8th Annual Earth System Grid Federation

December 2018

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.

8th Annual Earth System Grid Federation Face-to-Face Conference Report

December 3–7, 2018
Washington, DC, 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
Michael Lautenschlager (Co-Chair; German Climate Computing Centre)
Ghaleb Abdulla (DOE Lawrence Livermore National Laboratory)
Sasha Ames (DOE Lawrence Livermore National Laboratory)
Venkatramani Balaji (Princeton University)
Laura Carriere (NASA Goddard Space Flight Center)
Luca Cinquini (NASA Jet Propulsion Laboratory)
Sébastien Denvil (Institut Pierre-Simon Laplace)
Ben Evans (National Computational Infrastructure)
Robert Ferraro (NASA Jet Propulsion Laboratory)
Philip Kershaw (CEDA Science and Technology Facilities Council)
Tom Landry (CANARIE Centre de Recherche Information de Montréal)

ESGF Steering Committee
Justin (Jay) Hnilo (DOE, U.S.)
Sylvie Joussaume (ENES, Europe)
Tsengdar Lee (NASA, U.S.)
Clare Richards (NCI, Australia)
Dean N. Williams (Lawrence Livermore National Laboratory; Ex-officio member)

Disclaimer

Lawrence Livermore National Laboratory is operated by Lawrence Livermore National Security, LLC, for the U.S. Department of Energy, National Nuclear Security Administration under Contract DE-AC52-07NA27344. LLNL-TR-772409.

Contents

1. Executive Summary

Members of the Earth System Grid Federation (ESGF) gathered in Washington, DC, on December 3–7, 2018, for the 8th Annual ESGF Face-to-Face (F2F) Conference. The event packed 40 presentations, several plenary sessions, a poster session, guest speakers, an awards ceremony, and an Executive Committee (XC) meeting into the week. ESGF working teams (WTs) conduct daily operations remotely all over the world. Although most teams meet online regularly, in-person interaction is incredibly important for cementing relationships and evaluating progress.

Addressing "big data" challenges in Earth system research, ESGF is an international collaboration of computer scientists, data scientists, and climate researchers. The federation houses an enormous database of global observational and simulation data—more than 5 petabytes—and manages the high-performance computing hardware and software infrastructure necessary for scientific climate research. In the nearly two decades since its launch, ESGF has grown to serve 25,000 users on 6 continents.

ESGF's 2017 conference focused mainly on preparation for the release of the Coupled Model Intercomparison Project Phase 6 (CMIP6) dataset. CMIP6 presented a significant test of ESGF's infrastructure for the expected 20 petabytes of model output. Multiple components across the service stack needed to function under this level of stress—services such as publishing, search, download, and replication (i.e., moving data from one ESGF center to another). Incremental data challenges (DCs) in 2018 verified the integrity and robustness of ESGF infrastructure prior to the midsummer CMIP6 launch. Development teams reported on these readiness activities during the 2018 conference.

As part of CMIP6 readiness, the Input4MIPs initiative provided a key improvement in "forcing" dataset consistency when comparing with previous CMIP phases. Input4MIPs collects, archives, and documents climate datasets to support the coordinated modeling activities. ESGF hosts Input4MIPs data alongside CMIP datasets, enabling climate researchers to evaluate climate models with uniform standards under the same conditions. Lawrence Livermore National Laboratory's (LLNL's) Paul Durack recently won the World Climate Research Programme (WCRP) Data Prize for his leadership of Input4MIPs.

Similarly, the Obs4MIPs initiative began planning for CMIP6 in 2016. Co-led by LLNL and NASA's Jet Propulsion Laboratory (JPL), and hosted on ESGF servers, this project established a database used by the CMIP modeling community for comparing satellite observations with climate model predictions. In 2018, the Obs4MIPs team implemented several enhancements in data indicators and integration along with a prototype of color-coded quality indicators.

One crucial achievement during 2018 was the beta version 3.0 of the ESGF software stack installer, released during the conference. The Installation Working Team closed more than 200 issues for this version, and a conference poster detailed the team's efforts to stand up a Jenkins automation server for validating changes to installer code. The Python-based beta release addresses several long-standing problems such as a lack of error handling, lack of extensibility, and a complicated installation process.

Another progress report came from the Identity and Access Working Team, who demonstrated OAuth single sign-on integration to increase security and ensure proper user permissions. This work involved use cases for accessing data without authentication, using OAuth credentials with other platforms, handling different versions of OAuth, embedding the OAuth certificate in wget scripts, and confirming the OAuth access token workflow.

Most WTs made some headway in "containerizing" their areas of the ESGF infrastructure. For instance, a prototype of the search service was implemented with Docker and Kubernetes on the Solr Cloud. Similarly, the compute WT demonstrated a new container-based design for better scalability of server-side distributed computing.

Containerized architecture is compatible with Cloud deployment, which is another key effort under way among many ESGF teams. In addition to the search service prototype, ESGF's research partners at NASA plan to move their data into the Amazon Cloud for easier model intercomparisons with a containerized stack. Other research centers are also considering Cloud computing to support their ESGF nodes—namely, leveraging on-demand capabilities with some simplification of maintenance tasks—although European partners face unique challenges without native Cloud providers on their continent.

In 2019, ESGF WTs plan to complete the rollout of the installer and OAuth access; evaluate a trial run of the new search service; develop additional containerized packages; upgrade the user interface (UI) with standalone search; and stabilize CMIP6 operations. Additionally, compute node challenges are slated for spring and summer to finish containerizing the web processing service (WPS) application programming interface (API). These expanded capabilities will help ensure sustainability for eventual requirements from future CMIP datasets, user growth, exascale computing, machine learning (ML) technologies, and more.

Additionally, the XC plans to schedule additional F2F meetings and maximize ESGF's presence at relevant international conferences. Certain topics, such as infrastructure maintenance and opportunities with new scientific domains, require further deliberation.

Introduction and Keynotes

The conference 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. More than 60 professionals from 10 countries gathered together to share their knowledge and experiences gained over the past several years.

This year's emphasis was placed on compute services, service interoperability, topics from the WTs, and future ESGF architecture. In addition to presentations, plenary discussions on the day's topic were scheduled together with poster sessions at the end of each day. The conference closed with parallel WT meetings to coordinate work in the coming year and a closing plenary session to review the conference findings and to discuss the future conference schedule and focus.

Conference agenda, abstracts, and presentations are available from the ESGF website. The online version of this conference report will also be available from the ESGF Reports page.

Keynotes from representatives of the ESFG funding agencies and steering committee (SC) members outlined international cooperation, funding commitments, and potential future directions for the ESGF. On the European side, the third phase of the infrastructure project for ENES, known as IS-ENES3, started in January 2019 and continues until December 2022. IS-ENES3 provides continued support for ESGF from the European Union. Recommendations for infrastructure strategy were to prepare for exascale computing and DCs, maximize data transfer bandwidth, train all levels of experts, and two new topics: model evaluation and sustainable infrastructure.

On the U.S. side, the ESGF is one of the most important initiatives in LLNL's portfolio. New compute resources funded by the DOE will be installed to provide compute services at the LLNL node for subsetting and to aggregate CMIP data efficiently and reduce downloaded volume for remote analysis. NASA is making significant effort to move data into the Cloud and encourages the ESGF to explore this architecture change for the global climate data federation. The general consensus is that the ESGF compute services should integrate data analytics packages in support of model evaluation, given that the data volume is rapidly approaching the point where downloading to local compute resources becomes impossible. Another suggestion to consider is that the ESGF should coordinate at national and international levels with merging interdisciplinary research infrastructure activities. This coordination may open pathways for sustainable funding for the future.

The XC would like to thank each of conference attendees as well as those who could not attend but contributed to and/or supported the development of the ESGF software stack. This 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 commitments 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.

The ESGF-XC chair was unable to attend this year's conference. However, he forwarded his regards to the audience and hopes to be back soon:

Dear Colleagues:

My apologies for not attending this year's 2018 ESGF Conference. However, I know you are in great hands with the SC and XC committees. I am proud to be a part of this international team and you should be proud of another great year of contributions to the ESGF and climate change community efforts! Deadlines are being met and climate scientists are using ESGF to address one of the most pressing challenges of the day and the future. Without you NONE of this would be possible.

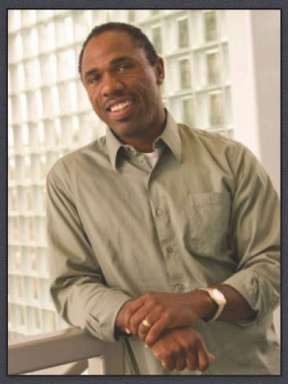
Again, I thank each and everyone of you for your contributions to ESGF and making it what it is today, "The leading climate simulation distributed data warehouse".

I hope to see everyone soon as I recover from my illness.

Best regards,

Dean N. Williams

ESGF Chair



Conference Findings

The federation's response to, and progress on, findings from the 2017 Seventh Annual ESGF F2F Conference Report are summarized above at the beginning of the Executive Summary.

