existing familiar interfaces.

We will also describe the application of the container technologies Docker and Kubernetes to underpin the provision and operation of new services including climate services for the EU Copernicus program—which re-uses the ESGF application stack—and the development of a new

Cluster-as-a-Service concept for JASMIN's cloud: the dynamic provision of clusters to host new compute and analysis applications such as Jupyter Notebooks, Dask, and PySpark.

Ophidia: An Interoperable 'Big Data' Framework for Climate Change Analytics Experiments

Sandro Fiore (CMCC), sandro.fiore@cmcc.it
Charles Doutriaux (LLNL/AIMS), doutriaux1@llnl.gov
Cosimo Palazzo (CMCC), cosimo.palazzo@cmcc.it
Alessandro d'Anca (CMCC), alessandro.danca@cmcc.it
Zeshawn Shaheen (LLNL/AIMS), shaheen2@llnl.gov
Donatello Elia (CMCC), donatello.elia@cmcc.it
Jason Boutte (LLNL/AIMS), boutte3@llnl.gov
Valentine Anantharaj (ORNL/E3SM), anantharajvg@ornl.gov
Dean N. Williams (LLNL/AIMS), williams13@llnl.gov
Giovanni Aloisio (CMCC), giovanni.aloisio@unisalento.it

The Ophidia project provides a complete environment for scientific data analysis on multidimensional datasets. It exploits data distribution and supports array-based primitives for mathematical and statistical operations, analytics jobs management and scheduling, and a native in-memory input/output (I/O) server for fast data analysis. It also provides access through standard interfaces like SOAP, GSI/VOMS, and OGC-WPS.

In the climate change domain, the Ophidia framework has been applied to support the implementation of real use cases on multi-model analysis, climate indicators, and processing chains for operational environments in different European projects.

A recent effort concerns a new interface implementing the ESGF WPS Extension Specification. In this regard, a complete Python-based interface has been developed to support Ophidia workflows submission by means of Python clients and applications. Authentication and authorization are guaranteed through a token-based approach. The remote submissions exploit the Ophidia workflow engine interface which exposes several constructs to implement different features (e.g., loops, automated workflows, arguments, interleaved mechanisms, parallelism).

The Ophidia stack along with the ESGF WPS compliant interface has been installed in OphidiaLab, a new multi-user environment for scientific data analysis deployed at the CMCC Supercomputing Center.

The demonstration will focus on the new OphidiaLab environment, the WPS-enabled interface, the workflow capabilities provided by Ophidia, and some use cases from European projects like EUBra-BIGSEA (Europe–Brazil Collaboration of Big Data Scientific Research Through Cloud-Centric Applications) and INDIGO-DataCloud.

Federated Data Usage Statistics in the Earth System Grid Federation

Alessandra Nuzzo (ENES/CMCC), alessandra.nuzzo@cmcc.it
Maria Mirto (CMCC) maria.mirto@cmcc.it
Paola Nassisi (CMCC) paola.nassisi@cmcc.it
Katharina Berger (DKRZ), berger@dkrz.de
Torsten Rathmann (DKRZ), rathmann@dkrz.de
Luca Cinquini (NASA/JPL), Luca.Cinquini@jpl.nasa.gov
Sébastien Denvil (ENES/IPSL), sebastien.denvil@ipsl.jussieu.fr

Sandro Fiore (CMCC), sandro.fiore@cmcc.it
Dean N. Williams (LLNL/AIMS), williams13@llnl.gov
Giovanni Aloisio (CMCC), giovanni.aloisio@unisalento.it

The federated monitoring system plays an important role in the context of the ESGF. This task is accomplished by the ESGF-dashboard component, which is composed by a backend and a front-end module: the former dedicated to managing data usage statistics at single site and federation level and the latter providing a flexible and usable web interface.

The main goal of the ESGF-dashboard is to provide a distributed and scalable monitoring framework responsible for capturing usage metrics at the single site level and at the global ESGF level.

The backend component of the ESGF-dashboard, included into the software stack of the ESGF data node, has a main role of collecting and storing a high volume of heterogeneous metrics, covering measures such as downloads and clients' statistics, aggregated cross and project-specific download statistics. With respect to the previous version of the dashboard front-end, its final implementation moves away from the previous desktop metaphor to approach a brand new one closer to the dashboard concept with a stronger usability.

The new UI, already deployed in production, provides a rich set of charts and reports through a web interface, allowing users and system managers to visualize the status of the infrastructure through a set of smart and attractive web gadgets. The key challenges of such concept are to communicate the most important information in a straightforward way and allow users to view specific details at the same time.

The collection of federated statistics is accomplished through a RESTful API that retrieves and aggregates metrics from all data nodes across the federation.

WPS-Based Processing Services for the Copernicus Climate Change Service (C3S)

Stephan Kindermann (DKRZ), kindermann@dkrz.de
Carsten Ehbrecht (DKRZ) ehbrecht@dkrz.de
Ag Stephens (CEDA) ag.stephens@stfc.ac.uk
Björn Brötz (DKRZ) Bjoern.Broetz@dlr.de
Wim Som de Cerff (KNMI) wim.som.de.cerff@knmi.nl
Maarten Plieger (KNMI) maarten.plieger@knmi.nl
Sébastien Denvil (ENES/IPSL) sebastien.denvil@ipsl.jussieu.fr

The C3S will integrate global and regional climate projections into the Climate Data Store[4]. The Climate Data Store will also provide consistent access to in situ and satellite-based climate observations, reanalysis data and multi-model seasonal forecasts. On the data access side, the ESGF and its data services (search, authentication, download, and subset) will provide the interface layer between C3S and the model data archives at DKRZ, the Science and Technology Facilities Council (STFC), and IPSL.

To provide data near processing services, a new service component will be developed and deployed near to the data archives. These services are supporting OGC WPS standardized interfaces and thus are supported by a wide range of different client tools and applications.

[4] https://www.ecmwf.int/en/about/what-we-do/environmental-services/copernicus-climate- change-service

We will provide an overview of the status of the Copernicus processing approach, including software packaging and deployment as well WPS development and deployment, which is based on the Birdhouse[5] open-source initiative. With respect to the processing codes to be made available via WPS, we are concentrating on climate data evaluation packages developed as part of the Copernicus C3S-Magic project. Key cornerstones of the approach presented are:

- A generic software packaging and deployment solution based on Conda and Docker
- A generic WPS component system supporting the flexible generation and deployment of WPS standardized web services (Birdhouse based)
- Support for parallel processing clusters based on different batch systems (e.g., SLURM, GridEngine)

A demo of a first test deployment will be presented.

Diagnostics Package for the E3SM Model
Chengzhu Zhang (LLNL/AIMS), zhang40@llnl.gov
Zeshawn Shaheen (LLNL/AIMS), shaheen2@llnl.gov
Chris Golaz (DOE/E3SM), golaz1@llnl.gov
Jerry Potter (NASA/GSFC), gerald.potter@nasa.gov

A new E3SM diagnostics package has been developed by the E3SM Workflow team to build a comprehensive diagnostics software that facilitates the diagnosis of the next-generation Earth system models. This package is embedded into the E3SM Automated Workflow for seamless transition between model run and diagnostics.

