# INTERCONNECTING
# A World of Petabytes

As the demand for global, broad-based climate change projections has grown, effectively managing the vast accompanying volumes of data represents a major challenge for computational scientists. In the area of understanding and predicting climate change and extreme weather events, advanced tools are required to securely store, manage, access, analyze, visualize, and process enormous and distributed data sets.

In 2017, Lawrence Livermore researchers received an R&D 100 Award for their work on the Earth System Grid Federation (ESGF)—a virtual, collaborative environment that links climate centers and users around the world to models and data via a computing grid powered by the world's supercomputing resources and the Internet. ESGF facilitates Earth system science and also provides an essential infrastructure for scientists to evaluate models through a common interface, regardless of the data's location. Although many large-scale data management systems exist, none rival ESGF's global accessibility and scientific capability.

The Earth System Grid Federation (ESGF) enables users to access global atmospheric, land, ocean, and sea-ice data generated by satellite, and in situ observations and complex computer simulations. Using modeling and data centers worldwide, ESGF forms a collaborative and interactive network for sharing massive amounts of data from climate and other science domains.

## Welcome to the Federation

ESGF's roots go back nearly two decades, when Livermore computer scientist Dean N. Williams collaborated with Argonne National Laboratory's Ian Foster and Steve Hammond at the National Center for Atmospheric Research to apply grid computing to climate science applications that required quickly transferring massive amounts of data. The result was a climate-dedicated cyber environment called the Earth System Grid (ESG), which they used to move data among Department of Energy (DOE) sites.

Before ESG, the climate community comprised various groups and organizations, each with its own methods for creating workflows, generating data file formats and conventions, and storing information. If researchers wanted to study the results from other groups, they first had to spend significant resources converting the data into formats with which they were familiar. This task was no less daunting for Livermore's Program for Climate Model Diagnosis and Intercomparison (PCMDI), which had been evaluating dozens of climate models from institutions all over the world since 1989.

Using grid computing to facilitate PCMDI activities, the ESG team was able to unify data formats and conventions, collect the standardized climate models, and disseminate the information throughout the climate community. Advancements in data management, distributed data sharing, and model archiving led to the full integration of ESG into PCMDI's Coupled Model Intercomparison Project Phase 3 (CMIP3), which was extensively used in the fourth assessment report by the Intergovernmental Panel on Climate Change (IPCC). The 2007 Nobel Peace Prize was co-awarded to IPCC for this work.

"CMIP3 was originally thought to consist of 1 terabyte of data. Ultimately, it was 35 times larger. The next phase, which was slated to be 100 terabytes, turned out to be nearly 2 petabytes," says Williams. "To retrieve data for CMIP3, we sent out massive disk file systems to the climate community that they then shipped back to us. Since this process was not scalable for petabytes of data, we sent our collaborators software rather than disks." Thus, in 2011, ESGF was born—a federated, distributed archive for Earth system data.

## A Prospering Federation

ESGF has become the largest-ever platform for collaborative data on Earth system science. At its heart is a peer-to-peer network of nodes distributed across several countries and united by common protocols and interfaces. Node sites span the globe including at NASA, the National Oceanic and Atmospheric Administration (NOAA), and Lawrence Livermore in the United States; the German Climate Computing Centre, the Institut Pierre Simon Laplace, and the Centre for Environmental Data Analysis in Europe; as well as institutions in Australia, China, and elsewhere abroad.

The federation's interoperability enables users to access global atmospheric, land, ocean, and sea-ice data generated by satellite, and in situ observations and complex computer simulations. With ESGF's networks, computers, and software, scientists can access and manage Earth system data more efficiently and robustly through newly developed user interfaces, distributed or local search protocols, federated security, server-side analysis tools, direct connections to high-performance networks, an open computation environment, and other community standards.

To guarantee data validity, ESGF uses quality control algorithms to ensure that the proper formats, units of measure, and conventions are used. Data are also given unique digital object identifiers so information can be tracked back to the source. Making the data easily traceable means researchers can reproduce the models of other researchers to test their repeatability. As models are rerun under different conditions, all variations of the output can be published concurrently to ESGF. The environment also uses versioning, which enables any updates to model data to supersede previous output while at the same time preserving the output history. As a result, data comparisons are possible between various outputs. Williams says, "Now that all these disparate data sets are in one place, scientists can for the first time do comparisons between observations, simulations, and reanalysis data. They can also utilize the data right away."

## Federation of the Future

Today, ESGF serves more than 25,000 users, including scientists and policymakers, and provides them access to a staggering 5 petabytes of data—and this number is continually increasing. Aside from work performed for IPCC, ESGF also supports at least 25 other projects, including DOE's new Energy Exascale Earth System Model (E3SM)—the most advanced Earth system model ever created. E3SM is designed to run on future exascale computing systems.

Expanding upon ESGF's capabilities, developers are now looking into machine learning as a way of searching for important connections, patterns, and occurrences, including those associated with naturally occurring climate oscillations such as an El Niño or La Niña, buried deep within the vast petabytes of data available. Another feature under development would allow ESGF to include other science domains. In a prototype version, ESGF has incorporated epidemiology and hydrology data, enabling researchers to study how changes in climate may affect the spread of disease in certain regions. Williams says, "Putting together these different data sets provides a bigger picture of the world and the potential challenges we face. Economic models and other types of models could also be used in this way. The future of ESGF and this type of modeling are promising."

*—Dan Linehan*

The ESGF development team is led by Laboratory scientist Dean N. Williams and brings together an international cadre of experts in high-performance computing and climate science to deliver globally accessible Earth system data. (Photo by George Kitrinos.)

**Key Words**: climate science, Coupled Model Intercomparison Project Phase 3 (CMIP3), data sets, Earth System Grid (ESG), Earth System Grid Federation (ESGF), Energy Exascale Earth System Model (E3SM), federated distributed archive, grid computing, Intergovernmental Panel on Climate Change (IPCC), large-scale data management system, petabyte, Program for Climate Model Diagnosis and Intercomparison (PCMDI).

***For further information contact Dean Williams (925) 423-0145 (williams13@llnl.gov).***