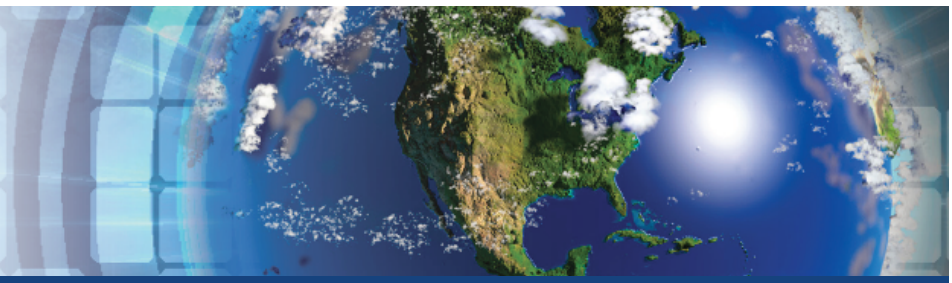
ESGF
Earth System Grid Federation
A Peer-to-Peer Enterprise



esgf.llnl.gov

What is ESGF?

The Earth System Grid Federation (ESGF) peer-to-peer (P2P) enterprise system is one of the largest-ever collaborative data efforts in Earth system science. Led by the U.S. Department of Energy's (DOE) Office of Biological and Environmental Research (BER), this international multi-agency federation—which heavily relies on European partners—develops, deploys, and maintains software to facilitate advancements in geophysical science. ESGF's open-source, operational code base disseminates model simulation, observational, and reanalysis data for research assessments and model validation via secure storage and dissemination of petabytes of data.

With its collection of independently funded national and international projects, ESGF manages the first-ever decentralized database for accessing geophysical data at dozens of federated sites. Its widespread adoption, federation capabilities, broad developer base, and focus on Earth system science data distinguish ESGF from other collaborative knowledge systems. Currently serving more than 25,000 users, the total ESGF archive manages over 5 petabytes of Earth system science datasets from more than 25 projects and ~70 model intercomparison projects; furthermore, it supports over 700,000 datasets from worldwide laboratories and universities (see "ESGF By the Numbers", next page).

ESGF provides the resources essential for global-scale research. Virtually all Earth system science researchers worldwide use ESGF to discover, access, and analyze data. In fact, many of today's most recognized Earth system science projects employ the valuable software and services developed by the ESGF team and its community. The federation will continue to expand access to relevant data integrated with tools for analysis and visualization that are supported by the necessary hardware and network capabilities to interpret peta- and exascale scientific data. **Figure 1** (next page) illustrates the ESGF infrastructure that enables users to access data and analysis tools through a system of distributed nodes.

Along with DOE BER, U.S. participants in ESGF are the National Aeronautics and Space Administration (NASA), National Oceanic and Atmospheric Administration (NOAA), and National Science Foundation (NSF). International laboratories in the federation include the European Network for Earth System Modelling (ENES), Australian National University's (ANU) National Computational Infrastructure (NCI), German Climate Computing Centre (DKRZ), Institut Pierre-Simon Laplace (IPSL), and the Centre for Environmental Data Analysis (CEDA).

ESGF Missions

- Support Coupled Model Intercomparison Project (CMIP) activities and prepare for future assessments. CMIP is an initiative of the World Climate Research Programme.
- Develop data and metadata facilities for inclusion of observations and reanalysis products for CMIP use.
- Enhance and improve current Earth system science research infrastructure capabilities through involvement of the software development community and through adherence to sound software principles.
- Foster collaboration across agency and political.
- Integrate and interoperate with other software designed to meet ESGF objectives such as those developed by NASA, NOAA, and the European ENES.
- Create software infrastructure and tools that facilitate scientific advancements.

Why is ESGF Important?

ESGF has transformed Earth system data into community resources available within a virtual, collaborative environment that links climate centers and users around the world to models and data via a computing grid power by the world's supercomputing resources and the Internet. Together with its international partners, ESGF provides a cohesive functional system that allows equal access to large disparate datasets that otherwise would have been accessible only with great difficulty, if at all, to some researchers. Through team efforts, ESGF's essential infrastructure enables scientists to evaluate models through a common interface, regardless of the data's location.

ESGF-Merged Independent Software Applications

- National and international network infrastructure integrating the world's climate model and measurement archives.
- Shared resources across multiple centers for high-performance computing and storage of tens of petabytes of transportable data.
- Easy-to-use and secure, federated web-based application programming interface and data infrastructure.
- Flexible infrastructure allowing participants to customize parameters.
- High-performance search, analysis, and visualization tools.
- Access to a broad set of data and tools for comparative and exploratory analysis.
- Virtual collaborative environment for analysis tasks demanding large, varied datasets.