This software is designed in a flexible, modular, and object-oriented fashion, enabling users to manipulate different processes in a diagnostics workflow. Numerous configuration options for metrics computation (i.e., regridding options) and visualization (i.e., graphical backend, color map, contour levels) are customizable. Built-in functions to generate derived variables and select diagnostics regions are supported and can be easily expanded.

The architecture of this package follows the CDP framework, which is also applied by two other DOE-funded diagnostics efforts (PCMDI metrics package and ARM diagnostics package), to facilitate effective interactions between different projects.

ESGF Errata Service
Guillaume Levavasseur (ENES/IPSL), glipsl@ipsl.jussieu.fr
Atef Ben-nasser (ENES/IPSL), abennasser@ipsl.fr
Mark A. Greenslade (ENES/IPSL), momipsl@ipsl.jussieu.fr

Due to the inherent complexity of the experimental protocols of projects such as CMIP5 and CMIP6, it becomes important to record and track reasons for dataset version changes.

The IPSL is finalizing a new ESGF Errata Service, currently under test phase at http://test.errata.es-doc.org/, to:

- provide timely information about known issues within the ES-DOC ecosystem;

[5] http://birdhouse.readthedocs.io/en/latest/

- allow identified and authorized actors to create, update, and close an issue using lightweight client (https://es-doc.github.io/esdoc-errata-client/); and
- enable users to query about modifications and/or corrections applied to the data in different ways through a dedicated API (https://es-doc.github.io/esdoc-errata-client/api.html).

The Errata Service exploits the PID attached to each dataset during the ESGF publication process. The PIDs request the Handle Service to get the version history of a (set of) file/dataset(s). Consequently, IPSL is closely working with DKRZ on the required connections and APIs between both services.

A first demonstration of the service has been very well received from the ESGF developer community. A release candidate of the service is currently delivering to potential users with a goal of deploying into production before the end of the year. This release will include two improvements:

- pyessv Controlled Vocabulary Manager, and
- issue registration support for any ESGF project.

DREAM Data Services for Biological Data and Beyond
Sasha Ames (LLNL/AIMS), ames4@llnl.gov
Luca Cinquini (NASA/JPL), Luca.Cinquini@jpl.nasa.gov
Dean N. Williams (LLNL/AIMS), williams13@llnl.gov

In this poster and demo, we introduce an alternate data service for DREAM. The TDS has been very effective for serving the netCDF data published in the ESGF. However, we need a service more specific for alternate data (e.g., ASCII-based) in other domains, such as the FASTA format used in bioinformatics to represent genomic and protein sequences. This service will allow a variety of content types to interoperate properly with a user's web browser. We will also show how non-netCDF data are published. Future work for the service will include random access for FASTA.

Community Data Analysis Tools
Charles Doutriaux (LLNL/AIMS), doutriaux1@llnl.gov
Denis Nadeau (LLNL/AIMS), nadeau1@llnl.gov
Dan Lipsa (Kitware), dan.lipsa@Kitware.com
Dean N. Williams (LLNL/AIMS), williams13@llnl.gov
Aashish Chaudhary (Kitware), aashish.chaudhary@kitware.com

CDAT is an open-source, Python-based suite of tools designed to provide many of the basic capabilities needed for validating, comparing, and diagnosing scientific data, with an emphasis on climate model behavior. It can be controlled either interactively or via a script file, or control can alternate between these modes during a session. Its strengths are that it allows users to: (1) build end-to-end complex data analysis and visualization workflows that use predefined components for data transformations; (2) collect data from disparate data sources; and (3) ingest user-defined local and remote processing steps.

CDAT's success can also be measured by its expanding use. It is now integrated with the international ESGF peer-to-peer enterprise system as a front-end access mechanism to acquire data for analysis and visualization and as a prototype backend tool to reduce data sets and return

visualization products. It is also expanding into other DOE-, NOAA-, and NASA-funded projects as the cornerstone of interagency proposed projects. DOE's E3SM project aims to use CDAT to deliver new capabilities that will further facilitate interactive and visual exploration and diagnostics of simulation and observational output. This project shares a collaborative vision for large-scale visualization and analysis of climate data and is working to organize and expand CDAT's capabilities. The design of CDAT incorporates the following requirements:

- Interactive and batch operations
- Workflow analysis and provenance management
- Parallel visualization and analysis tools (exploiting parallel I/O)
- Local and remote visualization and data access
- Comparative visualization and statistical analyses
- Robust tools for regridding, projection, data subsetting, and aggregation
- Support for unstructured grids and non-gridded observational data, including geospatial formats often used for observational datasets

The CDAT offers capabilities for climate scientists to manage big data analytics, sensitivity analyses, heterogeneous data sources, and multiple disciplinary domains, incorporating existing software components in combinations that were previously difficult or even impossible. The CDAT framework addresses challenges in analysis and visualization and incorporates new opportunities, including parallelism for better efficiency, higher speed, and more accurate scientific inferences. Today, the open-source CDAT provides hundreds of users access to increasing 1D, 2D, and 3D analysis and visualization products on many different operating system platforms (i.e., Linux/Unix, Windows, Mac OSX).

Visual Community Data Analysis Tools (vCDAT)

Matthew Harris (LLNL/AIMS), harris112@llnl.gov
Dan Lipsa (Kitware), dan.lipsa@kitware.com
James Crean (LLNL/AIMS), crean2@llnl.gov
Matthew Ma (Kitware), matthew.ma@Kitware.com
Charles Doutriaux (LLNL/AIMS), doutriaux1@llnl.gov
Dean N. Williams (LLNL/AIMS), williams13@llnl.gov
Aashish Chaudhary (Kitware), aashish.chaudhary@kitware.com

Parallel computing, workflows and provenance, exploratory analysis, big data processing for analysis, interactive analysis and visualization, and web informatics are some of the key features of the overall CDAT framework. To support these features, CDAT utilizes core technologies from open-source toolkits such as VTK, R, NumPy, SciPy, and a host of others. In its current format, the (vCDAT sits on the web server and provides a Python-based API, which provides the ability to read data from local or remote sources, run analysis algorithms on local or remote computing resources in serial or parallel mode, and visualize algorithm output in a thick client (e.g., desktop GUI) or a smart client (e.g., web browser). CDAT can use this computing server-side horsepower of a cluster or a supercomputer. The ability to connect to other instances of CDAT compute nodes is under development. On the client side, the deprecated desktop GUI used the CDAT Python API, whereas communication between its smart client replacement and

the Python framework uses the latest in web technologies, such as web-sockets and a RESTful API.

Our web-based analysis and visualization system, vCDAT, uses the traditional client–server architecture concept within the web-based model. It is similar to the thick-client concept in that the vCDAT smart clients are Internet-connected devices that allow a user's local applications to interact with server-based applications through the use of web services. This allows for more analysis and visualization interaction and software customization but without the hassle of software downloads and installation.

Integrating ES-DOC with the ESG Publisher
Alan Iwi (ENES/CEDA), alan.iwi@stfc.ac.uk
David Hassell (NCAS/UoR), david.hassell@ncas.ac.uk
Mark A. Greenslade (ENES/IPSL), momipsl@ipsl.jussieu.fr
Ag Stephens (ENES/CEDA), ag.stephens@stfc.ac.uk

The ES-DOC ecosystem has the capacity to capture and deliver essential information about climate modeling activities. Within CMIP6, scientists are describing their models and experiments in detail using a rich semantic model (Common Information Model [CIM] 2). Additionally, ES-DOC requires information about ensemble runs and each individual simulation. The extensive global metadata in CMIP6 netCDF data files will provide enough information to allow the ensemble and simulation records to be generated by scanning the file system directly.