Key findings from the 2018 conference were reviewed by the XC and are summarized here. These priorities are united by the need to be proactive with new and evolving architectures and technologies.

1. **Containerized architecture.** Now heading into its third year, the Containers Working Team is implementing ESGF's previously defined strategy of designing a new architecture based on containerization technology. Containers enable easier, more flexible, and scalable deployment and installation as well as Cloud compatibility. The team's version 1.0 release in September 2018 was based on Docker, Kubernetes, and Helm technologies. These development efforts will continue with updates to existing modules, containerization of additional modules, and larger scale test deployments.
2. **Compute node challenges.** Last year's successful DCs inspired a new set of challenges for the coming year. The Compute Working Team (CWT) will conduct iterative deployments and tests to stand up multiple production-ready, officially certified ESGF compute nodes. This effort will include data selection, security scans, API development, scalability tests, documentation updates, and regular status reports.

3. **OAuth 2.0 deployment.** With OAuth now in use at three sites and related services bundled with the installation, ESGF is primed to migrate fully from OpenID to OAuth. Planned work in the coming year includes implementation of OAuth functionality that improves, integrates, or otherwise supports client registration, access tokens, and single sign-on.
4. **CMIP6 data replication.** Now that CMIP6 operations are stabilizing, the next objective is to confirm timely, reliable distribution of CMIP6 data to users—i.e., replication across the Tier 1 ESGF sites. This work involves upgrading the Synda code base with Globus integration, monitoring and optimizing data transfer performance, and prototyping persistent identifier (PID)–based replication management.
5. **Next-generation search services.** The primary search functionality enhancements planned in 2019 are implementation of a standalone search API and development of a Solr-based "super-index." These efforts will provide greater scalability as the federation's datasets grow in size and complexity as well as offer resilience (less inter-dependency) at each site.
6. **Scalability on multiple fronts.** ESGF WTs have identified several opportunities for scalability improvements across the federation's architecture including publication processes, node management, notification and tracking services, errata services, and the user dashboard. In addition, after issuing a beta release during the conference, the Installation and Software Security Working Team expects to roll out version 3.0 of the software stack installer during 2019.
7. **ML tools.** Increased demand for sophisticated (and faster) data analysis makes ML and other data science techniques valuable tools in ESGF's growth trajectory. With the demonstration of neural networks (NNs) that detect and locate extreme climate events, the Machine Learning Working Team is investigating and refining the models' accuracy.

2. Technology Developments

The ESGF WTs integrate the real infrastructure development and cover the technological aspects. Individual WT reports are provided in the Appendix.

Team	Team Leads and Funding Agencies / Institutions	Description
1. User Interface, Search, and Dashboard Working Team	Luca Cinquini (JPL), Guillaume Levavasseur (IPSL), and Alessandra Nuzzo (CMCC)	Improve ESGF search and data cart management and interface; ESGF search engine based on Solr 5; discoverable search metadata; statistics related to user metrics
2. Compute Working Team (CWT)	Charles Doutriaux (LLNL)	Develop the capability to enable data analytics within ESGF
3. Identity, Entitlement, and Access (IdEA) Working Team	Philip Kershaw (CEDA) and Lukasz Lacinski (ANL)	Identity management and access control to enable resources (data and compute) to have appropriate access restrictions

Team	Team Leads and Funding Agencies / Institutions	Description
4. Installation and Software Security Working Team	Sasha Ames (LLNL), William Hill (LLNL), and Prashanth Dwarakanath (LiU)	Install components of the ESGF software stack; security scans to identify vulnerabilities in ESGF software
5. Containers Working Team	Luca Cinquini (JPL) and Sebastien Gardoll (IPSL)	Design and implement a new ESGF architecture based on containerization technologies
6. International Climate Network Working Team 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. ESGF Services: Node Manager and Tracking / Feedback Notification Working Team	Sasha Ames (LLNL) and Tobias Weigel (DKRZ)	Manage ESGF nodes and node communications
8. Publication, Quality Control, and Metadata Working Team	Sasha Ames (LLNL) and Katharina Berger (DKRZ)	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	Vacant (includes representatives from Tier 1 data centers, Tier 2 modeling centers, and the above WTs)	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), Sandro Fiore (CMCC)	Research in the applicability of various ML techniques and development of tools/analysis capabilities for domain scientists

3. Implementation Roadmap

The implementation roadmap highlights major topics that crystallized in the presentations and discussions over the conference and in the XC. They correspond with the more detailed WT reports found in the Appendices.

Short Term (2019)

New Developments

- The compute node challenge task is due on June 2019. However, we will be working on extensions and adding new capabilities beyond the current challenge.

    - The current compute challenge is due for release to public community by end of June 2019. This will involve deploying several compute nodes ready for production with limited set of core services (as agreed upon during the F2F meeting).
    - During the second challenge, we will extend the compute node's set of core services, deploy new compute nodes, and increase the number of sites with new capabilities. We will also work on an early prototype of federated compute services.
- Explore emerging technologies in preparation for the 2019 workshop on ESGF future architecture. This will be an investigative effort to understand the issues surrounding these technologies and how to apply them to climate data.
    - Distributed computing using Dask/Spark
    - Cloud computing
    - ML
    - Open software libraries from numfocus
    - Object storage

ESGF Operations with Emphasis on CMIP6

- Upgrade ESGF UI and create standalone search API. The decommissioning of CoG (i.e., composable graphical UIs) at the University Corporation for Atmospheric Research in the middle of this year imposes constraints on the ESGF UI in the short and medium terms. An additional requirement is to open the ESGF for community-based access interfaces.
    - Update to Python 3
    - Front-end modernization and standalone search interface as CoG replacement (maybe Jupyter), continues to 2020
    - Solr super-index search service
- Release: ESGF installer V3.0 operational rollout
    - Management of software package dependencies (migrate to Conda environments)
- Update and maintain ESGF/CMIP6 user support and documentation
    - Implementation and operation of the ESGF/CMIP6 Help Desk is organized in cooperation with the Working Team on Coupled Modelling, or WGCM, Infrastructure Panel (WIP). Experiences from CMIP5 infer that user requests cannot easily be separated into ESGF or CMIP6. First-level support will cover both topics and separate for the second-level support. The Help Desk will include online documentation and a Frequently Asked Questions section.
    - Reactivate the User Support and Documentation WT
- Stabilize CMIP6 operations
    - Globus integration (installer, Synda, ESGF UI)
    - Software maintenance and upgrades of different elements of the ESGF software stack
- Release: OAuth operational rollout
    - This has a specific application focus of supporting the compute node challenge. The goal is to enable token-based authentication to resources in place of short-lived certificates.
- Maintain ESGF publisher environment

    - Python 3 conversion
    - CMIP6 improvements (e.g., PID)

Medium Term (2020–2021)

New Developments

- Investigation of ML tools including the use cases, objectives, questions, scalable computer systems related work, statistical analysis, data retrieval and sub-setting, and ML algorithms.
- Provenance management to better track versions of data and replication history.
- Integration of and into Cloud environments, especially those developed in the context of Pangeo as well as Cloud-based network common data form (NetCDF) file storage solutions like Zarr.

Improvement of ESGF Infrastructure and Services

- Develop additional containerized packages alongside the containerized ESGF operational rollout.
- Support OAuth access tokens for accessing restricted resources.
- Implement services to better manage controlled vocabularies for ESGF projects (e.g., how to describe data with ontologies). This requires domain experts working with ESGF
- Develop an OpenSearch interface for ESGF search.
- Implement ESGF publisher improvements like parallel execution, decouple QC, improve for other data types.
    - CMIP6 will need either a climate model output rewriter (CMOR) tag or a valid NetCDF format file. The publisher will accept any NetCDF file.
    - Improvements (e.g., more robust, easy to use, parallel, other data types).
- Other ESGF maintenance.

Future of ESGF

- Future ESGF architecture
    - In continuation of the of the planned workshop on emerging technologies in 2019, the future ESGF architecture has to be drafted, discussed, and ideally decided. Additionally, 2020 should leave more room for the structural effort because the most active phase in CMIP6 is expected in 2019.
- Sustainability of ESGF
    - With respect to funding and the transition from implementation to an operational infrastructure, the role of ESGF in global climate research has to be determined. Reliability of a research infrastructure is closely related to sustainability. The current funding situation covers certain aspects ESGF development but has few options for maintenance and operations.

Acknowledgments

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—both within the U.S. DOE's Office of Science—NOAA, NASA, NSF, ENES, NCI, and CANARIE. Support also comes from other national and international agencies and private industry.