The climate modeling community has devoted many resources to produce large-scale computer simulations and collect vast amounts of observational data for producing assessment reports (ARs) such as the U.S. National Climate Assessment and the Intergovernmental Panel on Climate Change (IPCC) ARs.

There are many large-scale enterprise data management and retrieval systems. U.S. geophysical systems alone include the Global Change Master Directory, Network-Object Mobile-Agent Dynamic System (NOMADS), and the NASA Distributed Active Archive Centers (DAACs). However, none of them is a distributed data system, and none allows interoperability among different datasets (see **Fig. 2**, this page). Furthermore, an engine such as Google Earth Engine ingests only the most popular datasets of an area and is not interested in the long-tail data publishing used by ESGF and other scientific communities where individual scientists share their richly diverse and heterogeneous small datasets.

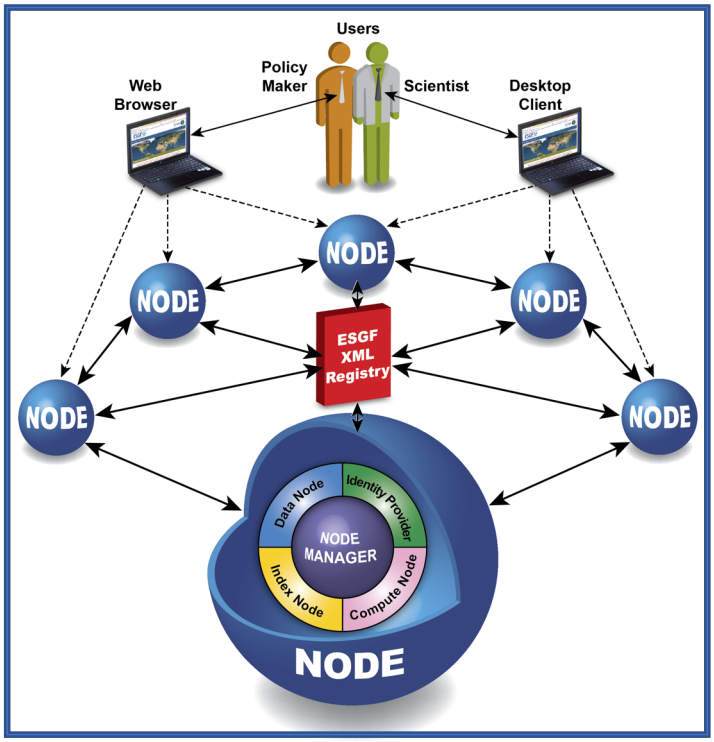


Fig. 1. ESGF ensures equal access to large disparate datasets. The ESGF infrastructure enables scientists to evaluate models, understand their differences, and explore the impacts of geophysical disturbances through a common interface, regardless of data location.

ESGF Comparison to Other Archives	ESGF	Google Earth Engine	NASA DAACs	NOAA NOMADS
Current capabilities / *Future capabilities*				
Data Management	■	■	■	■
Data Transfer	■	■	■	■
Analysis and Visualization	■	■	■	
Network	■	■	■	
Security	■	■	■	
Compute Facility (server-side)	▪	■	▪	
Distributed Search	■		■	
Provenance Capture	▪		■	
Dynamic Resources	▪	■		
Machine Learning	▪	■		
Federation	■			
Data Citation	■			
Long-tail Publication	▪			

Fig. 2. Summary of how ESGF compares with other governmental and commercial archives. Through its proposed work, ESGF will add capabilities for provenance capture, server-side computing access, dynamic resource management, and long-tail publishing—making ESGF the most capable system supporting Earth system science research.

ESGF By the Numbers

Usage Demographics

- Supports **>700,000** datasets from universities as well as national and international laboratories. **~4 million** datasets downloaded.
- Manages **>5 PB** of data in the total ESGF federated archive, which is expected to expand to >40 PB of uncompressed data, distributed across >25 projects and **~70** model intercomparison projects (MIPs).
- Services **18** highly visible national and international geophysical data products, including CMIP3, CMIP5, and soon CMIP6.