A command-line tool and Python library, cdf2cim, has been developed to manage the file-scanning, serialization to JSON, and upload to the ES-DOC server.

The ESG Publisher captures information from data files to generate aggregations and metadata summaries suitable for publishing to various sources, including THREDDS (Thematic Real-Time Environmental Distributed Data Services) and the ESGF Search system. Since all CMIP6 data (in the ESGF) will pass through the Publisher, it was considered appropriate to interface with cdf2cim in order to generate CIM 2 for ES-DOC. Staff at IPSL, the National Centre for Atmospheric Science (NCAS)/University of Reading (UoR), and STFC CEDA collaborated on building an extension to the Publisher that automates the generation of CIM 2 metadata and sends it to the server. We will describe how data node managers will work with these tools for CMIP6. This includes the use of GitHub tokens to authenticate with the ES-DOC server.

This solution further integrates the publication of data and metadata from detailed climate simulations. Beyond CMIP6, this approach would be applicable to other similar projects.

ESG Publisher: https://esgf.github.io/esg-publisher/
CIM 2: https://es-doc.org/cim
CDF2CIM: https://es-doc.org/utility-library-cdf2cim

Compute Working Team End-User Application Programming Interface
Jason Boutte (LLNL/AIMS), boutte3@llnl.gov
Charles Doutriaux (LLNL/AIMS), doutriaux1@llnl.gov

The ESGF CWT end-user API was created to leverage the power of the WPS interface standard. A WPS server can expose large-scale computational processes and data reduction that are location agnostic, allowing the computations and reductions to be performed where the data reside, 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 a WPS process, a well-defined climatology-specific API was planned, and an object-oriented Python end-user API and server implementation were created. With the API, users are eased into adoption of these WPS processes.

A Compliance-Checking Framework for CMIP7
Ag Stephens (ENES/CEDA), ag.stephens@stfc.ac.uk
Antony Wilson (STFC), antony.wilson@stfc.ac.uk
Guillaume Levavasseur (ENES/IPSL), glipsl@ipsl.jussieu.fr

The activity known as "compliance-checking" is distinct from "quality control/assurance" in that it is not concerned with the scientific credibility of the results but aims to check that data files adhere to a set of rules associated with a given project. There are many tools for compliance-checking that perform a very useful function for specific projects. However, it is commonplace for the written specification for a project to diverge from the software implementation.

A prototype compliance-checking "framework" is described that draws on the positives of its predecessors and is suggested as a model suitable for future MIPs. Built on the existing Integrated Ocean Observing System (IOOS) compliance-checker, which employs a plugin architecture per project, the framework attempts to provide a clear separation of concerns between the code implementing the "checks" and the project-specific configuration.

In the main library, each check is encoded in a Python class, including documentation (to describe what the check does), messages (to report successes/failures), and modifier parameters (so that each check has some defined flexibility).

On the configuration side, a separate library (compliance-check-maker) is concerned with generating the specific checks to be employed by a project. Each project writes a set of YAML files to describe a group of checks (e.g., to check global attributes or file names). The code then generates a Python plugin for the IOOS checker as well as a compliance specification document for the project. The latter is an essential feature of the framework, generating both the checks and specification from a single information source.

The modular approach employed by the framework makes it highly adaptable to different use cases, and the community is encouraged to add to the collection of supported checks. As the tool develops, it should take hours, rather than days or months, to configure a set of checks for a new project.

IOOS compliance-checker: https://github.com/ioos/compliance-checker

Google Earth Engine and Project Jupyter
Tyler Erickson (Google), tylere@google.com

The volume of Earth science data generated from models and by sensors (particularly those on satellites) continues to increase. For many analyses, managing this large volume of data is a barrier to progress, as it is difficult to explore and analyze large volumes of data using the traditional paradigm of first downloading datasets to a local computer. Furthermore, methods are needed that communicate Earth science algorithms that operate on large datasets in an easily understandable and replicable way.

This demo will highlight two technologies:

- Google Earth Engine – a cloud-based geospatial analysis platform that provides access to PBs of Earth science data and hundreds of geospatial operators via a JavaScript or Python API.
- Project Jupyter – an open-source project that supports interactive data science and scientific computing, including the Jupyter Notebook, a web-based environment that supports documents that combine code and computational results with text narrative, mathematics, images, and other media.

The technologies will be demonstrated by calculating climate indices from downscaled climate projections based on CCl/CLIVAR/JCOMM Expert Team on Climate Change Detection and Indices.

Day 2: Wednesday, December 6, 2017

ESGF Focus Areas

International Climate Network Working Group, Replication/Versioning and Data Transfer Working Team

Eli Dart (DOE/ESnet), dart@es.net
Lukasz Lacinski (DOE/ANL), lukasz@uchicago.edu
Stephan Kindermann (ENES/DKRZ), kindermann@dkrz.de

Efficient CMIP6 data analysis depends on the transfer and replication of high-volume datasets to data centers around the world. These data centers manage replica pools to support their user communities by, for example, redistributing the data or by providing data near processing facilities. The data transfer and replication are integrated into a complex workflow involving file systems, local networks, wide area networks, as well as dedicated DTNs, which are integrated into a data pipeline managed by dedicated data replication software installed at sites.

This session will provide an overview of the current status of the overall CMIP6 data replication pipeline and its different (technical and organizational) aspects. The session especially concentrates on the following:

- Data transfer and replication
    - Status and progress so far
    - Current problems
- CMIP6 replication strategy
    - Status of current discussion
    - Planning of the international data replication to well-established “CMIP6 data hubs” such that, for instance, transatlantic connections can be exploited efficiently
- Short-term action planning to support CMIP6 initially
    - Improve single-stream bandwidths to CMIP6 data servers from DTNs
    - Configuration issues at sites
    - Data publication to support download via DTNs

- Long-term action planning to support CMIP6+
    - Exploit Globus Transfer in replication pipeline
    - Expand DTN deployments to match data scale

Compute and Data Analytics Working Team

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

Compute Working Team Update: Server-Side Computing Progress
Charles Doutriaux and Daniel Duffy

The ESGF's main goal is to facilitate advancements in Earth system science with a primary mission of supporting CMIP activities. In preparation for emerging data analysis needs, such as future climate assessments, the CWT has been working to provide data-proximal analytics capabilities through the development of server-side APIs and client-side (end-user) APIs. This talk will provide a brief overview of ongoing development projects focused on the implementation of ESGF server-side analytics and discuss future goals of the working team.

ViSUS Streaming Analysis/Visualization: Interactive Analysis and Visualization of Arbitrarily Large, Disparately Located Climate Data Ensembles Using a Progressive Runtime Server, On-Demand Data Conversion, and an Embedded Domain Specific Language Suitable for Incremental Computation
Cameron Christensen, Giorgio Scorzelli, Peer-Timo Bremer, Shusen Liu, Ji-Woo Lee, Brian Summa, Valerio Pascucci

Massive datasets are becoming more common due to increasingly detailed simulations and higher resolution acquisition devices. Yet accessing and processing these huge data collections for scientific analysis is still a significant challenge. Solutions that rely on extensive data transfers are increasingly untenable and often impossible due to lack of sufficient storage at the client site as well as insufficient bandwidth to conduct such large transfers, that in some cases could entail PBs of data. Large-scale remote computing resources can be useful, but utilizing such systems typically entails some form of offline batch processing with long delays, data replications, and substantial cost for any mistakes. Both types of workflows can severely limit the flexible exploration and rapid evaluation of new hypotheses that are crucial to the scientific process and thereby impede scientific discovery.