Appendices

A. Working Team Reports

1. User Interface, Search, and Dashboard Working Team

Progress Report

The following activities took place in 2018:

- ESGF search services. Three major releases of the ESGF search services occurred (v4.15, v4.16, v4.17), which included bug fixes and a general upgrade of third-party libraries to address recently discovered software vulnerabilities. The most important work was updating all ESGF nodes to run the Solr 6.x release line, since the previous 5.x line is no longer supported and had become insecure. Because Solr releases are backward but not forward compatible, this required a coordinated effort from all nodes in the federation to prevent long interruptions in catalog replication.
- CoG. Several updates of the CoG UI took place to provide better searching capabilities and display of results for CMIP6 data and to hyperlink to the external documentation websites developed by ES-DOC (Earth system documentation) and the World Data Center for Climate (e.g., errata pages, PID pages, digital object identifier pages). This resulted in four major releases (v3.10 through v3.13) and several minor releases.
- Dashboard. Two major releases of esgf-dashboard (v1.5.14 and v1.5.21) and one for esgf-stats-api (v1.0.6) have been issued to ensure a more accurate data usage statistics delivery (by allowing to distinguish the downloads by users) and to provide a specific view for the CMIP6 project in terms of data downloads and published data. A REST (representational state transfer) API service has been deployed on the collector node at Centro Euro-Mediterraneo sui Cambiamenti Climatici (the Euro-Mediterranean Center on Climate Change, or CMCC) to provide, besides the web UI, the federated statistics in a programmatic way. Graphical restyling along with new views and metrics have been released, together with the option to export a file of comma-separated values from the graphical widgets for further analysis. Additionally, the installation of the esgf-dashboard has been fixed on CentOS 7, and documentation and configuration info have been produced.
- Solr Cloud index node. A new proposed architecture for the ESGF search services was prototyped and unveiled at the conference. This architecture is based on deploying a single ESGF "super-index" node on the commercial Cloud (for high availability), harvesting all metadata catalogs from individual ESGF nodes, and pointing all clients (e.g., CoGs, other UIs, scripts) to this instance. Internally, the super-index is based on Solr Cloud, running as a system of individual Docker containers on a distributed Kubernetes cluster. This architecture would be highly scalable—because of the use of Solr Cloud and because it is hosted on a single scalable environment, as opposed to all ESGF nodes in the federation—and would greatly facilitate upgrading to new versions of Solr as previous versions become obsolete and insecure. A prototype installation was

deployed on Amazon Web Services and has been faithfully tracking the publication of data throughout the ESGF federation for several months.

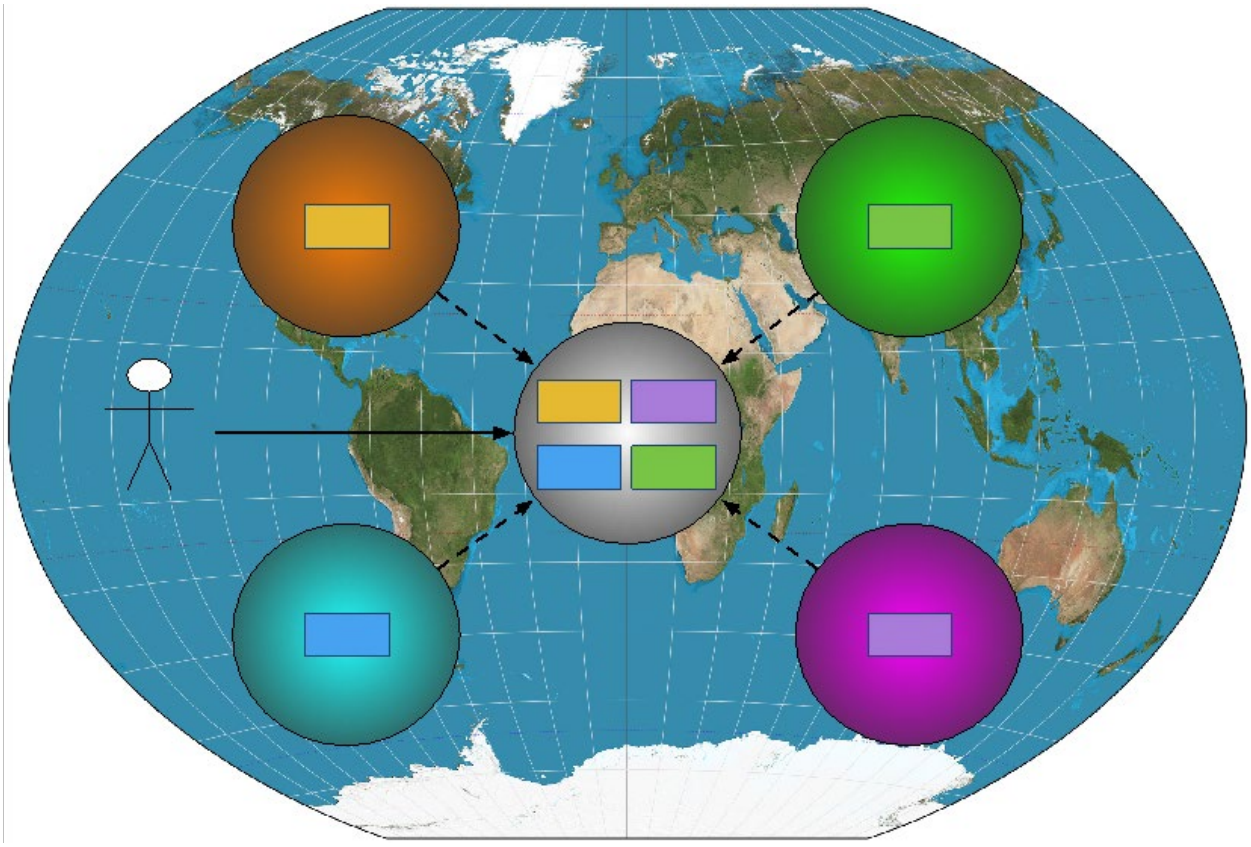


Figure 1. *Proposed next-generation ESGF search architecture, where metadata catalogs are harvested from individual ESGF nodes into a "super-index" node.*

Future Roadmap

The following work is planned for 2019:

- CoG. The ESGF-XC and ESGF-SC need to make some critical decisions regarding the future directions of the ESGF UI. The current CoG code base has become very big, containing a lot of functionality not directly needed by ESGF. A decision must be made as to whether decompose and upgrade the current code base (based on Python and Django), or completely replace it with one or more new web applications. Once a decision is made, appropriate funding and human resources must be allocated.
- Solr Cloud super-index. Pending final approval by the ESGF-XC, this WT plans to release a stable version of the ESGF super-index on the Amazon Cloud, and instrument it to periodically harvest metadata from all individual ESGF index nodes. This work can proceed in parallel with normal ESGF publishing and replication operations and will not disrupt the management of CMIP6 data.

- Dashboard. Several activities will be supported by IS-ENES3 as part of the core data services foreseen in the project. In particular, (1) there will be a discussion with the CWT to evaluate the integration of metrics about computation into the dashboard; and (2) net statistics from the PerfSONAR (Performance Focused Service Oriented Network Monitoring Architecture) service will be integrated into the dashboard UI. Also, (3) the downloads made by Synda for replication will be visualized into the dashboard UI and distinguished from the regular ones; (4) the dashboard will start collecting statistics coming from other download services (e.g., wget, GridFTP); and (5) new views for specific projects (e.g., CORDEX [Coordinated Regional Climate Downscaling Experiment], E3SM [Energy Exascale Earth System Model], CMIP3, Input4MIPs) will be deployed. The libraries included into the software packages will be periodically checked and updated.
- Replicas and versioning synchronization. Discrepancies on datasets versions and their replicas between sites will persist with the Solr Cloud architecture. This causes harm for the end user that can, for instance, still download replicas of retracted data because of critical issues. A script made for CMIP5 data will be extend to CMIP6 to synchronize the dataset versions across the ESGF indexes harvested by the Solr Cloud super-index. A strategy has to be defined to also synchronize the data retractions across the index nodes to avoid the super-index serving data with critical issues.

2. Compute Working Team

Progress Report

This was a growth year for the CWT, as many institutions developed solutions that are either production ready or nearing it. The following activities took place in 2018:

- LLNL polished the Python-based compute_api and is nearing completion of a Docker/Kubernetes-based architecture that will hopefully become the de facto “ESGF compute-node” (though centers will still be free to deploy a compute node any way they want). LLNL also developed a JavaScript-based UI to orchestrate workflows.
- NASA’s Goddard Space Flight Center (GSFC) re-implemented its Earth Data Analytics Services backend as a Python/DASK-based solution. This implementation is much easier for developers to build on while offering performance comparable to the original MPI/Spark implementation for the typical size of data processed by DASK.
- CMCC readied their Ophidia backend and performed many optimizations, notably in the input/output area. A new release, expected by early 2019, will represent the basis for a CWT-compliant version of Ophidia for ESGF.
- CERFACS (Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique) improved on Climate4Impact with seamless delegation of calculation to CWT, while adding multiple auth schemas to Climate4Impact (i.e., B2ACCESS, ESGF OpenID, EGI FedCloud Certificates, CWT API Key). Deployment of the calculations was done using the European Data Infrastructure B2 Generic Execution Framework Service. Also, significant work has been done in the icclim backend that calculates climate indices/indicators (large performance improvements, bugs fixed, Python 3 full compatibility). Icclim is also being used by CRIM (Centre de Recherche Informatique de Montréal)/Ouranos.