150,824
CMIP5 total number of datasets

4,261.921 TB
CMIP5 total data volume

365
Obs4MIPs total number of datasets

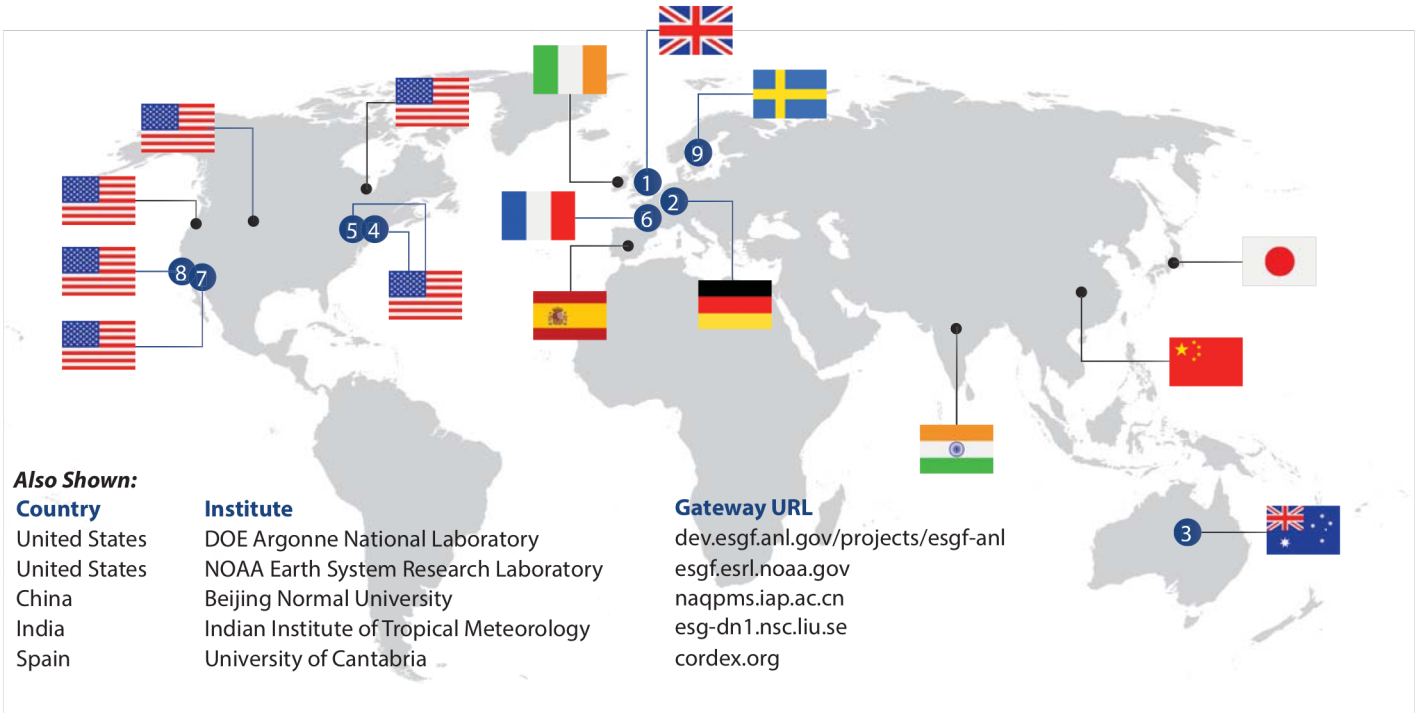
0.285 TB
Obs4MIPs total data volume

68.709
CORDEX total number of datasets

59.813 TB
CORDEX total data volume

Major ESGF Node Sites

Institution	Gateway URL	Version	Country	Project(s)	Contact
1 CEDA	esgf-index1.ceda.ac.uk	2.4.0	U.K.	CMIP5, CORDEX, Obs4MIPs, SPECS, ESA CCI, EUCLEIA, CLIPC	alan.iwi@stfc.ac.uk
2 DKRZ	esgf-data.dkrz.de	2.4.0	Germany	CMIP5, CORDEX, Obs4MIPs, ISI-MIP	berger@dkrz.de
3 ANU NCI	esgf.nci.org.au	2.4.0	Australia	CMIP5	ben.evans@anu.edu.au
4 NOAA GFDL	esgdata.gfdl.noaa.gov	2.4.0	U.S.	CMIP5, ncpp2013, Obs4MIPs	hans.vahlenkamp@noaa.gov
5 NASA GSFC	esgf.nccs.nasa.gov	2.4.0	U.S.	CMIP5, Obs4MIPs, Ana4MIPs, NEX-GDDP, NEX-DCP30, CREATE-IP	daniel.q.duffy@nasa.gov
6 IPSL	esgf-node.ipsl.upmc.fr	2.4.0	France	CMIP5, CORDEX, Obs4MIPs	sebastien.denvil@ipsl.jussieu.fr
7 NASA JPL	esgf-node.jpl.nasa.gov	2.4.0	U.S.	Obs4MIPs, GASS-YoTC, CMAC	luca.cinquini@jpl.nasa.gov
8 DOE LLNL	esgf-node.llnl.gov	2.4.0	U.S.	CMIP5, CMIP3, input4MIPS, ACME	sasha@llnl.gov
9 LiU	esg-dn1.nsc.liu.se	2.4.0	Sweden	CMIP5, CORDEX, SPECS, CLIPC	pchengi@nsc.liu.se