To facilitate interactive analysis and visualization of these data ensembles, we introduce a dynamic runtime system suitable for progressive computation and interactive visualization of arbitrarily large, disparately located spatiotemporal datasets. This system is based on the streaming IDX data format, which utilizes a hierarchical z-order to facilitate fast loading of coarse resolution data as well as better spatial locality for more efficient sub-region reads.

We provide an on-demand IDX data server to enable access to existing datasets, designed to provide streaming hierarchical versions of equivalent netCDF climate data volumes in a user-directed manner such that specific timestep fields are converted just-in-time. This permits the bulk of the data to remain on the server and facilitates interactive analysis and visualization by immediately sending results for specific data requests. Initial conversions are cached for future use, amortizing the cost across successive requests.

Our system includes an embedded domain-specific language that 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 processing. Computations involving large amounts of data can be performed remotely in an incremental fashion that dramatically reduces data movement, while the client receives updates progressively, thereby remaining robust to fluctuating network latency or limited bandwidth.

This system is integrated with the ESGF software stack using a Docker-based deployment, and facilitates interactive, incremental analysis and visualization of massive remote datasets up to PBs in size.

Identity, Entitlement, and Access Working Team

Phil Kershaw (ENES/CEDA), philip.kershaw@stfc.ac.uk
Lukasz Lacinski (DOE/ANL), lukasz@uchicago.edu

Over the past year, the IdEA working team has focused on the packaging of OAuth2 support into the ESGF release in order to replace the legacy OpenID 2.0 system and bring new capabilities to applications and services in the operational federation. This work has centered on two main aspects: (1) the incorporation of the standalone OAuth2 Authorization and Resource services implementation from CEDA, and (2) work to embed OAuth2 with the various dependent ESGF services—the compute, index, and data nodes. We will describe the new core identity services including the SLCS, which with OAuth2 provides a means to get delegated certificates. We will also describe the provision of the various integration hooks to other ESGF services: the refactoring needed to integrate OAuth with the ORP—the access control filter system overlaying the TDS—and the development of a generic Python OAuth Client package by ANL.

In addition to integration with the ESGF installer, work with JPL has been undertaken to run the OAuth, SLCS, and dependent services as Docker containers. We will review the roadmap for making this and other functionality operational and outline our plans for the future evolution of the IdEA architecture.

User Interface, Search, and Dashboard Working Teams

Luca Cinquini (NASA/JPL), Luca.Cinquini@jpl.nasa.gov
Guillaume Levavasseur (ENES/IPSL), glipsl@ipsl.jussieu.fr
Alessandra Nuzzo (ENES/CMCC), alessandra.nuzzo@cmcc.it

This presentation will provide a progress report and future roadmap for the recently unified working group that includes the CoG UI, the search backend services, and the Dashboard and metrics functionality.

CoG development has been focused on integrating new features in support of critical community projects such as the upcoming CMIP6 and the ongoing Obs4MIPs. If funding is provided, we plan to completely re-factor the CoG software to enhance its modularity, functionality, security, and extensibility. Also, because the front-end is more and more requested by non-scientific users from different backgrounds, future efforts must lead to a friendlier interface with intuitive layouts and helpful tutorials.

The backend search services have been mostly stable, with some development effort again focused on supporting CMIP6 features, as well as addressing newly discovered security

vulnerabilities. Future work must address necessary Solr upgrades and perhaps moving to Solr Cloud.

The dashboard UI application has been completely re-written since the ESGF shut-down. The dashboard is deployed as an information provider as part of each data node, and since the last release it includes a RESTful API. Federation-level metrics are provided by aggregator applications that will be deployed at selected Tier 1 sites.

Installation and Software Security Working Team

William Hill (LLNL/AIMS), hill119@llnl.gov
Sasha Ames (LLNL/AIMS), ames4@llnl.gov
Prashanth Dwarakanath (ENES/LiU), pchengi@nsc.liu.se
Luca Cinquini (NASA/JPL), Luca.Cinquini@jpl.nasa.gov
George Rumney (NASA/GSFC), george.rumney@nasa.gov

Software Installation Working Team
William Hill (LLNL/AIMS) Sasha Ames (LLNL/AIMS) and Prashanth Dwarakanath (ENES/LiU)

This presentation will provide an update on the current state of the ESGF installation process and the future direction of the installer. The installer version 2.5.13 is currently written as a collection of Bash scripts. Steady progress has been made in porting these Bash scripts over to Python. Refactoring the script to Python will benefit both the ESGF developers and users. Python enhances the codebase's readability, maintainability, and testability, thus speeding up the development cycle for future releases. The code will be written to be more modular in structure. Additionally, the refactor opens the possibility for creating a web interface for the installer.

Software Container Architecture (i.e., Docker)
Luca Cinquini (NASA/JPL)

This presentation will report on the current state of the effort to design and implement a next-generation ESGF architecture based on Docker containers. Such a model presents great advantages with respect to the current "monolithic" architecture supported by the shell-based installer, such as easier to install and upgrade, scalable onto multiple hosts, and deployable both on internal clusters and commercial Cloud. This work has so far been supported by the DREAM project and is now joining forces with the new European Copernicus project.

Software Security Working Team
George Rumney (NASA/GSFC), Daniel Duffy (NASA/GSFC), Luca Cinquini (NASA/JPL), and Dean N. Williams (LLNL/AIMS)

The Executive Committee of the ESGF chartered a Software Security Working Team (SSWT) to oversee the security of the ESGF software stack and to provide guidance for a continuous improvement path consistent with federal controls. The SSWT maintains the security review procedure for all ESGF software releases and is responsible for ensuring best practices are maintained across the federation. For more information, the software security plan can be found at http://esgf.llnl.gov/media/pdf/ESGF-Software-Security-Plan-V1.0.pdf. While progress has been made, significant challenges remain within such a complex software stack. This short talk will highlight those challenges, the need for more involvement across the community, and near-term goals.

Day 3: Thursday, December 7, 2017

Coordinated Efforts with Community Software Projects

Publication, Quality Control, Metadata, and Provenance Capture Working Team

Sasha Ames (LLNL/AIMS), ames4@llnl.gov
Heinz-Dieter Hollweg (ENES/DKRZ), hollweg@dkrz.de
Bibi Raju (PNNL), mailto:bibi.raju@pnnl.gov

Publication Working Team
Sasha Ames (LLNL/AIMS)

The ESGF publication team supports many different projects, although the most recent focus has been on CMIP6 readiness. This talk will give an overview of the current state of the process and our future directions. CMIP6 is an example of a project where the Publisher is enabled for attribute-controlled vocabulary checking. Another example project, Input4MIPs, make use of several unconventional features. These projects introduced the PID assignment functionality aspect of the process. Python 3 conversion will become an action item for the next calendar year.