- The University of Utah ported the Visualization Streams for Ultimate Scalability (ViSUS) software to a container-based architecture, allowing for easy adoption and deployment. ViSUS's IDX library now has a companion library (xidx) that allows ViSUS to serve the companion metadata.
- CRIM implemented WPS-T 2.0 REST JSON (JavaScript object notation) server, created a new AuthZ/AuthN module for birdhouse, made progress toward interoperability with ESGF-CWT API, and started to look into use cases for ML.
- Ouranos launched an operational WPS for climate analysis (Power Analytics and Visualization for Climate Science, or PAVICS), started to embed hydrological models and related analytics as WPS, developed "xclim" a climate indices library based on icclim+xarray+dask+WPS, added native OpenDAP support to PyWPS, and wrote a Python interface to WPS based on OWSLib.
- Finally, the Centre for Environmental Data Analysis (CEDA), Institut Pierre-Simon Laplace (IPSL), and DRKZ established components supporting future Open Geospatial Consortium WPS deployments at ENES sites.
- CWT came together to address concerns about many centers starting to offer "ESGF" service and hurt the ESGF brand name. As a result, a Certification document was approved by the XC. This certification process for both server and operator will be the base of next year's main CWT task—the compute grand challenge(s). This should address the ESGF community's concern for seeing operational proof of CWT's activities.

Future Roadmap

During the F2F conference, the CWT came together to plan a series of compute grand challenge(s) with the goal of having three or four production-ready, officially certified ESGF compute nodes. The complete roadmap includes the following tasks:

- Identification of CMIP6 and CMIP6 datasets for testing
- Identification of existing security scans for other ESGF components
- Communication with affected institutions regarding the exact path for the production node to open
- API development, baselining, and compliance
- Scaling the service to multiple servers
- Documentation
- Monthly progress reports for the XC
- Beta release in April 2019
- Public release in July 2019

3. Identity, Entitlement, and Access Working Team

Progress Report

The focus for 2018 was integration of OAuth 2.0 into the ESGF code base to migrate away from the legacy OpenID 2.0 single sign-on technology. Use of OAuth 2.0 increases the usability of the system while also supporting new use cases involving user delegation. The following activities took place in 2018:

- OAuth code development has been completed for all major components, including CoG. The OAuth service has been deployed at three sites in the federation: CEDA, LLNL's

Program for Climate Model Diagnosis and Intercomparison, and JPL. It is being used by other services in the federation for single sign-on and user delegation (e.g., with the ENES Climate Impacts Portal and the CWT node software stack).

- OAuth services are bundled with the ESGF installation. Node deployers can select whether to enable them.
- The changes allow simultaneous support for OAuth and legacy OpenID 2.0. This will enable the federation to conduct a staged migration from OpenID 2.0 to OAuth as and when centers are ready.
- MyProxyCA is replaced with the short-lived credential service (SLCS). The SLCS is a web service implementation of MyProxy that allows integration of OAuth so that clients can get delegated user certificates.
- Replacement of the current Wget scripts is under development. This takes advantage of OAuth, embedding a certificate directly into the script when it is issued and avoiding the need for the user to authenticate when the script is invoked.

Future Roadmap

Full deployment of OAuth services at all sites in the federation is the priority for the coming year. It is important that OpenID 2.0 is retired and the new capabilities with OAuth 2.0 can be exploited. The following work is planned for 2019:

- Semi-automation for OAuth client registration. Clients (i.e., services using OAuth 2.0) must register with the OAuth authorization services that they wish to use. For any service to take advantage of single sign-on across all the federation, it must register with the authorization services of all the identity providers (IdPs). This process should be automated to some degree to reduce the management burden for service deployers and the maintainers of IdPs.
- Provide support for OAuth access tokens. Support to access secured resources such as processes on compute nodes and data from data nodes would greatly simplify APIs for access-secured resources and reduce the burden of complexity with the existing system, which uses X.509 user certificates. The latter capability could still be retained for specific use cases as required.
- Migrate to OpenID Connect (OIDC). A WT on OIDC for research federations has been established. There is a need to track progress with this and also present ESGF's use cases. OIDC builds on top of the OAuth 2.0 framework and so the changes required for any migration would be reduced to some degree.
- Retire SAML (security assertion markup language) attribute service interface and replace with OAuth 2.0. OAuth Client Credential flow can implement this and doing so would reduce the overall burden of supported code but is not essential for future development of the system.
- Explore additional technologies. We plan to explore use of alternate off-the-shelf implementations for parts of the IdEA system such as the JASMIN Accounts Portal (accounts.jasmin.ac.uk) and Keycloak (keycloak.org).

4. Installation and Software Security Working Team

Progress Report

The Installation Working Team continued to improve the performance and reliability of the ESGF installation process. The following activities took place in 2018:

- The ESGF 2.6.x releases were completed during the DC period and addressed several shortcomings in the software as we tested the components needed for CMIP6 publication, access, and metrics.
- v2.7.1 upgraded to Solr 6.6.5 to address a vulnerability.
- Considerable effort went into v2.8.x testing, with v2.8.1 released in January 2019. The goal of these releases was to identify and remove components susceptible to known vulnerabilities.
- Significant progress was made on the Python-based ESGF installer script. Some issues the Python-based installer addressed are poor code architecture, codebase fragmentation, lack of error handling, paltry automated testing, and a lack of documentation. An alpha version of the v3.0 installer was released in March 2018, and a beta version was released in December 2018 at the conference. The refactor effort for the v3.0 installer was significant with more than 3,300 commits, 220 closed GitHub issues, and 120 merged pull requests. The development process gave the team new insights on how components of the legacy code interacted together, and we were able to use that knowledge to streamline some of the installation steps.
- ESGF 3.0 has moved away from hosting installation scripts on distribution mirrors. All installation scripts for v3.0 are kept in the GitHub repository. Only larger files such as compiled jar files are fetched from the distribution mirror during a v3.0 installation.
- As part of the activities of the Software Security Working Team, a tool called ESGF Scanner was developed, to allow for regular checking of ESGF software dependencies against known lists of vulnerabilities. The reports generated from ESGF Scanner were used during testing of v2.8.x, to try and address as many known vulnerabilities as was feasible. v2.7.1 had more than 75 open vulnerabilities and exposures against it, while v2.8.2 has 13, of which at least 8 are known to be practically unexploitable. Work was done to integrate ESGF Scanner runs as part of a Jenkins workflow, allowing for a vulnerability scan to be performed as soon as a new release is available.
- To improve our development, we adopted a more formal code review process for contributing to ESGF. No code should be pushed to the master or devel branch directly. Instead, a pull request (and associated GitHub issue) should be made that will ideally be reviewed and approved by two peers. Whenever a new pull request is merged into the devel branch, a new tag will be cut. the bump in the tag will be determined according to the semantic versioning methodology.
- After presenting on ESGF 3.0 at the conference, we had conversations with collaborators (specifically with Matt Pryor of CEDA) about the next steps to release. We decided to explore Ansible as a proof of concept, and it appears to be a much cleaner solution. Moving forward, we are ceasing development on ESGF 3.0 and shifting our efforts to implementing the installer in Ansible. Ansible is a tool designed for managing deployments to multiple servers and will allow us to refactor the ESGF installer configuration scripts written in YAML. The development time for implementing features

in Ansible has been dramatically shorter than in Python, in part due to the knowledge we obtained by first implementing the features in Python.

Future Roadmap

The following work is planned for 2019:

- The primary focus in January 2019 is ensuring we are able to bring up all services. This ideally includes both CentOS 6 and 7, and there is no immediate intention to begin testing on Redhat. This effort will be guided more by the documentation for each service rather than by referencing previous installers. However, there are often undocumented and somewhat hidden catches, and in these cases, the previous installers are referenced. At this point, only Globus services (specifically GridFTP) are being troublesome on CentOS 7.
- Another important aspect of using Ansible is to not only "make it work right now" but to make the deployment work in a reliable way, resilient to external factors. It also should be easily understood and modified. While the initial Ansible configuration layout was designed to support these ideas, and they have been followed thus far, it is important to continue doing so. Opportunities exist for improvements and simplifications, and the installation challenge will be a good way to refine our configuration.
- Ongoing work will strive toward improving the Ansible interface and gaining a comprehensive understanding. Eventually, installation should be a release test process.