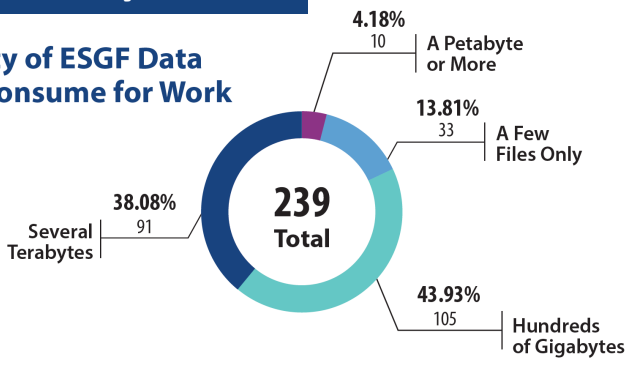
Also Shown:

Country	Institute	Gateway URL
United States	DOE Argonne National Laboratory	dev.esgf.anl.gov/projects/esgf-anl
United States	NOAA Earth System Research Laboratory	esgf.esrl.noaa.gov
China	Beijing Normal University	naqpms.iap.ac.cn
India	Indian Institute of Tropical Meteorology	esg-dn1.nsc.liu.se
Spain	University of Cantabria	cordex.org

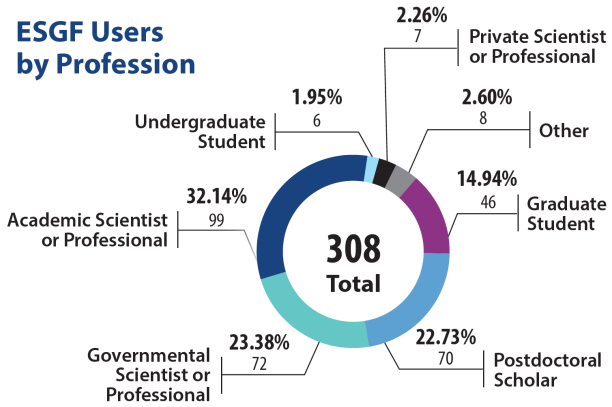
Fig. 3. National and international scientific gateways for the dissemination and secure access of geophysical data and resources of the ESGF peer-to-peer enterprise system. This ESGF collaboration develops, deploys, and maintains software infrastructure for the management, dissemination, and analysis of model output and observational data. As illustrated by the breadth of projects and geographical reach, ESGF supports a broad range of activities that require ongoing access to a full spectrum of Earth system science data.

User Survey Results

Quantity of ESGF Data Users Consume for Work



ESGF Users by Profession



The ESGF system is built specifically to support scientific research, comprising peer nodes distributed across several countries and united by common protocols and interfaces (see **Fig. 3**, previous page). This 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 compute environment, and other community standards.

Users also can easily leverage ESGF's component architecture to access data from other scientific domains, such as satellite, instrument, and other forms of observational data. Many Earth science projects have adopted the full ESGF infrastructure. They include the Coupled Model Intercomparison Project (CMIP), whose output is to be used in upcoming IPCC ARs; MIPs endorsed by the World Climate Research Programme; the Accelerated Climate Modeling for Energy (ACME) project; and the Centre of Excellence in Simulation of Weather and Climate in Europe (ESiWACE) project, which uses ESGF in its overarching workflow process to store and to analyze model output.

As projections of global geophysical disturbances have become critical to informing decision making, effectively managing vast volumes of data presents a major challenge for computational and data scientists who support those projections. ESGF fills this vital worldwide infrastructural role. It also delivers data for a wide variety of purposes and projects, including both model simulation and observational data, and delivers ultrascale capabilities for cataloging, accessing, and analyzing large datasets.

