2016 Earth System Grid Federation Annual Report

June 1, 2016

Dean N. Williams
ESGF Executive Committee Chair

Abstract

The Earth System Grid Federation (ESGF) experienced a major setback in June 2015, when it experienced a security incident that brought all systems to a halt for more than half a year. However, federation developers and management committee members turned the incident into an opportunity to dramatically upgrade system security and functionality and to develop planning and policy documents to guide ESGF evolution and success. Moreover, despite the incident, ESGF developer working teams continue to make strong and significant progress on various enhancement projects that will help ensure ESGF can meet the needs of the climate community in the coming years.

Brief project summary

ESGF is primarily funded by the Department of Energy's (DOE's) Office of Science (the Office of Biological and Environmental Research [BER] Climate Data Informatics Program and the Office of Advanced Scientific Computing Research Next Generation Network for Science Program), with support from other U.S. federal and international agencies. The federation works across multiple worldwide data centers and spans seven international network organizations to provide users with the ability to access, analyze, and visualize data using a globally federated collection of networks, computers, and software. Its architecture employs a series of geographically distributed peer nodes that are independently administered and united by common federation protocols and application programming interfaces (APIs). The full ESGF infrastructure has now been adopted by multiple Earth science projects and allows access to petabytes of geophysical data, including the Coupled Model Intercomparison Project (CMIP; output used by the Intergovernmental Panel on Climate Change assessment reports), all model intercomparison projects (MIPs; endorsed by the World Climate Research Programme [WCRP]), and the Accelerated Climate Modeling for Energy (ACME; ESGF is included in the overarching ACME workflow process to store model output). ESGF is a successful example of integration of disparate open-source technologies into a cohesive functional system that serves the needs of BER and the global climate science community.

Data served by ESGF includes not only model output but also observational data from satellites and instruments, reanalysis, and generated images.

Objectives

The internationally distributed peer-to-peer ESGF "data cloud" archive represents the culmination of an effort that began in the late 1990s to

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

ESGF was established in 1999 by the DOE, in affiliation with the WCRP Working Group on Coupled Modeling (WGCM), to support CMIP. It provides a community-based infrastructure for climate model diagnosis, validation, intercomparison, documentation, and data access. Through ESGF, scientists are able to analyze general circulation models in a systematic fashion, a process that facilitates model

improvement. Virtually the entire international climate modeling community has participated in CMIP since its inception.

ESGF portals are gateways to scientific data collections hosted at sites around the globe; they allow the user to register[1] and potentially access all data and services within ESGF. Currently, more than 40 portals are in use, including several at Lawrence Livermore National Laboratory (http://pcmdi.llnl.gov). Other portals include DKRZ – https://esgf-data.dkrz.de, CEDA – https://esgf-index1.ceda.ac.uk, IPSL – https://esgf-node.ipsl.upmc.fr, NASA/JPL – https://esgf-node.jpl.nasa.gov, NOAA/ESRL – https://esgf.esrl.noaa.gov, to name a few. ESGF allows international climate research teams to work in highly distributed research environments, using unique scientific instruments, petascale-class computers, and extreme amounts of data. Keys to ESGF's success is its ability to effectively and securely catalog, disseminate, and analyze research results in a globally federated environment. For example, new results generated by one team member are immediately accessible to the rest of the team, and they can annotate, comment on, and otherwise interact with those results.

Desired outcomes and deliverables

The growing international interest in ESGF development efforts has attracted many others who want to make their data more widely available and easy to use. For example, WCRP, which provides governance for CMIP, has now endorsed the ESGF software foundation to be used for ~70 other MIPs, such as obs4MIPs, TAMIP, CFMIP, and GeoMIP. At present, more than 40 projects are supported by ESGF, including data from 25 worldwide climate research centers spanning 21 countries. Virtually all climate researchers now have used ESGF directly or indirectly.

The world is full of large-scale data management and retrieval enterprise systems. In the U.S., for climate science alone, systems include the Global Change Master Directory, the Network-Object Mobile-Agent Dynamic System, the Atmospheric Radiation Measurement Archive, the Regional Climate Model Evaluation System, and the National Aeronautics and Space Administration Distributed Active Archive Centers, to name a few. However, ESGF is the sole distributed data system and the only one to offer interoperability among disparate data sets (i.e., simulations, observations, and reanalysis data) for assessment study.

