LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-TR-780063

Annual Earth System Grid Federation 2019 Progress Report

G. M. Abdulla

July 1, 2019

Disclaimer

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

Annual Earth System Grid Federation 2019 Progress Report

July 1st, 2019

Ghaleb Abdulla (DOE Lawrence Livermore National Laboratory)

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.

Executive Summary

The ESGF facilitates advancements in geoscience by providing:

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

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

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

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

ESGF is continuously adding new data sets based on community requests and needs and in the future, it intends to widen its scope to include other climate-related datasets such as downscaled model data, climate predictions from both operational and experimental systems, and other derived data sets. Over the next few years, we propose to:

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

The information emerging from collaboration between interagency partner meetings influences requirements, development, and operations. In December of 2018, representatives from a significant fraction of projects utilizing ESGF to disseminate and analyze data attended the sixth annual ESGF Face-to-Face (F2F) Conference (https://esgf.llnl.gov/esgf-media/pdf/2018_8th_Annual_ESGF_Conference_Report_final.pdf). Attendees provided important feedback regarding current and future community data use

cases. Discussions focused on maintaining essential operations while developing new and improved software to handle ever-increasing data variety, complexity, velocity, and volume. Focusing on federation resiliency and reaffirming the consortium's dedication to extend the existing capabilities needed for large-scale data management, analysis, and distribution of highly visible community data and managed resources, the ESGF Executive Committee decided to reorganize the working teams for better synergy and greater alignment. See Table 1 for the working team list and representatives approved in the F2F meeting December of 2018.

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

Team	Team Leads and Funding Agencies / Institutions	Description
1. User Interface, Search, and Dashboard Working Team	Sasha Ames (LLNL), 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)	Identity management and access control to enable resources (data and compute) to have appropriate access restrictions
4. Installation and Software Security Working Team	Sasha Ames (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), Sebastien Gardoll (IPSL), Jason Boute (LLNL)	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	Katharina Berger (DKRZ)	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), Ghaleb Abdulla (LLNL), and Sandro Fiore (CMCC)	Research in the applicability of various ML techniques and development of tools/analysis capabilities for domain scientists

ESGF recent statistics show an active user community; between January 1st 2016 and May 29, 2019 ESGF was accessed by over 15,000 unique IP addresses and a total of 1.1 PB of data was downloaded from the main LLNL ESGF data node (see Figure1).

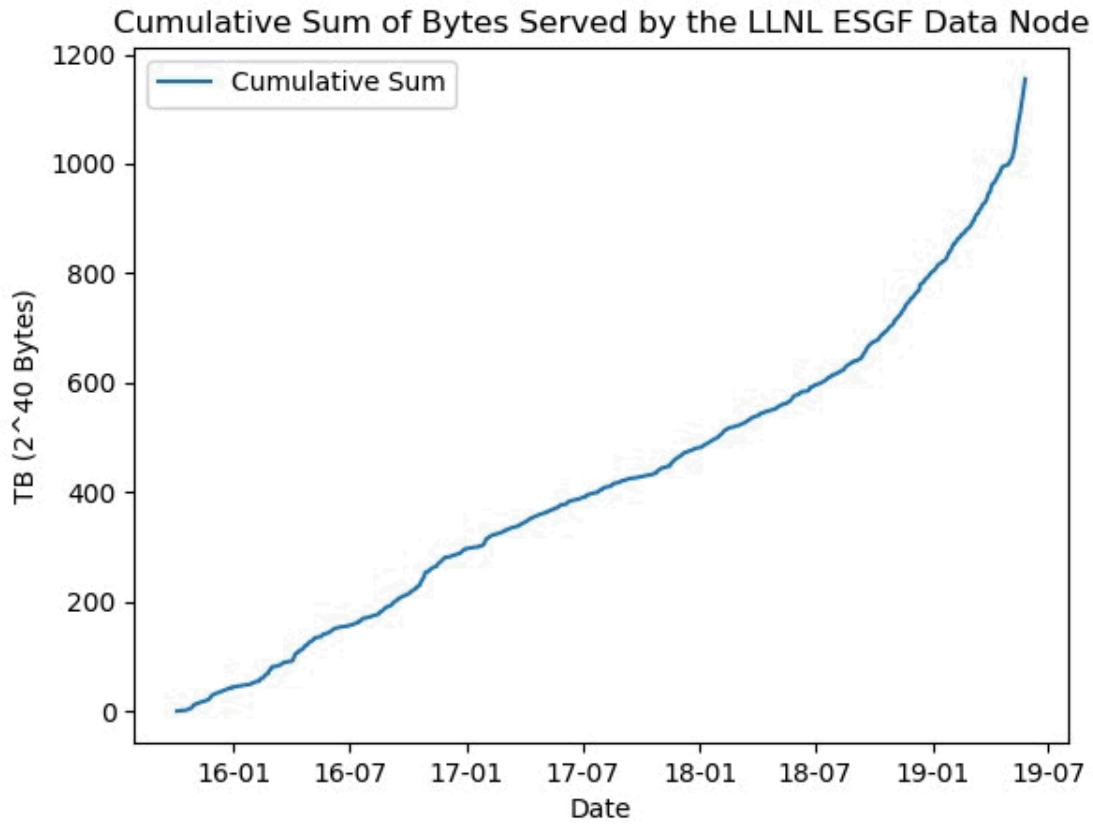

Figure 1: Cumulative sum of bytes served by LLNL ESGF data node. Total amount of bytes downloaded exceeds 1 PB.