5. Containers Working Team

Progress Report

The following activities took place in 2018:

- Containerization of the ESGF software stack. Over the course of 12 months, the WT provided an alternative implementation of the ESGF software stack, where each service is installed and deployed as a Docker container. A full ESGF node can be deployed on a single server using docker-compose or on a cluster of nodes using a Helm chart (i.e., a convenient package of Kubernetes API objects). This architecture was tested using an onsite cluster at JPL as well as Kubernetes clusters on Amazon Web Services and Google Cloud. The first release v1.0 of ESGF/Docker (stable but not yet feature complete) was released in September 2018.
- Solr Cloud index node. A new architecture for the ESGF search services was prototyped and presented at the conference. This work is described in the progress report of the UI and Search WT.
- Continuous integration (CI) setup. The production instance of the ESGF/Docker images (components) implements CI best practices. ESGF-jenkins, a scalable cluster of nodes of Jenkins, runs the CI job that automatically builds, tests (ESGF-test-suite), packages, and makes the images available on Dockerhub when the ESGF Container Working Team makes a successful pull request into a branch of the GitHub repository of ESGF/Docker. The CI process includes preventing software regression, improving reactive security, and making ESGF/Docker production less error prone.
- ESGF/Pangeo testbed. The ESGF collaborated with members of the Pangeo project to demonstrate a proof of concept for scalable analysis of ESGF data holdings via the

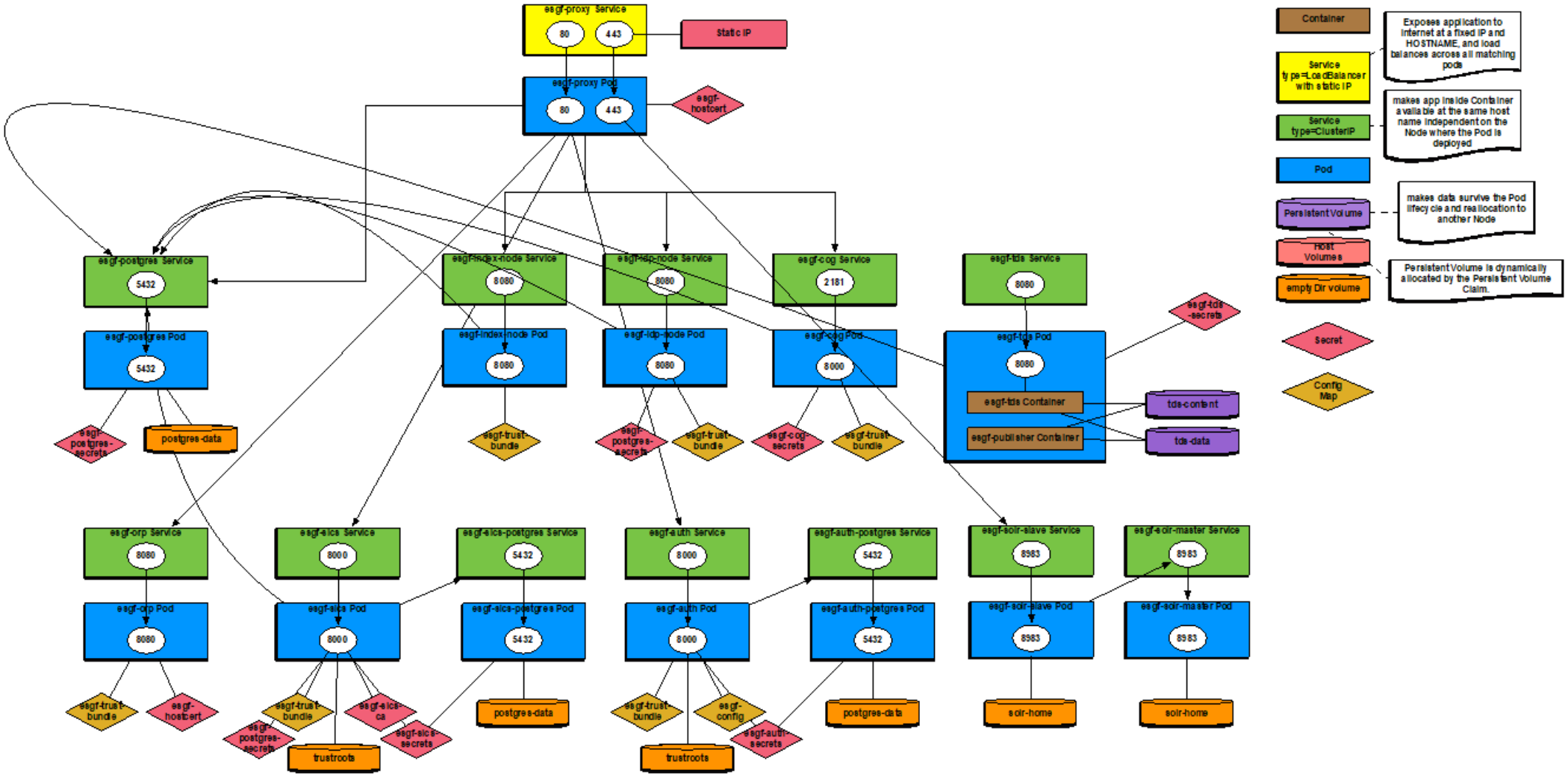
Pangeo infrastructure. A test ESGF node was deployed on the Google Cloud by Geophysical Fluid Dynamics Laboratory (GFDL) and JPL staff, and populated with sample CMIP6 data. A Pangeo notebook was developed that can access these data via OpenDAP (Open-Source Project for a Network Data Access Protocol) and execute climate science algorithms on a cluster of distributed computing nodes.

Future Roadmap

The following work is planned for 2019:

- Bring the ESGF/Docker installation up to par with the classic installer (either the current shell-based installer or the Python based installer), so that it can be a viable—and preferable—alternative to the standard installation method for an ESGF node. This will require updating all Docker images to run the latest release of the service as well as implementing an image for the Globus data transfer service.
- Enable deployment of different topologies using the shared pool of ESGF/Docker images—that is, the capability to deploy a full ESGF node, just an Index node, a Data node, and so forth.
- GSFC and JPL plan to use the ESGF/Docker container distribution to set up and operate a combined ESGF/NASA node on the Amazon Cloud as a replacement for the two nodes currently deployed separately at the respective institutions. This will require the Docker stack to be brought to a production-level standard.
- ESGF-jenkins plans to support the dynamic updating of ESGF/Docker deployments.

Figure 2. *Diagram showing containerization of the ESGF software stack.*



6. International Climate Network Working Team and Replication/Versioning and Data Transfer Working Team

Progress Report

After replication testing as part of the DCs in 2018, CMIP6 data collections were replicated at LLNL and at Deutsches Klimarechenzentrum (German Climate Computing Centre, or DKRZ). As part of these testing activities as well as this first transition to production replication between sites, a number of technical and organizational issues were identified and addressed:

- The replication software stack relies on the Synda replication tool, which is currently based on the (old) 2017 code base. A large number of issues and pull requests for this code base has emerged since then. Based on a priority ordering of bug fixes and needed feature updates, new Synda releases are scheduled for 2019.
- Replication testing based on globus-url-copy as well as Globus online showed inefficient use of Globus features in the current Synda code base. Better exploitation of Globus online features requires collaboration with the Globus team. A work plan for this has been collected and Globus support was ensured such that these issues can be addressed in 2019.
- Based on test infrastructure experiences, the core replication sites (Tier 1 sites: DKRZ, LLNL, IPSL, CEDA, NCI) deployed production data transfer nodes (DTNs) that will be used in CMIP6 replication activities. While tests showed the possibility to support approximately 300-megabyte transfer rates between sites, a number of configuration and optimization issues must be addressed to sustain these rates for very large data collections.
- Operational issues were identified in relation with the consistency of data replicas across sites in the case of un-publication of datasets as well as the publication of new versions. The “latest-version” problem was addressed based on an operational agreement between sites to update their search index regularly with respect to “latest-flag” information. Approaches to address the un-publication problem are currently developed by LLNL and IPSL.
- To support future replication activities, the exploitation of PID-based tools was discussed, and first steps were identified for implementation in 2019.
- As a basis for replication planning between sites—to ensure overall replication requirements, such as at least one copy available across the federation or at least one copy of the most important data-collections available at a center in each continent—information was collected with respect to the most often used variables based on CMIP5 experiences. This information is maintained and managed on GitHub.

Future Roadmap

In 2019, work will concentrate on the replication of CMIP6 data across Tier 1 replica centers to support the reliable and timely distribution of CMIP6 data to end users.

- The Synda code base will be improved and redesigned, optimizing Globus-based data transfers. To support this Synda restructuring, the WT will prioritize issues to be resolved in future Synda releases.

- The ESGF DTN performance monitoring and optimization will continue at Tier 1 sites as the expected 2019 CMIP6 volume to be published will constitute an unprecedented data replication exercise for all Tier 1 sites.
- As a basis for future data replication management tools, PID-based replication management tools will be prototyped. This will allow a clear identification and synchronization of replication sets based on unique identifiers for (eventually large) data collections.
- Additionally, replica centers will make their replicated CMIP5 data collections also accessible via their DTNs.

7. Node Manager and Tracking/Feedback Notification Working Team

Progress Report

With regard to the errata service, the following activities took place in 2018:

- CMIP6 DCs. The ESGF errata service beta release (v0.6.2.0) was part of the CMIP6 DCs. This leads to the improvement of the errata command-line client and helped us gather useful feedback for the front-end component of the system. Best practices about the errata and issue registration have been communicated to all concerned actors during the DCs.
- Web forms. The new web forms were implemented and deployed to the front-end during the CMIP6 DC to facilitate the issue management (creation, update, and closure). Users can now choose to manage their issues through the web forms or the command-line client.
- Production release. Since June 2018, the ESGF errata service has been in production (errata.es-doc.org) and supports issue registration for CMIP6, CMIP5, and CORDEX projects. Two minor issues appeared since opening the ESGF errata service to CMIP6:
    - a stopped PID ingestion due to a password change on the IPSL RabbitMQ instance, and
    - minor bugs related to the front-end display features.