An important design criterion for ESGF was that it be open source and take advantage of open-source tools. The system leverages web servers and a few key protocols to provide secure communication. ESGF is not only the cornerstone platform for climate research and information sharing and collaboration, but is also designed to be a growing research platform for further investigation into new computing paradigms, including smart, autonomic, self-managing, and self-repairing computing infrastructures.

Management plan

The management of ESGF continues to be a committee-based decision-making process involving a Steering Committee and an Executive Committee[2], led by Dean N. Williams. Members of these committees are normally tied to one or more projects in the community with a specific mission or goal. To better coordinate international efforts, the Executive Committee has written a strategic planning document that describes ESGF development over the next ten years. This living document was presented to the Steering Committee and posted on ESGF's website for community review in May 2015. Other policy and planning documents are listed in the *2015 progress summary* section of this report.

[1] Users have a shared identity and authentication to all sites, thus, they register only once to gain access.
[2] The ESGF governance and governance structure can be found at http://esgf.llnl.gov/governance.html.

The Steering Committee provides a forum for ESGF funding agencies to communicate, engage with, and coordinate their support for ESGF, and to help fashion a common vision for its evolution. The Executive Committee, which accepts guidance from and reports to the Steering Committee, provides general guidance and makes high-level decisions in directing and coordinating the course of ESGF, ensuring that they are consistent with multiple sponsor needs. In this process, individual PIs can make decisions, but their actions must be relayed to the Steering Committee. This governance model drives innovation and quality of services and helps to balance the conflict that exists between new development and day-to-day operations. The Executive Committee is scheduled to hold regular meetings every month at minimum, and the Steering Committee will meet at least twice per year.

Additionally, the ESGF management and approximately 20 development teams have adopted the following tools to improve the coordination among developers and enhance the long-term stewardship and documentation of the software stack:

- For project task organization, the national and international team of ESGF developers and managers use the Atlassian on-demand Confluence and JIRA Agile software tools.
- Git is used for version control of the software stack repository.
- GitHub is used for online bug/enhancement tracking and reporting.

Each code-release version is tagged in Git and released to the community. The latest official release of the 2016 ESGF software stack is version 2.3. The development teams comprise 5 to 20 members; each member commits between 10% and 50% of his or her time. Two leads are assigned to each of the aforementioned teams, and the leads report progress to the ESGF Executive Committee over a designated time period.

Recent annual progress reports and presentations from the ESGF Executive Committee and sub-team leads can be found on the ESGF website (http://esgf.llnl.gov) along with the conference report for the 2016 ESGF Face-to-Face (F2F) Conference Report (http://esgf.llnl.gov/media/pdf/2015-ESGF_F2FConference_report_web.pdf).

Data plan

As discussed at the 2015 ESGF F2F conference, governance and use cases are real issues that determine how requirements affect operations and software development as they relate to projects and data. Therefore, with encouragement from many supporting funding agencies, representatives from a significant fraction of projects utilizing ESGF to disseminate and analyze data attended the conference to provide 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. Data use cases for computing and data science activities that are critical to the community meeting its scientific mission (both as individual projects and as a federation of projects) are summarized in a series of presentations (found at http://esgf.llnl.gov/facetoface.html) and in the 2016 ESGF F2F Conference Report http://esgf.llnl.gov/media/pdf/2015-ESGF_F2FConference_report_web.pdf). **Section 2** of the conference report describes four primary project data use cases and plans—for CMIP6, the Coordinated Regional Climate Downscaling Experiment, Obs4MIPs, and ACME.

Web presence

A web presence is intrinsic to ESGF. It exists on three levels:

1. The ESGF user interface, which enables scientists to host, manage, and share scientific projects seamlessly from any location;
2. Software development websites, such as the Confluence/JIRA Agile development tool and the GitHub code management and software repository; and
3. The official ESGF site, where anyone can learn about ESGF, uncover the latest software events, and download the ESGF software stack (http://esgf.llnl.gov).

In January 2016, the ESGF user interface (or portal front end) officially changed to the CoG content management system. This new web interface allows individual projects to be more identifiable and to customize their interface as needed. At the same time, it recognizes the many sponsors that support the distributed archive and the development of its underlying software. To do so, a single project logo or label identifies each project at the space near the top of each webpage. As the user switches to other sites and projects in the federation to retrieve data or utilize resources, the project logo changes.