Quality Control Progress
Heinz-Dieter Hollweg (ENES/DKRZ)
The current state of the QA-DKRZ tool is presented. A Best Practices procedure for installation/update is given as well as configuration, operating the tool for projects like CMIP6 and eventually summarizing the results.

Provenance Data Harvest and Scientific Results Reproducibility
Bibi Raju (PNNL)

Data provenance provides a way for scientists to observe how experimental data originates, conveys process history, and explains influential factors such as experimental rationale and associated environmental factors from system metrics measured at runtime. PNNL developed a provenance harvester that is capable of extracting already existing file-based information produced by applications. File based information is extracted and transformed into an intermediate data format inspired in part by W3C CSV on the Web recommendations, called the HAPI syntax. This syntax provides a general means to pre-stage provenance into messages that are both human readable and capable of being written to a provenance store, ProvEn. The harvested provenance data can later be retrieved from ProvEn store and can be used for various purposes. This extracted information greatly helps to reproduce a simulation either by the same user or a different user in the same host environment. The harvested provenance information can be also used to compare different application runs.

HAPI is being applied to harvest provenance from climate ensemble runs for (E3SM project funded under the U.S. DOE's Office of BER Earth System Modeling program. E3SM informally provides provenance in a native form through configuration files, directory structures, and log files that contain success/failure indicators, code traces, and performance measurements. HAPI is a generic format and can be applied to harvest provenance from relational database tables as well as other scientific applications that log provenance related information.

Machine Learning

Sookyung Kim (LLNL/AIMS), kim79@llnl.gov
Sebastien Denvil (ENES/IPSL), sebastien.denvil@ipsl.fr

Philip Kershaw (ENES/CEDA), philip.kershaw@stfc.ac.uk
Tom Landry (CRIM), tom.landry@crim.ca

Deep Learning Application for Community Machine Learning
Sookyung Kim (LLNL/AIMS)

This presentation will report on the progress of current effort to leverage DL techniques to detect, localize, and track extreme climate events using the ESGF framework. Specifically, we present recent results of the system we developed to detect and locate extreme climate events by CNNs. Our system can capture the pattern of extreme climate events from pre-existing coarse reanalysis data corresponding to only 16,000 grid points—without an expensive downscaling process and with fewer than 5 hours to training using 5-layered CNNs. As the use case of our framework, we tested tropical cyclone detection with labeled reanalysis data and achieved 99.98% of detection accuracy with localization accuracy within 4.5 degrees of longitude/latitude. In addition, we will introduce the prototype of the DL system to track the extreme climate events by considering spatiotemporal evolution of an event using LSTM, which can track the event in time-series reanalysis data.

Presentation on Copernicus and H2020 Programme Machine Learning Efforts
Sébastien Denvil (ENES/IPSL), Sandro Fiore (ENES/CMCC), Philip Kershaw (ENES/CEDA)

The past ten years have witnessed ancient ML and DL algorithms awaken. This trend was accompanied by large disruption in so called “big data” technology (cloud, GPU, Docker, and alike).

There are many applications of ML in the field of EO. Those algorithms are very well placed to fill gaps in observations. In the field of modelization, there are a few examples of ML usage for model parameters tuning but the full potential of ML with respect to climate modeling has not yet been fully realized.

This talk will discuss how those trends in ML and DL and how the can be leveraged in the climate community and the role ESGF could play.

Imagery, Text, and Geospatial Machine Learning Applications in Montreal’s Booming ML Landscape
Tom Landry (CRIM)

Montreal's technological landscape is currently transformed by an Artificial Intelligence revolution. Several major world-class tech corporations recently opened laboratories and offices in the city. New research chairs, super-clusters and specialized institutes are living proof of increasing provincial and federal public investments in the domain. Prospects for startups are also getting better; the talent pool is large and venture capital is receptive.

CRIM is an applied research center positioned in the middle of academia and industry. Three research teams at CRIM—Vision and Imaging (VISI), Emerging Technologies and Data Science (TESD), and Speech and Text (PATX)—have been delivering and transferring ML expertise and applications for several years. We will present of few of our ML projects that are susceptible to be of use for ESGF.

VISI demonstrated ML in target detection, classification, super-resolution and filtering from several sensing types: either SAR, optical satellites, LIDAR, aerial images or video sequences.

TESD produced a highly scalable grid-density clustering algorithm for Spark MLLib and unsupervised ML on climate products with SciSpark. The team also works closely with the city of Montreal in Smart City scenarios for both descriptive and predictive analytics. PATX's natural language understanding expertise recently allowed them to propose future work for climate and EO data. This future work includes metadata annotation for active learning, query understanding interface, and workflow recommendations.

Diagnostics

Zeshawn Shaheen (LLNL/AIMS), shaheen2@llnl.gov
Tom Landry (CRIM), tom.landry@crim.ca

Community Diagnostics Package (CDP)
Zeshawn Shaheen (LLNL/AIMS)

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 Analytics and Informatics Management Systems (AIMS) team at LLNL is developing the CDP, a framework for creating new diagnostics packages in a generalized manner. Designed in an object-oriented method, CDP allows for a modular implementation of the components required for running diagnostics. The design of CDP consists of modules to handle the user-defined parameters, metrics, provenance, file I/O, output of results, and algorithms for calculating the diagnostics.

Copernicus and H2020 Programme Diagnostics Needs and Overview
S. Denvil (IPSL), M. Lautenschlager (DKRZ), S. Fiore (CMCC), F. Guglielmo (IPSL), M. Juckes (CEDA), S. Kindermann (DKRZ), M. Kolax (SMHI), C. Pagé (CERFACS), W. Som de Cerff (KNMI)

A diverse set of software, tools, frameworks, and technologies are available around the globe for computing diagnostics relevant for climate modeling simulation results analysis: Birdhouse, Climate4Impact, NCO, ESMValTools, climate data operators, Climaf, vCDAT, WPS, MAGICS, and so on. Diagnostics themselves can be tailored towards climate modelers as well as to scientific researchers interested in climate data products but who are not climate modelers themselves. It can also be tailored to scientific researchers (or non-researchers) from other domains who need those products to assess the impacts of climate change on ecosystems, on economic activities, or for other applications.

The fact that such a large diversity exists makes it hard to think that only one diagnostic toolbox "to rule them all" can emerge. The focus of our talk will be to summarize the various needs we can anticipate, to describe diagnostic tool boxes we have at our disposal, and to identify potential gaps. Copernicus, H2020, and national initiatives, past or present, have largely contributed ideas and software packages on this topic. We will give an overview of related ENES activities, together with the implications and potential benefits for EGSF.

Canada Diagnostics
Tom Landry (CRIM)

Our computation framework is largely reliant on Birdhouse and its extensive WPS logging and monitoring capabilities. Most services and processes are called by a client web platform offering its users tools and workspaces. Data access and computation are also currently conducted separately on a Spark cluster at CRIM. Different jobs schedulers were used—for instance,

SLURM on HPC and RabbitMQ on hybrid clouds. In order to interoperate between systems, the ESGF CWT API is being evaluated.