  Those issues have been resolved, and the service is fully operational. The errata service currently counts 20 issues related to CMIP6 (13 are resolved, 1 will not be fixed, and 6 are new or on hold).
- Documentation upgrade. The documentation is now easily accessible through the front-end of the service. It has been entirely updated with step-by-step tutorials including screenshots. Information is still available to guide the user through the command-line client usage. The API endpoints have been detailed.

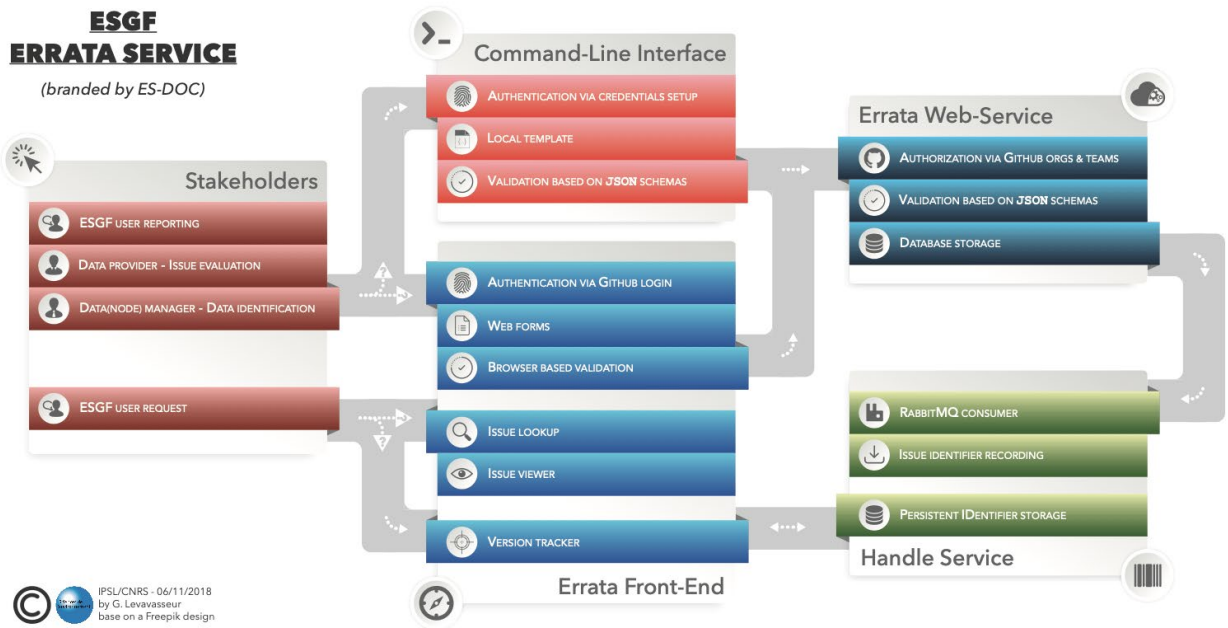


Figure 3. *ESGF errata service overview.*

This WT's additional 2018 accomplishments include:

- Node manager. The ESGF node manager daemon has been phased out of the software stack as of the v2.6 release. While we have considered keeping the esgf-nm API component in the stack for use in a registry service, until we have a use case for such a service, we opt not to maintain the component. Its installation will be phased out as we switch from v2.x to the next-generation ESGF deployments (to be based on Ansible, version number to be determined).
- User notification. For user notification services, we have implemented a prototype of the subscription-based notification. This system completes email notifications based on new publications, updates to existing datasets and retracted publications. These are matched based on experiment or variable criteria in a test project based on CMIP6. To test the prototype effectively in a real-time environment, we have implemented a publication simulator that automatically publishes at regular intervals.

Future Roadmap

The following work is planned for 2019:

- Error handling. Make the error messages more readable for the users when something is wrong in the web forms or the command-line client returns. The user has to understand the error before sending any request to the errata support.
- Check PID through the command-line client. To allow automated processes in third-party tools (e.g., publisher, CliMAF, CoG), the PID lookup feature will be copied from the

front-end to the command-line client. The results should be lightweight (e.g., using Boolean) with useful prompts if needed (e.g., issue exists, latest version).

- Informational issue. We decided to expand the errata service to non-error "issues." In some cases, ESGF data can be published or not published for other reasons than an error (e.g., simulation extension, correction prior to a first version). Itis useful for the data provider to document such an approach in order to facilitate the user support.
- Statistics endpoint. Statistics are often requested in the ESGF or project reports/survey in order to know is a service useful to the users. We decided to implement a new API endpoint and a corresponding webpage to serve basic statistics about registered issues (e.g., number of issues, number of resolved, issues grouped by project, severity, status).
- Communication. IPSL staff committed to providing support and best practices to all CMIP6 climate modeling groups. The door is also open to any ESGF project that would benefit from the errata service.
- Notification system. We decided to start designing a notification system to allow users to subscribe to an "issue flux." The users would be notified about major issue changes (e.g., severity, status, affected datasets) instead of regularly requesting the service. This will be part of a more general discussion in the ESGF as a notification service becomes necessary for other components.
- Monitoring service. We will be exploring the deployment of a monitoring service module in the software stack, leaning toward using the Prometheus framework, as previously adopted by the LLNL compute software stack reference implementation. This will address the gap of the node manager's role as a self-monitoring module.
- Beyond the prototype. We will work toward moving the subscription-based service from a prototype to a fully working system. To accomplish this, first we will add the functionality for users to subscribe to updates, most likely within CoG. Second, we will improve the backend scalability by searching for updates using the search API enhanced with range searches over timestamps enabled. A smaller effort will involve implementing several standing queries to show recently published datasets.
- Download tracking. Moving on from the subscription and standing query-based approaches, we will prototype a means to track downloads without the need to force user logons. Given this ability, we can provide notifications based on new versions or retractions of datasets previously downloaded. Given the success of this prototype, we will work toward an eventual release of this system as well. Enhancing notifications beyond email is a goal but likely beyond the scope of 2019 work.

8. Publication, Quality Control, and Metadata Working Team

Progress Report

The following activities took place in 2018:

- The primary focus of this WT was ensuring the reliability of the ESGF software infrastructure for CMIP6 publication, both at the sites of our participants and of external collaborators who have less familiarity with the software. A considerable obstacle that was addressed was the integration of PrePARE and how importing CMOR tables with conflicting versions are managed. Given our experience, we have produced

comprehensive documentation, including step-by-step guides in Jupyter notebooks to aid the broader community.

- In addition to CMIP6, 2018 saw the republication and ongoing work within the Input4MIPs project. E3SM published in its native output format and, given the sheer size of the data for some datasets, stressed the software to its limits, giving us an opportunity to learn what may help with future publication jobs of comparable size.
- Several other software enhancements have been made to the esg-publisher. We have completely switched over to the esg-search REST API, leaving the hessian API as a deprecated legacy option. To facilitate code development, we have ported all http requests to use the well-known, easy to use "requests" module. We have added several options to enable the publication of large datasets much more efficiently—namely the disabling of aggregations and a more aggressive commit schedule to PostgreSQL.

Future Roadmap

The following work is planned for 2019:

- Given the end date for Python 2.7 support, we will convert the publisher and its dependencies to Python 3 in 2019.
- We will consider an integration of the pyeesv controlled vocabulary service, which would alleviate the need to configure vocabularies within project-specific ini files.
- We will explore mechanisms for enhancing publication:
    - remote operation of the esg-publisher software;
    - running publication in parallel; and
    - Conda environment for publication that can be imported independent of ESGF software stack releases.

9. User Support and Documentation Working Team

No report is available because the WT lead changed responsibilities in 2018. The day-to-day work has been covered by other WTs. In 2019, ESGF will appoint a new team lead and reactivate this WT, which is important for the stability of ESGF.

10. Machine Learning Working Team

Progress Report

The following activities took place in 2018 regarding deep learning for climate science:

- Detection of extreme climate events using convolutional NNs (CNNs). Conventional extreme climate event detection relies on high spatial resolution climate model output for improved accuracy. As a cost-efficient alternative, we developed a system to detect and locate extreme climate events using the five-layered CNN, which is trained for binary classification and location regression tasks for hurricanes. Our cross-validation results show 99.98% detection accuracy, and the localization accuracy is within 4.5 degrees of longitude/latitude (around 500 km and three times the data resolution).
- Resolution reconstruction of climate data with pixel recursive model. Our results using CNNs for extreme climate events detection show that simple NNs can capture the pattern of extreme climate events with high accuracy from very coarse reanalysis data. However, localization accuracy is relatively low due to the low resolution of input climate images.