Sites not using the official ESGF front end may choose whether to display a project logo for their user interface in a fashion similar to CoG. However, since ESGF is the means by which the data are distributed to the community, these sites must have the icon/logo or words "powered by ESGF" displayed at all times to indicate ESGF's presence.

2015 progress summary

For the federation, 2015 was dominated by the June security incident, which prompted node managers to take all ESGF nodes offline and bring system operations to a halt. While the incident was an obvious setback to the ESGF brand, it was an opportunity for developers to work on hardening, improving, and upgrading the software stack. All ESGF modules have been subjected to dynamic and static security scans and protected against all known common vulnerabilities and exposures. Since the incident, the ESGF team has formulated a plan for executing these scans periodically, including before deployment of every major and minor ESGF release. The team also has installed the latest versions of the underlying software libraries and engines used by ESGF (e.g., Postgres, Apache, Tomcat™, Java™, Python, Django, Solr, and OpenSSL) and implemented a process to maintain and keep them currently moving forward.

Other critical upgrades relate to the ESGF software itself. For example, because all data had to be republished anyway, developers integrated the new metadata archives with the newest Solr version, which offers better performance and more seamless upgrades. Additionally, ESGF decided to stop using and supporting the old ESGF "web front-end" user interface and instead brought the system back up with CoG as the new web user interface. The new software stack, called "ESGF 2.0," incorporates a drastic overhaul of security and functionality.

Perhaps the most important outcome involved the formulation of planning documents for ESGF 2.0 deployment across all nodes in the federation and for bringing the system back to regular operations in the spring of 2016. These living planning and policy documents (listed below) help to guide the evolution and ensure the success of ESGF:

- ESGF Governance Policy (http://esgf.llnl.gov/governance.html);
- Logo Requirement and Usage Guidelines (http://esgf.llnl.gov/logo_requirements.html);
- ESGF Strategic Roadmap (http://esgf.llnl.gov/media/pdf/2015-ESGF-Strategic-Plan.pdf);
- Software Security Plan (http://esgf.llnl.gov/media/pdf/ESGF-Software-Security-Plan-V1.0.pdf);
- ESGF Implementation Plan (http://esgf.llnl.gov/media/pdf/ESGF-Implementation-Plan-V1.0.pdf);

ESGF is committed to using well-established quality standards as a foundation for new subcomponent design and providing the community of users with the most up-to-date documentation and training possible. Such standards are already being employed to ensure that current subcomponent offerings meet best practices in software design, engineering, and implementation. ESGF is systematically and rigorously developing, testing, evaluating, and documenting its subcomponents and associated tasks via teleconferences, face-to-face workshops and conferences, written reports, and journal publications.

To expedite the execution of key ESGF features in preparation for the CMIP6 archive—an important driver for ongoing development—we have prioritized the development efforts of each subcomponent in the *ESGF Implementation Plan*. The first CMIP6 data sets are scheduled to arrive at the beginning of 2017, and thus most short-term ESGF subcomponent implementation activities must be completed by December 31, 2016.

ESGF network activities can be explored at the International Climate Network Working Group (ICNWG) website: http://incwg.llnl.gov and the 2016 ESGF F2F Conference Report.

Project progress toward objectives

All ESGF working teams made substantial progress in 2015, starting with the complete overhaul of every component in the ESGF software stack for security compliance and better usability. Looking forward into 2016 and beyond, ESGF developers will be fully engaged in a wide range of activities that will expand the system's functionality, reliability, and performance to better meet the evolving needs of the climate community in the coming years. Ongoing working teams and their tasks include:

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

- ***Provenance Capture Working Team:*** Implementing a framework for capturing detailed information through data publication, analysis, and generation of derived products.
- ***Publication Working Team:*** Working on finalizing metadata requirements in support of CMIP6, as well as enabling a new publishing service to support "long-tail" data providers.
- ***Quality Control Working Team:*** Improving the quality of ESGF user services with regard to additional (external) documentation and coordinating the implementation of errata and citation pilots.
- ***Replication and Versioning Working Team:*** Setting up an infrastructure to execute massive automatic data transfer and publishing across major climate centers, using tools such as Synda and Globus.
- ***Software Security Working Team:*** Supporting ESGF software security scans, which are critical for minor and major releases of the ESGF software stack.
- ***User Support Working Team:*** Redefining tools and services that can be better integrated with the new ESGF 2.x software stack.