New implementations are deployed on a staging TB infrastructure at CRIM, composed of a dozen Open Stack virtual machines. In that TB, CRIM conducted several technical interaction experiments in the OGC TB-13 EOC thread. Production systems are deployed on Ouranos infrastructure, on a bare metal server located at Calcul Quebec premises. Both CRIM and Ouranos infrastructure are tied to the CANARIE network, the national backbone of Canada's ultra-high-speed National Research and Education Network. All services and major constitutive elements of PAVICS are placed in CANARIE's science gateway registry, where a mandatory REST API documents the component, compiles usage statistics, collects reliability metrics, and notifies administrators of downtime.

We improved Birdhouse workflow capabilities and added several data validation tests. Specific workflows are executed regularly to test system integrity of both CRIM and Ouranos data and processing resources. Any change in the outcome of the integration workflows triggers a warning in the team's #pavics Slack channel, better exposing the system's state to a concentration of developers. To help manage and monitor its numerous Docker instances, PAVICS uses a lightweight Docker host management tool called Portainer. Queues are monitored and controlled with the Flower library.

CMIP6 Data Node Operations Team (CDNOT)

Sébastien Denvil (ENES/IPSL), sebastien.denvil@ipsl.jussieu.fr

A CDNOT has been appointed by the WIP and will include representatives from groups responsible for CMIP6 ESGF data nodes. The CDNOT is charged with implementing a federation of data nodes responsive to the requirements of the evolving CMIP6 process as articulated by the WIP.

The CDNOT scope covers functions and resource issues (hardware, networks, people) related to: installing, configuring and operating all the nodes and node services in the CMIP6 data federation and the policies and processes involved in both managing data on a data node (including acquisition, quality assurance, citation, versioning, and publishing) and providing access services.

The talk will summarize ongoing and foreseen actions and will describe the groups objectives.

Node Manager and Tracking/Feedback Notification

Sasha Ames (LLNL/AIMS), ames4@llnl.gov
Tobias Weigel (ENES/DKRZ), weigel@dkrz.de

Node Manager and Services Working Team (NWT)

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

After operating for several years without a node manager software component, this module has been redeployed with software stack in v2.5.x. This gives the ESGF a registry of components running at several sites and feeds information into the dashboard UI module. Moreover, there are several additional APIs running from the node manager, namely node status pages and one to distribute PID server (RabbitMQ) credentials. We aim to tighten security and to determine additional APIs for the node manager to contain. We will also give an update regarding a subscription service for user notification.

PID Services and Tracking/Feedback
Tobias Weigel (ENES/DKRZ), weigel@dkrz.de

The goal of the ESGF PID services is to record PIDs (Handles) for all files and datasets in CMIP6 and, more recently, for Obs4MIPs and CORDEX.

PID services for the ESGF consist of multiple components:

1 A message queue federation, based on RabbitMQ installations at DKRZ, IPSL, and PCMDI, which provides failover and load-balancing capacities in view of massive CMIP6 data object numbers.

2 A Python library (esgf-pid) used by the ESG Publisher to create and dispatch PID operation messages.

3 A Java servlet (queue consumer) which runs locally at DKRZ, connected to the PID servers (Handle server deployments) that execute the actual PID operations.

This talk will explain these components and their current status and embedding in the overall ESGF workflow, which also touches on concerns of CMOR and the CoG UI. We will also explain a special part of the PID services: the custom collection-building facility geared towards end users.

The talk will also showcase existing and future developments toward providing data tracking through PIDs and providing end users with tools that help them understand the state of data at hand (e.g., whether new versions are available). Relevance to other efforts such as the Research Data Alliance, EUDAT, and EOSC will be highlighted.

User Support and Documentation
Matthew Harris (LLNL/AIMS), harris112@llnl.gov

The ESGF Support Working Team is a collection of people from around the globe who aim to give ESGF users the best experience possible. This team includes representatives from Tier 1 and Tier 2 data and modeling centers, respectively. We have learned a lot over the last few years as the ESGF has had transitions and changes in both the wiki and website. We will cover our experiences and the direction our group should go to maintain the quality of user experience.

The ESGF Documentation Working Team is responsible for managing the documentation generated by other working teams. The team manages esgf.llnl.gov, which offers additional documents, such as for sponsors and committees, acknowledgments, governance, publications, tutorials, supported projects, wikis, and much more. We will cover the usability of our documentation and the direction moving forward using Sphinx and Read the Docs.

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)

- Energy Exascale Earth System Model (E3SM)
- 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)
- Input Datasets for Model Intercomparison Projects (Input4MIPs)
- European Space Agency's 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)
- Collaborative REAnalysis Technical Environment Intercomparison Project (CREATE IP)
- NASA NEX Global Daily Downscaled Climate Projections (NEX GDDP)
- NASA NEX Downscaled Climate Projections (NEX-DCP30)
- Coupled NEMS
- Climate Model Development Task Force (CMDTF)

Figure 14. *Major federated ESGF worldwide sites.*

D. Conference Participants and Report Contributors

Figure 15. *Conference attendees*.

Joint International Agency Conference and Report Organizers

- Dean N. Williams – Chair of the ESGF Executive Committee, U.S. DOE, LLNL
- Michael Lautenschlager – Co-Chair of the ESGF Executive Committee, ENES/DKRZ
- Luca Cinquini – ESGF Executive Committee, NASA/JPL
- Daniel Duffy – ESGF Executive Committee, NASA
- Sébastien Denvil – ESGF Executive Committee, IPSL
- Robert Ferraro – ESGF Executive Committee, NAS

ESGF Program Managers in Attendance

- Justin Hnilo – Chair of the ESGF Steering Committee, DOE Office of BER
- Sylvie Joussaume – ESGF Steering Committee, ENES
- Tsengdar Lee – ESGF Steering Committee, NASA
- Ben Evans – ESGF Steering Committee, NCI

Attendees and Contributors

No.	Name	Affiliation
1	Aloisio, Giovanni	CMCC/University of Salento
2	Ames Sasha	LLNL/AIMS
3	Anantharaj Valentine	Oak Ridge National Laboratory (ORNL)

No.	Name	Affiliation
4	Balaji Venkatramani	Princeton University
5	Ben Nasser, Atef	IPSL
6	Berger, Katharina	DKRZ
7	Bhatia, Karan	Google
8	Boutte, Jason	ESGF
9	Byrns, David	CRIM
10	Carriere, Laura	NASA
11	Cholia, Shreyas	LBNL
12	Christensen, Cameron	University of Utah
13	Cinquini, Luca	NASA/JPL
14	Dart, Eli	ESnet/LBNL
15	Denvil, Sébastien	Centre Nationale de la Recherche Scientifique (French National Centre for Scientific Research)/IPSL
16	Doutriaux, Charles	LLNL
17	Duffy, Daniel	NASA
18	Durack, Paul	LLNL
19	Dwarakanath, Prashanth	Linkoping University
20	Erickson, Tyler	Google, Inc.
21	Evans, Ben	NCI
22	Ferraro, Robert	NASA/JPL
23	Fiore, Sandro	Euro-Mediterranean Center on Climate Change Foundation
24	Gleckler, Peter	PCMDI
25	Glover, Rod	Pacific Climate Impacts Consortium
26	Greenslade, Mark	IPSL
27	Greguska, Frank	NASA/JPL, California Institute of Technology
28	Harr, Cameron	LLNL