Working Team Reports

1. User Interface, Search, and Dashboard Working Team

- 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. Added enhancements to the search interface to speed up response time, fixed a gridftp performance issue, and added capabilities to generate general statistics from log files and populate an ESGF dashboard.
- 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 five major releases (v3.10 through v3.14) 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.
- We developed a means to capture coarse-grained data usage metrics from apache httpd log files as backup replacement / supplemental service to the esgf-dashboard. This service is capable of producing usage rates of data in real-time, in addition to providing historical information from a bulk-ingest of raw log data. We use the Prometheus framework to serve the data and the Grafana dashboard for data visualization.
- We have added a node status view to CoG that reports which data nodes are on/offline and integrated with search results
- Several issues with the Globus integration have been corrected in CoG
- Solr Cloud index node. A new proposed architecture for the ESGF search services was prototyped and unveiled at the F-2-F conference (Figure 2). 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.

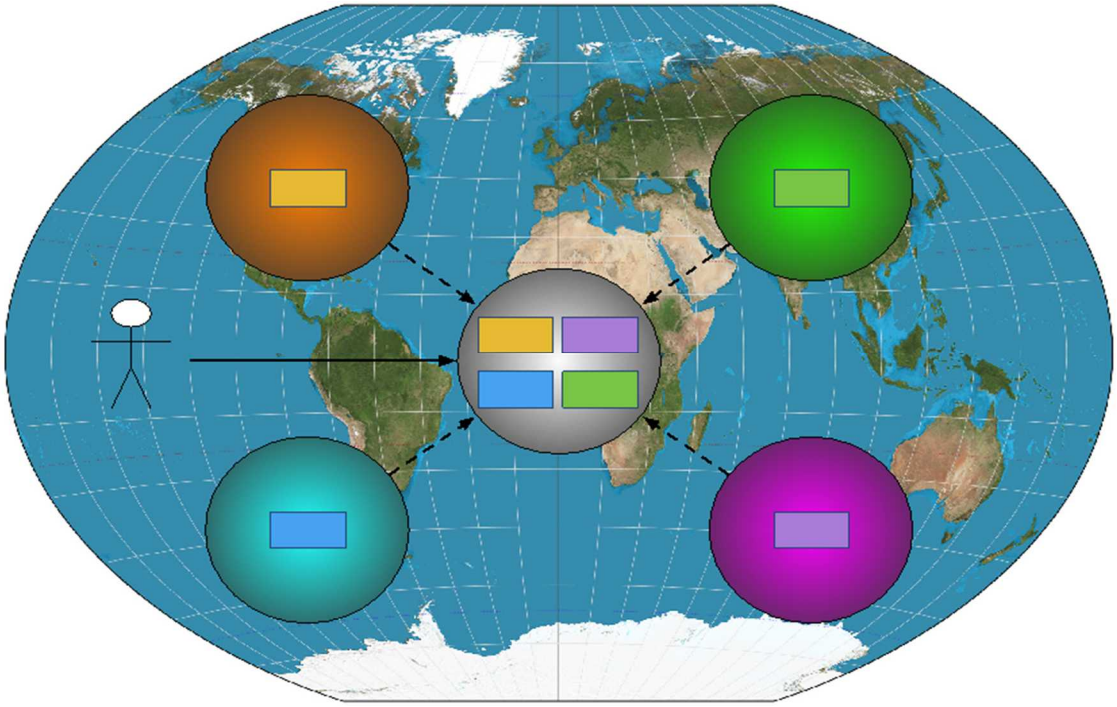

Figure 2: Proposed next-generation ESGF search architecture, where metadata catalogs are harvested from individual ESGF nodes into a "super-index" node.

2. Compute Working Team

This was a really productive year for the CWT Team which grew considerably in maturity.