On the ESGF Horizon

ESGF functionality (See **Fig. 4**, this page) will enable scientific discovery:

- **Dynamic resources.** Manage and enhance user accessibility.
- ***In situ* analysis.** Enable users to obtain real-time computational results by performing calculations at the site where data reside.
- **Machine learning.** Find patterns and features in exabyte-scale data, providing users with new insights based on the vast wealth of available data.
- **Uncertainty quantification.** Connect uncertainty in underlying data with the reliability of conclusions.
- **Cloud development.** Enable easy setup of ESGF nodes and services on the cloud, either for permanent management or for transitional computational tasks ("cloud bursting").

ESGF Roadmap

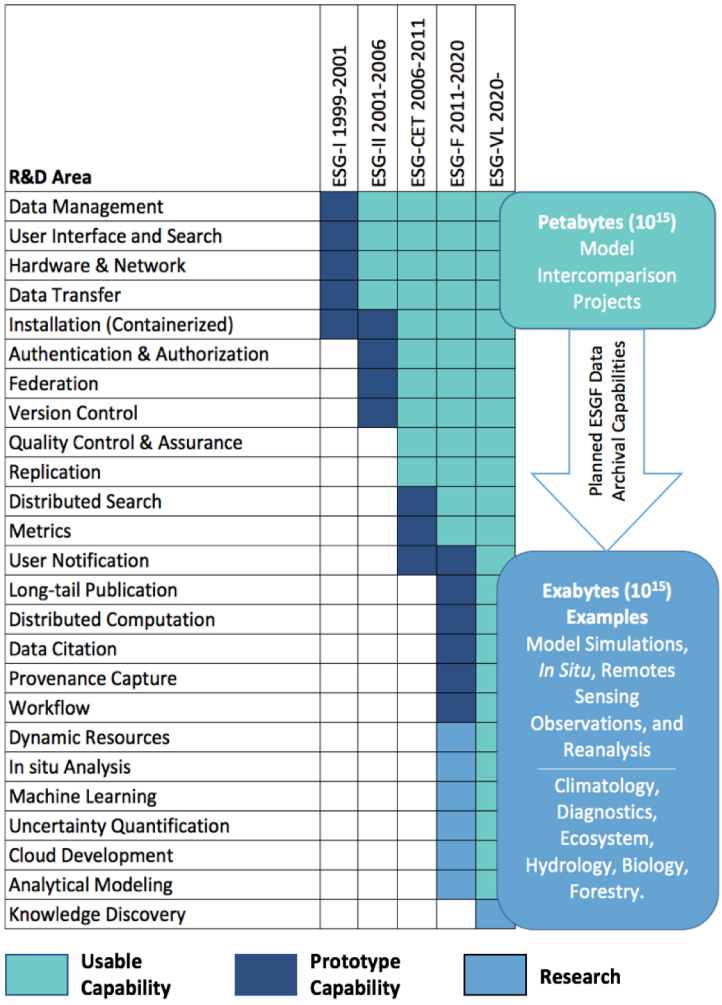


Fig. 4. Needed ESGF data system evolution. ESGF development rests on several specific research and development (R&D) areas, listed at left. These efforts will strengthen several existing capabilities, bring many prototyped capabilities into full community use, and introduce several R&D areas that set the stage for knowledge discovery.

- **Analytical modeling.** Gauge how long calculations will take on various platforms, so that users can optimize their use of worldwide computational resources.

ESGF Contacts and Websites

DOE BER Program Manager

Justin Hnilo
justin.hnilo@science.doe.gov
301-903-1399

ESGF PI and Chair

Dean N. Williams
DOE Lawrence Livermore National Laboratory (LLNL)
williams13@llnl.gov
925-423-0145

Websites

Earth System Grid Federation
esgf.llnl.gov

- **ESGF Committee Members**
  esgf.llnl.gov/committee.html
- **Reports**
  esgf.llnl.gov/reports.html
- **Conference**
  esgf.llnl.gov/conferences.html
- **Letters of support**
  esgf.llnl.gov/letters-of-support.html

LLNL ESGF Node
esgf-node.llnl.gov

Biological and Environmental Research Community Resources
science.energy.gov/ber/community-resources

Portal to CMIP3, CMIP5, input4MIP, and ACME datasets, ESG Center for Enabling Technologies
esgf-node.llnl.gov

June 2017