For a much more detailed report of the activities of each of the ESGF working teams, please see the 2016 5th Annual ESGF Face-to-Face Conference Report[3], **Appendix G, p. 85**.

For additional information regarding the ESGF working teams or the annual report, please contact Dean N. Williams at williams13@llnl.gov.

Appendix 1: List of publications and presentations primarily about ESGF or one of its components

Journal papers and conference proceedings:

1. Marcin Plociennik et al., "Two-level dynamic workflow orchestration in the INDIGO DataCloud for large-scale, climate change data analytics experiments," ICCS 2016 paper.
2. Jim McEnerney et al., "Parallelization of Diagnostics for Climate Model Development," Accepted by the *IEEE Journal of Software Engineering and Applications* (2016), http://www.scirp.org/journal/jsea, http://dx.doi.org/10.4236/jsea.2016.*****.
3. C.P. Covey et al., "Metrics for the Diurnal Cycle of Precipitation: Toward Routine Benchmarks for Climate Models," Accepted by *Journal of Climate* (2016), http://journals.ametsoc.org/doi/pdf/10.1175/JCLI-D-15-0664.1.
4. Dean N. Williams et al., "A Global Repository for Planet-Sized Experiments and Observations," *Bulletin of the American Meteorological Society* (2016), http://dx.doi.org/10.1175/BAMS-D-15-00132.1.
5. D. N. Williams, "Better tools to build better climate models," Cover of *Eos* **97** (2016), https://eos.org/project-updates/better-tools-to-build-better-climate-models.
6. Dean N. Williams et al., *Working Group on Virtual Data Integration: A Report from the August 13–14, 2015, Workshop*, U.S. Department of Energy Office of Science, LLNL-TR-678127-REV-1 (2016), https://e-reports-ext.llnl.gov/pdf/801504.pdf.
7. Dean N. Williams et al., "Strategic Roadmap for the earth system grid federation," in *Big Data, 2015 IEEE International Conference Proceedings* (Santa Clara, CA, USA, 2015), pp. 2182–2190, http://ieeexplore.ieee.org/xpl/articleDetails.jsp?reload=true&arnumber=7364005.
8. John. L. Schnase et al., "Big Data Challenges in Climate Science," *IEEE Geoscience and Remote Sensing Magazine* (2015).
9. Matthew B. Harris et al., "Nerd Herding: Practical Project Management in the Field," in *Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science* (San Francisco, CA, USA, 2015), pp. 123–126,

[3] http://esgf.llnl.gov/media/pdf/2015-ESGF_F2FConference_report_web.pdf.

http://www.iaeng.org/publication/WCECS2015/WCECS2015_pp123-126.pdf.
10. C. Palazzo et al., "A Workflow-Enabled Big Data Analytics Software Stack for eScience," in *The Second International Symposium on Big Data Principles, Architectures & Applications* (BDAA 2015), (Amsterdam, the Netherlands, 2015).
11. Hashim Iqbal Chunpir et al., "Evolution of e-Research: From Infrastructure Development to Service Orientation," in *HCI International Conference* (2015).
12. Chris A. Mattmann et al., "Next Generation Cyber-infrastructure to Support Comparison of Satellite Observations with Climate Models," in *Conference on Big Data from Space*, *Research, Technology and Innovation* (Frascati, Italy, 2015).
13. Dean N. Williams, "Visualization and Analysis Tools for Ultrascale Climate Data," *Eos, Transactions American Geophysical Union* **95** (42), 377–378 (2014), http://onlinelibrary.wiley.com/doi/10.1002/2014EO420002/abstract.
14. D.N. Williams et al., "Department of Energy Strategic Roadmap for Earth System Science Data Integration," in *Big Data, 2014 IEEE International Conference Proceedings* (Washington D.C., USA, 2014), pp. 772–777, http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7004304.
15. Matthew Harris, "Webengine," in *Proceedings of the World Congress on Engineering and Computer Science 2014* (San Francisco, CA, USA, 2014), vol. I, pp. 131–135, http://www.iaeng.org/publication/WCECS2014/.
16. Sandro Fiore et al., "Ophidia: A full software stack for scientific data analytics," in *IEEE High Performance Computing & Simulation* (2014), pp. 343–350, http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6903706.
17. L. Cinquini et al., "The Earth System Grid Federation: An open infrastructure for access to distributed geospatial data," *Future Generation Computer System* **36**, 400–417 (2014), http://www.sciencedirect.com/science/article/pii/S0167739X13001477.
18. S. Fiore et al., "Ophidia: A Full Software Stack for Scientific Data Analytics," in *Proc. of the 2014 International Conference on High Performance Computing & Simulation* (Bologna, Italy, 2014), pp. 343–350, ISBN: 978-1-4799-5311-0.