To resolve this issue, we developed the pixel-recursive super-resolution model reconstructs the resolution of climate images, so we can potentially increase the accuracy of localization task using NNs. Using this model, we developed the novel networks that can synthesize details of tropical cyclones in ground truth data while enhancing their resolution. Therefore, this approach suggests the possibility of reducing the computing cost required for downscaling process to increase resolution of data. With best of our knowledge, this is the first model using NN-based super-resolution techniques to enhance the quality of climate data.

- Tracking tropical cyclones using long short-term memory (LSTM). In the spatiotemporal CAM5 climate simulation data containing the single trajectory of a tropical cyclone, we developed the tracking framework with CNN and LSTM to track the trajectory of a tropical cyclone. The CNN first embeds an image of each time frame and the embedding of the image feed to the LSTM cell as the input. The hidden state of LSTM cells following the fully connected network predicts the latitude and the longitude of the tropical cyclone by the regression operation. We performed qualitative analysis that shows promising potentials but also several limitations of the primitive LSTM.
- Tracking and forecasting tropical cyclones using ConvLSTM. We developed Convolutional LSTM (ConvLSTM)–based spatiotemporal models to track and predict hurricane trajectories from large-scale climate data—namely, pixel-level spatiotemporal history of tropical cyclones. To address the tracking problem, we model time-sequential density maps of hurricane trajectories, enabling capture of not only the temporal dynamics but also spatial distribution of the trajectories. Furthermore, we introduced a new trajectory prediction approach as a problem of sequential forecasting from past to future hurricane density map sequences. Extensive experiments on actual 20 years' record shows that our ConvLSTM-based tracking model significantly outperforms existing approaches, and that the proposed forecasting model achieves successful mapping from predicted density map to ground truth.

Future Roadmap

The following work is planned for 2019:

- We will be looking at spatiotemporal prediction of climate data including short-term weather prediction.
- We will investigate applying deep reinforcement learning for numerical weather prediction.

B. Current Data Holdings

- Coupled Model Intercomparison Project Phase 6 (CMIP6) – contributions will continue through 2020
- 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)

C. Conference Participants and Report Contributors

Figure 4. *Conference attendees.*

Joint International Agency Conference and Report Organizers

- Michael Lautenschlager – Co-Chair of the ESGF-XC, ENES/DKRZ
- Ghaleb Abdulla – ESGF-XC, LLNL
- Sasha Ames – ESGF-XC (acting), LLNL
- Venkatramani. Balaji – ESGF-XC, Princeton University
- Laura Carriere – ESGF-XC, GSFC
- Luca Cinquini – ESGF-XC, JPL
- Sébastien Denvil – ESGF-XC, IPSL
- Ben Evans – ESGF-XC, NCI
- Robert Ferraro – ESGF-XC, JPL
- Tom Landry – ESGF-XC, CRIM
- Philip Kershaw – ESGF-XC, CEDA, STFC

ESGF Program Managers in Attendance

- Justin Hnilo – Chair of the ESGF-SC, DOE Office of BER
- Sylvie Joussaume – ESGF-SC, ENES
- Claire Richards – ESGF-SC, NCI
- Tsengdar Lee – ESGF-SC, NASA (by telecon)

Attendees and Contributors

No.	Name	Affiliation
1	Abdulla, Ghaleb	LLNL
2	Aloisio, Giovanni	Foundation Euro-Mediterranean Center on Climate Change
3	Ames, Sasha	LLNL
4	Auten, Holly	LLNL
5	Bader, David	LLNL
6	Bai, Yuqi	Tsinghua University
7	Balaji, Venkatramani	Princeton University
8	Ben Nasser, Atef	IPSL
9	Boutte, Jason	LLNL
10	Carlson, Nathan	LLNL
11	Carriere, Laura	GSFC
12	Choi, Jaewon	APEC Climate Center

No.	Name	Affiliation
13	Cimadevilla Alvarez, Ezequiel	University of Cantabria, Spain
14	Cinquini, Luca	JPL
15	Cofiño, Antonio S.	University of Cantabria
16	Denvil, Sébastien	IPSL
17	Doutriaux, Charles	LLNL
18	Downie, Carlos	LLNL
19	Durack, Paul	LLNL
20	Dwarakanath, Prashanth	LiU
21	Easterbrook, Steve	University of Toronto
22	Ferraro, Robert	JPL
23	Fiore, Sandro	Foundation Euro-Mediterranean Center on Climate Change
24	Gardoll, Sébastien	IPSL
25	Gleckler, Peter	LLNL
26	Greenslade, Mark	IPSL
27	Gruver, Patricia	Québec Government Office in Boston
28	Han, Jeongmin	APEC Climate Center
29	Hill, William	LLNL
30	Huard, David	Ouranos
31	Inoue, Takahiro	Research Organization for Information Science & Technology
32	Iwi, Alan	United Kingdom Research and Innovation
33	Jefferson, Angela	LLNL
34	Joussaume, Sylvie	Centre National de la Recherche Scientifique/IPSL
35	Juckes, Martin	CEDA

No.	Name	Affiliation
36	Kershaw, Philip	CEDA, STFC
37	Kim, Sookyung	LLNL
38	Kindermann, Stephan	DKRZ
39	Landry, Tom	CRIM
40	Lautenschlager, Michael	DKRZ
41	Lee, Huikyo	JPL
42	Levavasseur, Guillaume	IPSL
43	Maxwell, Thomas	GSFC
44	McCoy, Renata	LLNL
45	Muryanto, Lina	LLNL
46	Nadeau, Denis	LLNL
47	Nikonov, Serguei	GFDL/Princeton University
48	Nunez, Cheryll	LLNL
49	Nuzzo, Alessandra	Foundation Euro-Mediterranean Center on Climate Change
50	Pagé, Christian	CERFACS
51	Percivall, George	Open Geospatial Consortium
52	Peterschmitt, Jean-Yves	Laboratoire des Sciences du Climat et de l'Environnement
53	Petruzza, Steve	University of Utah
54	Pritchard, Matt	CEDA
55	Pryor, Matthew	CEDA
56	Richards, Clare	National Computational Infrastructure
57	Shaheen, Zeshawn	LLNL
58	Shen, Yingshuo	GSFC
59	Sim, Alex	Lawrence Berkeley National Laboratory

No.	Name	Affiliation
60	Stockhause, Martina	DKRZ
61	Story, Matthew	LLNL
62	Taylor, Karl	LLNL
63	Vahlenkamp, Hans	GFDL
64	Weigel, Tobias	DKRZ
65	Yan, Jinghui	Beijing Climate Center

D. Awards

Individuals

Name: Guillaume Levavasseur (IPSL) and Alan Iwi (CEDA)
Award title: Exceptional contribution to CDNOT DCs

The ESGF-XC would like to bestow an award for exceptional contributions during the CMIP6 DCs to Guillaume Levavasseur and Alan Iwi. During this year, Guillaume and Alan have demonstrated week after week their ability to grasp and solve complex technical questions thanks to their thoroughness and technical skills. Without respite, they polished and dominated the CMIP6 publication chain in all its dimensions. Additionally, Guillaume and Alan have been long-time contributors to ESGF. Thanks to their impressive problem-solving mindset, they have just been tackling any problem that came along, and we sincerely thank them for this.

Name: Prashanth Dwarakanath (LiU)
Award title: Exceptional contribution to the ESGF release cycle

The ESGF-XC would like to bestow an award for exceptional contribution during the last year to Prashanth Dwarakanath. During this year, Prashanth provided us with more than a dozen ESGF releases. The CMIP6 DCs required from Prashanth a very high level of commitment, week after week during several months. Also very important were his strong communication skills when it came to timeline schedule that were necessary to make the best of the group's workforce.

Name: Ruth Petrie (NCAS/CEDA, STFC Rutherford Appleton Laboratory)
Award title: Exceptional service for CDNOT

We wish to recognize Ruth Petrie for her indispensable contributions to the operation of the CDNOT, particularly during the ESGF DCs that started in early 2018 in preparation for handling the CMIP6 model output. Without her contribution, it would be very hard to see that the DCs could have been run successfully, and that the community would be able to make our goal of June 2018 readiness for a CMIP6 opening. Specifically, she was outstanding as an incredible, diligent scribe capturing the DC discussions every week, providing all the CDNOT documents to organize the DCs, and consistent and timely with

needed communication. We would like to thank Ruth for all the hard work she put in for the CDNOT and welcome her continued contribution to the effort in 2019.

Name: Martina Stockhause (DKRZ)
Award title: Implementation of the CMIP6 early citation service

The ESGF-XC would like to recognize Martina's outstanding contributions to design, implementation, and starting operations of the CMIP6 early citation service. Since the beginning of the CMIP6 data management discussions, she spent a lot of effort in designing the citation service and coordinating its implementation in collaboration with the WIP and the ESGF Working Teams, as well as with the other CMIP6 services in the ESGF. The early citation service has been prototyped for the first CMIP6 data, allowing for direct visibility of these data in the recently launched Google Data Portal. We would like to thank Martina for her great work and hope that she will continue her engagement in ESGF at least throughout CMIP6.