No.	Name	Affiliation
29	Hiebert, James	Pacific Climate Impacts Consortium/University of Victoria
30	Hill, William	LLNL
31	Hollweg, Heinz-Dieter	DKRZ
32	Huard, David	Ouranos
33	Inoue, Takahiro	Research Organization for Information Science and Technology
34	Kershaw, Philip	STFC/CEDA
35	Kim, Soo	LLNL
36	Kindermann, Stephan	DKRZ
37	Kolax, Michael	SMHI
38	Lacinski, Lukasz	ANL/University of Chicago
39	Landry, Tom	CRIM
40	Lee, Ji-Woo	LLNL
41	Levavasseur, Guillaume	Institute Pierre Simon Laplace
42	Lu, Kai	Linkoping University
43	Maxwell, Thomas	NASA Goddard Space Flight Center
44	McCoy, Renata	LLNL/AIMS
45	Nadeau, Denis	LLNL/AIMS
46	Nikonov, Serguei	Geophysical Fluid Dynamic Laboratory
47	Pagé, Christian	Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique
48	Peterschmitt, Jean-Yves	Climate and Environment Sciences Laboratory (LSCE)/IPSL
49	Pobre, Alakom-Zed	NASA Goddard Space Flight Center
50	Potter, Gerald	NASA
51	Raju, Bibi	PNNL
52	Richards, Clare	Australian National University
53	Serradell, Kim	Barcelona Supercomputing Center

No.	Name	Affiliation
54	Shaheen, Zeshawn	LLNL
55	Stephens, Ag	Science & Technology Facilities Council (U.K.)
56	Taylor, Karl	LLNL/PCMDI
57	Tsengdar, Lee	NASA
58	Vahlenkamp, Hans	GFDL
59	Weigel, Tobias	DKRZ
60	Zhang, Chengzhu	LLNL

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 ESGF data science domain. This year, the ESGF Executive Committee decided on 15 winners of the ESGF Achievement Awards. These awards recognize dedicated members of the ESGF community who contribute nationally and internationally to federation efforts. Award recipients exemplify the community's spirit and determination to succeed. The Executive Committee's recognition of these members' efforts is a small token of appreciation and does not diminish the efforts of others who also work hard to make the ESGF a success. Achievement Award winners for 2017 include:

- **William Hill** and **Alexander "Sasha" Ames** (LLNL/AIMS, U.S.) won a group award for taking over the leadership of the ESGF Installation Working Team, filling a vacuum in one of the most critical areas of the ESGF development efforts. Through their perseverance, skills, and dedication, they managed to get this effort back on track, resulting in successive releases of ESGF software to support data forthcoming from CMIP6 and many other important community projects. In addition, William started the conversion of the current installation of Shell-based software to a new Python-based framework, which will make installation easier, less error-prone, and more extensible.
- **Eli Dart** (Energy Sciences Network [ESnet], U.S.) won an award for his unparalleled efforts to span seven international network organizations to form the ICNWG. Under the ICNWG banner, Eli leads a network team responsible for the networks and computers that enable the geographically distributed ESGF peer nodes to communicate and distribute tens of PBs of data to the global climate community. In addition, his team improves scientific productivity through effective, efficient use of networks to exchange data between CMIP6 Tier 1 replication data centers (i.e., LLNL, CEDA, DKRZ) and by providing network and transfer diagnostic tests needed to monitor network performance. This year, Eli coordinated the large-scale CMIP6 replication process between Tier 1 sites. This includes the integration of the Synda automated data transfers and publication process with ESGF DTNs and disparate network infrastructures.

- **Alessandra Nuzzo**, **Maria Mirto**, **Paola Nassisi**, and **Sandro Fiore** (CMCC, Italy) won a group award for developing the new ESGF dashboard. The dashboard shows up-to-the-moment usage demographics and a statistical overview of data use, such as total number of registered users by continent and country, downloads by continent and country, and total number of datasets and data volume. Capturing federated usage metrics, the dashboard provides a rich set of charts and reports through a web interface, allowing users and system managers to visualize the status of the ESGF infrastructure through smart, user-friendly web gadgets. The CMCC group addressed key challenges such as communicating the most important information in a straightforward way and allowing different users to view specific details simultaneously. Without their critical work in displaying automatic real-time data usage, scientists would have no clear way to determine the importance of their projects' data contributions to the community.
- **Jason Boutte** (LLNL/AIMS, U.S.) won an award for completion of the server-side compute process in time for CMIP6 deployment. To assist in the development of remote computing (i.e., aggregation, subsetting, regridding), Jason modularized compute capabilities to allow use of multiple backend analysis engines within the ESGF framework. That is, the CWT end-user API is infused with community analysis and visualization tools, such as CDAT, the Earth Data Analysis System (EDAS), and Ophidia. Through his efforts, on-demand computing at CMIP6 Tier 1 sites will provide new ways for managing complex analytical workflows and allow scientists to more readily share information and collaborate with others. Server-side computing can be accessed via command line, Jupyter Notebook, or the ESGF CoG web browser UI.
- **Philip Kershaw**, **Matt Pryor**, **William Tucker** (ENES/CEDA, U.K.), **Lukasz Lacinski** (ANL, U.S.) and **Sasha Ames** (LLNL/AIMS, U.S.) won a group achievement award for transitioning the ESGF authentication and authorization services from the current OpenID-based system, to the new OAuth2 industry standard. This was truly a collaborative effort: Phil led the system design and protocol specification; Matt implemented the ESGF OAuth2 server and the SLCS; Lukasz implemented the ESGF-OAuth client, a replacement for the current ORP; William wrote the TDS filter to trigger the full workflow upon data download; and Sasha led the integration of the new security components into the ESGF installer. This effort is one of the best examples to date on how ESGF developers from different institutions and continents can proficiently work together to accomplish a long-term goal.
- **David Byrns** (CRIM, Canada), as a CRIM senior advisor and an ESGF technical lead, has been instrumental in the creation of a new Platform for the Analysis and Visualization of Climate Science, known as PAVICS. David supervised a twelve-person team, including four new junior developers, in a Kanban process and played a pivotal role in DevOps, supporting both his peers and his client. In PAVICS, he led the efforts in software architecture, documentation, application packaging and deployment, workflows, platform integration, and system diagnostics. His work enabled the adoption and adaptation of Birdhouse in PAVICS, as well as the testing of novel workflows in EO and LIDAR. David's remarkable efforts drastically improved PAVICS for successful, sustained sharing with the ESGF in the coming years.

- **Guillaume Levavasseur** (IPSL), as an IPSL research engineer, received an award for his sustained contribution to many ESGF aspects over the past 5 years. In 2017 he has been supporting continuously the replication and versioning team so that they can exercise and learn how to use the Synda replication software in their environment. In 2017 he released the esg-prep python program that will help dozens of data managers around the globe to easily version their datasets. And finally in 2017, as one of the fathers of the errata service (together with Atef Ben Nasser), he supported the release of the first version of the errata service that will enable the community to reach an unprecedented level of traceability and provenance of the CMIP data repository.
- **Michael Lautenschlager** (DKRZ, Germany) won an award for his longstanding commitment to the ESGF. Michael serves as the ESGF Co-Chair and a member of the CMIP6 WIP. His work and leadership are critical to the overall success of the ESGF, including helping to incorporate (PID tracking integration, citation information, and errata annotation—to name only a few.