Reports:

1. Wes Bethel et al., *Report of the DOE Workshop on Data Management, Visualization, and Analysis of Experimental and Observational Data*, U.S. Department of Energy Office of Science. (In process 2016).
2. D.N. Williams et al., *5th Annual Earth System Grid Federation Face-to-Face Conference Report,* U.S. Department of Energy Office of Science, DOE/SC-0181, LLNL-TR-689917 (2016), http://esgf.llnl.gov/media/pdf/2015-ESGF_F2FConference_report_web.pdf.
3. Dean N. Williams et al., *Working Group on Virtual Data Integration: A Report from the August 13–14, 2015, Workshop*. U.S. Department of Energy Office of Science, DOE/SC-0180 (2016), http://www.osti.gov/scitech/biblio/1239196/.
4. Sandrine Bony et al., *Report on the Nineteenth Session of the Working Group on Coupled Modelling (WGCM)* (2016), http://www.wcrp-climate.org/images/modelling/WGCM/WGCM19/documents/WGCM19_reportv4.pdf.
5. David Moulton et al., *Building a Cyber-infrastructure for Environmental System Science: Modeling Frameworks, Data Management, and Scientific Workflows; Workshop Report,* U.S. Department of Energy Office of Science, DOE/SC-0178. (2015), http://doesbr.org/ESS-WorkingGroups/.
6. Eli Dart et al., *Biological and Environmental Research Network Requirements Review Final Report* (2015), http://www.es.net/assets/pubs_presos/BER-Net-Req-Review-2015-Final-Report.pdf.
7. D.N. Williams et al., *4th Annual Earth System Grid Federation and Ultrascale Visualization Climate Data Analysis Tools Conference Report*, Lawrence Livermore National Laboratory,

Livermore, CA (2015), LLNL-TR-666753, http://aims-group.github.io/pdf/2014-ESGF_UV-CDAT_Conference_Report.pdf.

8. D.N. Williams et al., *3th Annual Earth System Grid Federation and Ultrascale Visualization Climate Data Analysis Tools Conference Report*, Lawrence Livermore National Laboratory, Livermore, CA (2014), LLNL-TR-650500, http://uvcdat.llnl.gov/media/pdf/ESGF_UV-CDAT_Meeting_Report_December2013.pdf.

Presentations:

1. The latest presentations for the 2015 ESGF F2F Conference can be found on the following website: http://esgf.llnl.gov/2015-F2F.html.
2. The latest presentations for the 2014 ESGF & UV-CDAT F2F Conference can be found on the following website: http://esgf.llnl.gov/2014-F2F.html.

Posters:

1. "Distributed Computation Resources for Earth System Grid Federation (ESGF)," Daniel Duffy et al., at the *2014 American Meteorological Society Conference* (Phoenix, AZ, USA, 2015).
2. "Distributed Computation Resources for Earth System Grid Federation (ESGF)," Daniel Duffy et al., at the *2014 American Geophysical Union Conference* (San Francisco, CA, USA, 2014).

Awards:

1. 2015 Federal Laboratory Consortium (National) Award for Advancing Federal Research and Technology for the Ultrascale Visualization for Climate Data Analysis Tools (UV-CDAT).
2. 2014 Federal Laboratory Consortium (Far West Regional) Award for Advancing Federal Research and Technology for the Ultrascale Visualization for Climate Data Analysis Tools (UV-CDAT).
3. 2013 Federal Laboratory Consortium (Far West Regional) Award for Advancing Federal Research and Technology for the Earth System Grid Federation (ESGF).

Appendix 2: Science highlights reported during the past year

1. Data used from the ESGF archives have helped to generate ~1600 peer-reviewed CMIP5 articles (http://cmip.llnl.gov/cmip5/publications/) and over ~600 peer-reviewed CMIP3 articles (http://www-pcmdi.llnl.gov/ipcc/subproject_publications.php).