Name: Stephan Kindermann (DKRZ)
Award title: Leading the CMIP6 data replication service

The ESGF-XC would like to recognize Stephan's outstanding and continuous support of the ESGF data replication activities over the last few years. The work started with the survey and optimization of the wide-area network connection between ESGF Tier 1 sites. The implementation effort continued for CMIP6 as part of the DCs, and it is now an integral part of the CDNOT tasks for coordination of the data operations. Potential for performance optimization in the responsibility of the ESGF has been identified in local networks at data node sites and in the parallelization of the data replication software stack. We would like to thank Stephan for his persistence and dedication to CMIP6 activities, even if they are more behind-the-scenes, and we hope that he will continue his work for ESGF and CMIP6 in the future.

Name: Matt Pryor (NCAS/CEDA, STFC Rutherford Appleton Laboratory)
Award title: Technical contribution to the development of container-based ESGF deployment

We would like to recognize Matt's excellent technical contribution to the Container Working Team. Over the past year the WT has been developing a fully containerized version of the ESGF software. This has required the integration of a complex set of application components into a completely new deployment and operational framework. Matt has worked closely with the team and incorporated key innovations and applied best practice in this new and fasting moving field—particularly in the application of Kubernetes and the use of Helm charts. He has made major contributions to the development of a modular architecture for the deployment and operation of a container-based ESGF services stack, and we are seeing the fruit of this in the first operational installations.

Name: Michael Lautenschlager (DKRZ)
Award title: ESGF-XC exceptional leadership

The ESGF-XC would like to bestow an award for exceptional leadership to Michael Lautenschlager, head of the Data Management department at DKRZ. Following Dean William's prolonged absence, Michael has assumed primary responsibility for managing

and directing the ESGF-XC including leading the monthly conference calls, charting the program of the 2018 F2F conference, and leading European contributions to ESGF. Additionally, Michael has been a long-time supporter of ESGF from its inception, and the group he leads at DKRZ is one of the most important contributors to the software and data management infrastructure for CMIP6. Michael, please accept our most sincere thanks.

Name: Sébastien Denvil (IPSL)
Award title: Extraordinary achievement in CDNOT leadership

The ESGF community would like to recognize Sébastien Denvil of IPSL for his singular efforts in bringing forth a functioning federation of data nodes hosting data for CMIP6. In his leadership role within the CDNOT, he orchestrated the very complex process of operationalizing the federation for CMIP6. This was done during 2017–2018 when everything was in flux—the scientific design of CMIP6, the design and implementation of the ESGF software stack, and the number of participating institutions (currently standing at four data nodes hosting output from six models from four institutions). The challenges included ensuring participation from a quorum of data nodes, getting consensus on a software stack from development and deployment teams, and most importantly, orchestrating a series of DCs to ensure that the federation indeed functioned as designed. Without Sebastien's indefatigable efforts, we would not be where we are today. The ESGF community therefore salutes Sebastien with an Extraordinary Achievement Award at the 2018 ESGF F2F.

Groups

Names: Karl Taylor (LLNL) and Venkatramani Balaji (GFDL/Princeton)
Award title: CMIP6 recognition

On behalf of the climate change community, the ESGF-XC would like to recognize Venkatramani Balaji and Karl Taylor for their continued leadership in the WIP. Balaji and Karl have worked very hard to bridge the gap between the modeling groups that generate CMIP6 model data, and the ESGF developer community who is providing the software infrastructure to organize and make these data accessible worldwide. Among other activities, they have been providing constant feedback on the usability of the ESGF UI and APIs, as well as leading the community in defining the data and metadata standards for the CMIP6 model output. Both Karl and Balaji have been steady supporters of the ESGF collaboration from the very beginning. A big thanks to both of them.

Names: Paul Durack (LLNL), Peter Gleckler (LLNL), Robert Ferraro (JPL), and Duane Waliser (JPL)
Award title: Obs4MIPs recognition

The ESGF-XC would like to present a group award to the people responsible for the success of the Obs4MIPs project, in alphabetical order: Paul Durack, Robert Ferraro, Peter Gleckler, and Duane Waliser. The main goal of Obs4MIPs is to enhance the exposure and usability of critical observational datasets from several agencies (e.g., NASA, NOAA, ESA) for climate change research, making them accessible through the same interfaces and APIs as the model data. The Obs4MIPs leaders have worked for several years to make this possible by defining data and metadata standards, coordinating

work across international groups, securing funding for data processing, and transforming current datasets to comply with the same standards as CMIP6 data. This project is quickly gaining in popularity among observational data providers and will guarantee a more systematic and wider use of observational data in the next Intergovernmental Panel on Climate Change assessment report on climate change.

E. Acronyms

Acronym	Definition
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 the U.S. DOE (anl.gov).
API	Application programming interface
BER	U.S. 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/).
CANARIE	Canadian Network for the Advancement of Research, Industry and Education
CCI	ESA's Climate Change Initiative
CDNOT	Coupled Model Intercomparison Project Data Node Operations Team
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).
CERFACS	Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique
CI	Continuous integration

Acronym	Definition
Climate4Impact	Web portal that enables visualization of climate model datasets 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 WCRP's Working Team 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 UIs—Collaborative software enabling projects to create dedicated workspaces, network with other projects, and share and consolidate information within those networks (earthsystemcog.org/projects/cog/).
ConvLSTM	Convolutional long short-term memory—A neural network approach that enables spatiotemporal models to track and predict hurricane trajectories from large-scale climate data.
CORDEX	Coordinated Regional Climate Downscaling Experiment—Provides global coordination of regional climate downscaling for improved regional climate change adaptation and impact assessment (cordex.org).
CREATE	Collaborative REAnalysis Technical Environment—NASA project that centralizes numerous global reanalysis datasets into a single advanced data analytics platform.
CREATE-IP	Collaborative REAnalysis Technical Environment Intercomparison Project—Data collection, standardization, and ESGF distribution component of CREATE (earthsystemcog.org/projects/create-ip/).
CRIM	Centre de Recherche Informatique de Montréal—Computer Research Institute of Montréal (crim.ca)

Acronym	Definition
CWT	Compute Working Team
DC	Data challenge
DKRZ	Deutsches Klimarechenzentrum (German Climate Computing Centre)—Provides high-performance computing platforms and sophisticated, high-capacity data management and services for climate science (dkrz.de).
DOE	U.S. Department of Energy—Government agency chiefly responsible for implementing energy policy (energy.gov).
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 (climatemodeling.science.energy.gov/projects/energy-exascale-earth-system-model).
ENES	European Network for Earth System Modelling—Common infrastructure for distributed climate research and modeling in Europe, integrating community Earth system models and their hardware, software, and data environments (verc.enes.org).
ES-DOC	Earth system documentation
ESA	European Space Agency—International organization coordinating the development of Europe's space capability, with programs to develop satellite-based technologies and services and to learn more about Earth, its immediate space environment, the solar system, and universe (esa.int/ESA/).
ESGF	Earth System Grid Federation—Led by LLNL, a worldwide federation of climate and computer scientists deploying a distributed multi-petabyte archive for climate science (esgf.llnl.gov).
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).
F2F	Face-to-face

Acronym	Definition
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).
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/).
ICNWG	International Climate Network Working Team (Group)—Formed under the ESGF to help set up and optimize network infrastructure for climate data sites around the world (icnwg.llnl.gov).
IdEA	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/).
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).
Jenkins	An open-source, Java-based automation server that provides CI capabilities for software development.
JPL	Jet Propulsion Laboratory—A federally funded research and development laboratory and NASA field center in Pasadena, California (jpl.nasa.gov).

Acronym	Definition
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).
LSTM	Long short-term memory—A recurrent neural network with feedback connections among nodes.
MIP	Model Intercomparison Project
ML	Machine learning
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).
NCI	National Computational Infrastructure—Australia's high-performance supercomputing, Cloud, and data repository (nci.org.au).
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).
NSF	National Science Foundation
OAuth	Open standard for authorization (oauth.net)

Acronym	Definition
Obs4MIPs	Observations for Model Intercomparisons—Database used by the CMIP modeling community for comparing satellite observations with climate model predictions (earthsystemcog.org/projects/Obs4MIPs/).
OIDC	OpenID Connect—Authentication layer of the OAuth 2.0 framework.
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).
OpenID	An open standard and decentralized authentication protocol. (CoG uses an ESGF OpenID as its authentication mechanism.)
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)
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).
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
SC	Steering committee
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,

Acronym	Definition
	and international benefits to the United Kingdom and, more broadly, the world (stfc.ac.uk). Also includes the Rutherford Appleton Laboratory.
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
ViSUS	Visualization Streams for Ultimate Scalability
WCRP	World Climate Research Programme—Facilitates 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 Team (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/).
WPS	Web processing services
WT	Working Teams (also known as Working Groups) responsible for developing, and implementing, and managing different aspects of the ESGF ecosystem.
XC	Executive Committee