- The first Compute Grand Challenge was designed during the 2018 F2F, the goal of this Challenge was to deliver production quality compute services by July 1st. Namely the services to be delivered were: subset and aggregation. The Challenge successfully completed on July 1st. 3 institutions provided compute services to the public: LLNL (USA), NASA (USA) and CMCC (Italy)
- A certification process was established to ensure regulation of services provided on behalf of ESGF and preserve ESGF's reputation. ESGF will certify servers and operators (see: https://docs.google.com/document/d/1pxz1Kd3JHfFp8vR2JCVBfApbsHmbUQQstifhGNdc6U0/edit)
- Version 2.0 of the Python-based API was officially released at the same time, allowing users to programmatically call services on certified servers. See: https://github.com/ESGF/esgf-compute-api
- LLNL completed its Docker/Kubernetes-based architecture (https://github.com/ESGF/esgf-compute-wps). Version 2.0 was deployed on a dedicated cluster (accessible at: https://aims2.llnl.gov.) as their official production compute server.
- NASA's Goddard Space Flight Center (GSFC) re-implemented its Earth Data Analytics Services backend as a Python/DASK-based solution. The older (EDAS) version went officially online but will be replaced by this version after receiving clearance from NASA's security.
- CMCC polished their Ophidia cluster implementation along with user facing wps, as part of the Compute Grand Challenge
- 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.
- CWT's work was also integrated into Canada's public facing website: https://climatedata.ca
- Analytics got an upgrade both in data I/O and visualization with the release of CDAT 8.1 (https://github.com/CDAT/cdat/releases/tag/v8.1)

- Data exploration was made a bit easier with the release of VCDAT 2.0 a Jupyter-based front end to CDAT. (https://github.com/CDAT/jupyter-vcdat)

3. Identity, Entitlement, and Access Working Team

The focus for the last year 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.

4. Installation and Software Security Working Team

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

- 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 and beyond 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.
- The final releases of the ESGF bash-based installation v2.8.1-4 were completed in early 2019.
- The effort to complete ESGF 3.0 based on Python provided the experience to the rapid completion of Ansible playbooks to install the ESGF Software Stack. The playbooks reduced the amount of code needed for installation by a factor of 6. The releases using Ansible playbooks start at major version 4 as they shift the deployment of several of the components in a manner incompatible with previous major version, and to date, we have released versions 4.0.1-3.

5. Containers Working Team

- 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.
-
- 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.

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

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.
- LLNL Replication and Publishing: we have continuous operation of Synda and have implemented an automated replica publishing pipeline at LLNL. We have published more than 320000 CMIP6 datasets to date.

7. Node Manager and Tracking/Feedback Notification Working Team

- 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.
- 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.
- To replace the monitoring functionality of the Node Manager, we have integrated Prometheus server metric exporter into the software stack deployment using Ansible. Prometheus provides an API, of which other ESGF clients and services can use to determine the online status of nodes in federation. To facilitate

monitoring specifically of (THREDDS) data services, we have implemented external monitors of the service to feed data to the API ass

- 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. We have developed a prototype integration of the User Notification framework in CoG to manage user subscriptions (add/list/delete).

8. Publication, Quality Control, and Metadata Working Team

- 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.

10. Machine Learning Working Team

- 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.
- Learning to Focus and Track Extreme Climate Events. We tackle extreme climate event tracking problem. It has unique challenges to other visual object tracking problems, including wider range of spatio-temporal dynamics, blur boundary of the target, and shortage of labeled dataset. In this work, we proposed a simple but robust end-to-end model based on multi-layered ConvLSTM, suitable for the climate event tracking problem. The algorithm first learns to imprint location and appearance of the target at the first frame in auto-encoding fashion, and then, the learned feature is consumed by the tracking module to track the target in subsequent time frames. To tackle the data shortage problem, we propose data augmentation based on conditional GAN. Extensive experiments show that the proposed framework significantly improves tracking performance on hurricane tracking task over several state-of-the-art methods.
- Focus and Track: pixel-wise spatio-temporal hurricane tracking**.** We propose a pixel-wise extreme climate event tracking framework to track a target in the multiple moving objects scenario. We applied our model to tackle the challenging hurricane tracking problem.
  The proposed framework consists of two sub-models based on multi-layered ConvLSTM: a focus learning and a tracking model. Focus learning model learns location and appearance of target at first frame of video with auto-encoding fashion, and then, learned feature is fed into tracking model to follows the target in consecutive time frames. Extensive experiments show that the proposed tracking framework significantly outperforms against state-of-the-art tracking algorithms.

Publications and presentations

1- ESGF XC committee, "8th Annual Earth System Grid Federation Face-to-Face Conference Report", https://esgf.llnl.gov/esgf-media/pdf/2018_8th_Annual_ESGF_Conference_Report_final.pdf
2- Ghaleb Abdulla, "Tales of Big and Small Data", at *the Round table on Data for AI, success stories and frustration stories*, DOE office of Science Workshop, June, 5 2019
3- Ghaleb Abdulla, "The 80/20 rule: can and should we break it using efficient data management tools?", *DATAWorks workshop,* April 10, 2019
4- Ghaleb Abdulla, "The Earth System Grid Federation (ESGF) Overview", *Open Science Workshop for the DOE Office of Defense Nuclear Nonproliferation R&D,* February 7, 2019
5- Soo Kyung Kim. et al. , "Learning to Focus and Track Extreme Climate Events" , submitted to *BMVC* 2019
6- Soo Kyung Kim. et al., "Focus and Track: pixel-wise spatio-temporal hurricane tracking", *AI for Climate Change, ICML workshop,* June 2019