In addition, **Dean N. Williams** was inducted into the **ESGF Hall of Fame** for his longtime leadership and commitment to the ESGF. Dean has been the ESGF Principal Investigator and recognized leader since the beginning of the project, when it was funded by DOE over 19 years ago. Throughout the years, Dean has been the most critical force and steadfast supporter of the ESGF, in both good and difficult times. There would be no ESGF without Dean. Because of his leadership, strategic vision, and everyday involvement, the collaboration has evolved from a small DOE research project to a global partnership of dozens of institutions that work together to support climate research. Nineteen years ago, it would have been impossible to predict this success. All of us are in his debt, as is the community. Thank you, Dean, and please keep doing what you are doing for many years to come.

Finally, in November 2017, the ESGF was recognized by *R&D Magazine* for an R&D 100 Award. These "Oscars of invention" are awarded annually to only 100 products and technologies in the scientific community with commercial potential. This prestigious award is among the highest honors given to inventors, and the ESGF Executive Committee acknowledges and thanks everyone who has contributed to the ESGF's success.

F. Acknowledgments

The 2017 ESGF FF2F Conference organizers wish to thank national and international agencies for providing travel funding for attendees to join the conference in person, the U.S. DOE's 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 Holly Auten for taking the detailed conference notes used in this report.

ESGF development and operation continue to be supported by the efforts of principal investigators, software engineers, data managers, projects (e.g., CMIP, E3SM, CORDEX, MIPs in general, and many others), and system administrators from many agencies and institutions worldwide. Primary contributors to these open-source software products include: ANL;

Australian National University; British Atmospheric Data Centre; Computer Research Institute of Montreal; Euro-Mediterranean Center on Climate Change; German Climate Computing Centre; Earth System Research Laboratory; GFDL; Goddard Space Flight Center; Institut Pierre-Simon Laplace; JPL; Kitware, Inc.; National Center for Atmospheric Research; New York University; ORNL; Los Alamos National Laboratory; LBNL; 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. NASA, U.S. National Oceanic and Atmospheric Administration, U.S. National Science Foundation, Infrastructure for the ENES, and the Australian NCI provide major funding for the ESGF community hardware, software, and network infrastructure efforts.

G. Acronyms

Acronym	Definition
AIMS	Analytics and Informatics Management Systems—Program at LLNL enabling data discovery and knowledge integration across the scientific climate community (aims.llnl.gov).
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).
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).
API	Application Programming Interface (en.wikipedia.org/wiki/Application_programming_interface/).
ARM	Atmospheric Radiation Measurement
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/).
C3S	Copernicus Climate Change Service
CCI	European Space Agency's Climate Change Initiative
CDAT	Community Data Analysis Tools
CDI	EUDAT's Collaborative Data Infrastructure

Acronym	Definition
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.
CDNOT	Coupled Model Intercomparison Project Data Node Operations Team
CDP	Community Diagnostics Package
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).
CIM	Climate Information Model
Climate4Impact	Web portal that enables visualization of climate model data sets targeted to the climate change impact assessment and adaptation communities (climate4impact.eu/impactportal/general/).
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).
CNN	Convolutional Neural Network
CoG	Composable Graphical User Interfaces—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

Acronym	Definition
CREATE	Collaborative REAnalysis Technical Environment—NASA project that centralizes numerous global reanalysis data sets 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/).
CREATE-V	Collaborative REAnalysis Technical Environment Visualization Web Tool
CRIM	Centre de Recherche Informatique de Montréal—Computer Research Institute of Montréal (crim.ca)
CV	Controlled Vocabulary
CWT	ESGF Compute and Data Analytics Working Team
DKRZ	Deutsches Klimarechenzentrum (German Climate Computing Centre)—Provides HPC platforms and sophisticated, high-capacity data management and services for climate science (dkrz.de).
DL	Deep Learning
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).
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 data sets wherever they might be located within the distributed ESGF archive.
DSC	Data and Service Challenge
DTN	Data Transfer Node—Internet location providing data access, processing, or transfer (fasterdata.es.net/science-dmz/DTN/).
E3SM	Energy Exascale Earth System Model—DOE's effort to build an Earth system modeling capability tailored to meet the climate change research strategic objectives (https://e3sm.org/).

Acronym	Definition
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).
EOC	Earth Observation Clouds
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/).
ESFRI	European Strategy Forum on Research Infrastructures
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).
ESMValTools	Earth System Model eValuation Tools
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).
EU	European Union
EuBra-BIGSEA	Europe–Brazil Collaboration of Big Data Scientific Research Through Cloud-Centric Applications (eubra-bigsea.eu).
EUDAT	European Data Infrastructure
F2F	Face to Face
GEF	Generic Execution Framework
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).
GIS	Geographic Information System

Acronym	Definition
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/).
GUI	Graphical User Interface
HAPI	HArvester Provenance Interface
HDF	Hierarchical Data Format—Data model, library, and file format for storing and managing a wide variety of high-volume and complex data types
HPC	High-Performance Computing
HPSS	High-Performance Storage System—Modern, flexible, performance-oriented mass storage system (hpss-collaboration.org).
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, and Access Working Team
IdP	Identity Provider
IDX	A type of multiresolution file format
Input4MIPs	Input Datasets for Model Intercomparison Projects—A database used for preparing forcing datasets and boundary conditions for CMIP6 (pcmdi.llnl.gov/projects/input4mips/).
I/O	Input/Output
IOOS	Integrated Ocean Observing System
IPCC	Intergovernmental Panel on Climate Change—Scientific body of the United Nations that periodically issues assessment reports on climate change (ipcc.ch).

Acronym	Definition
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).
JPL	Jet Propulsion Laboratory—A federally funded research and development laboratory and NASA field center in Pasadena, California (jpl.nasa.gov).
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).
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).
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/).
LSTM	Long Short-Term Memory
MIP	Model Intercomparison Project
ML	Machine Learning

Acronym	Definition
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).
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).
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).
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).
netCDF	Network Common Data Form—Machine-independent, self-describing binary data format (unidata.ucar.edu/software/netcdf/).
NN	Neural Network
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).
NRCan	Natural Resources Canada
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/).
OGC	Open Geospatial Consortium—International nonprofit organization that develops quality open standards to improve sharing of the world's geospatial data (opengeospatial.org).
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
PATX	CRIM's Speech and Text Research Team
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).
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
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.
SAML	Security Assertion Markup Language

Acronym	Definition
SAR	Synthetic Aperture Radar
SLCS	Short-Lived Credential Service
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).
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	Testbed
TDS	THREDDS Data Server
TESD	CRIM's Emerging Technologies and Data Science Research Team
THREDDS	Thematic Real-Time Environmental Distributed Data Services—Web server that provides metadata and data access for scientific data sets using a variety of remote data access protocols (dataone.org/software-tools/thematic-realtime-environmental-distributed-data-services-thredds/).
UI	User Interface
UV-CDAT	Ultrascale Visualization–Climate Data Analysis Tools—Provides access to large-scale data analysis and visualization tools for the climate modeling and observational communities (uvcdat.llnl.gov).
vCDAT	Visual Community Data Analysis Tools
VISI	CRIM's Vision and Imaging Research Team
ViSUS	Visualization Streams for Ultimate Scalability
W3C	World Wide Web Consortium—An international community that develops web standards.

Acronym	Definition
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).
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).
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.
WPS	Web Processing Service—Provides rules for standardizing inputs and outputs (requests and responses) for geospatial processing services (opengeospatial.org/standards/wps/).