International Consortium Developing the Next Generation Earth System Grid Federation (ESGF) Distributed Data Infrastructure

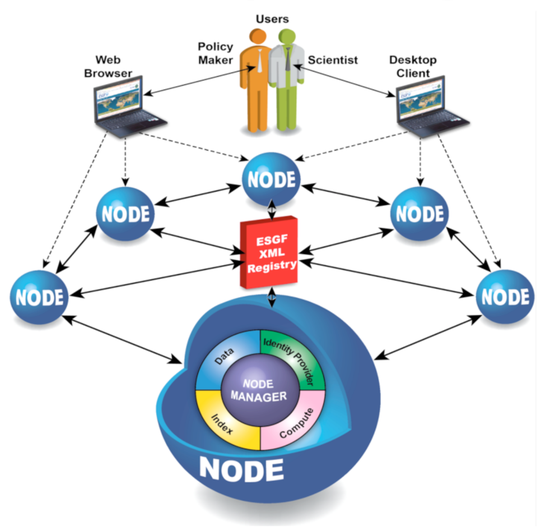
Forrest M. Hoffman[1], Philip Kershaw[2], Sasha Ames[3], Rachana Ananthakrishnan[4], Laura Carriere[5], Ben Evans[6], Stephan Kindermann[7], Christian Pagé[8], and Aparna Radhakrishnan[9]

[1]Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA; [2]Centre for Environmental Data Analysis, Harwell Oxford, Didcot, United Kingdom; [3]Lawrence Livermore National Laboratory, Livermore, California, USA; [4]University of Chicago, Chicago, Illinois, USA; [5]National Aeronautics and Space Administration, Greenbelt, Maryland, USA; [6]Australian National University National Computational Infrastructure, Canberra, Australia; [7]German Climate Computing Centre, Hamburg, Germany; [8]Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique, Toulouse, France; and [9]National Oceanic and Atmospheric Administration Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey, USA

Introduction to the Earth System Grid (ESGF)

- The **Earth System Grid Federation (ESGF)** is an *international consortium* and a *globally distributed peer-to-peer network of data servers* using a common set of protocols and interfaces to archive and distribute climate and Earth system model output and related input, observational, and reanalysis data to the research community.
- These **Open Science data** were produced by modeling centers participating in the **World Climate Research Programme's (WCRP's) Coupled Model Intercomparison Projects (CMIP)**.
- The data are used by scientists all over the world to **investigate consequences of possible climate change scenarios and the Earth system feedbacks** that could occur as a result of continued or increasing anthropogenic emissions.
- Many of these studies form the basis for the *Assessment Reports* produced by the **United Nations Intergovernmental Panel on Climate Change (IPCC)**, including the IPCC Sixth Assessment Report from Working Group I, which as released 9 August 2021.

ESGF Conceptual Diagram



37	189	26	15,842,865	272,040,492	47.95
Projects	Countries	Data nodes	published datasets	files downloaded	PB downloaded

The distributed ESGF network, consisting of 26 data nodes that host over 15 million datasets from 37 data projects, has delivered more than 272 million files totalling 47.95 petabytes to 189 countries since January 2018.

*Primary international contributors to the development of ESGF, represented by the logos above, include the **Department of Energy (DOE)**, the **National Aeronautics and Space Administration (NASA)**, the **National Oceanographic and Atmospheric Administration (NOAA)**, and the **National Science Foundation (NSF)** in the United States; the **Infrastructure for the European Network for Earth System modelling (IS-ENES)** program in Europe; and the **National Computational Infrastructure (NCI)** in Australia.*

Coupled Model Intercomparison Project (CMIP) Output

- The **Coupled Model Intercomparison Project (CMIP)** begin in 1995 under the auspices of the **Working Group on Coupled Modelling (WGCM)**.
- The objective of CMIP is to **better understand past, present and future climate changes** arising from natural, unforced variability or in response to changes in radiative forcing in a multi-model context.
- This understanding includes assessments of model performance during the historical period and quantifications of the causes of the spread in future projections.
- In addition to the major simulation experiments, CMIP has also included a series of smaller model intercomparison efforts, called the **Coordinated CMIP Experiments**, designed to understand specific aspects of model responses.
- Under the guidance and at the direction of the WGCM, all CMIP activities are overseen by a coordinated pair of subcommittees:
  - **CMIP Panel** – works with those organizing various focused model intercomparisons to integrate them with the set of standard CMIP experiments to forge a synergistic experiment design for each new phase of CMIP
  - **WGCM Infrastructure Panel (WIP)** – promotes coordinated development of infrastructure needed to support CMIP, most notably the archiving and serving of CMIP data
- In March 2022, the **CMIP International Project Office (CMIP-IPO)** opened alongside European Space Agency's Climate Office at its European Centre for Space Applications and Telecommunications (ECSAT) facility in Oxfordshire, United Kingdom. Contact the IPO team at cmip-ipo@esa.int.
- The CMIP-IPO coordinates the project under the governance of the **WCRP Working Group on Coupled Modelling (WGCM)** and is part of the developing **Earth System Modelling and Observations (ESMO) Core Project**, which coordinates all modelling, data and observations activities across WCRP and their key partners.

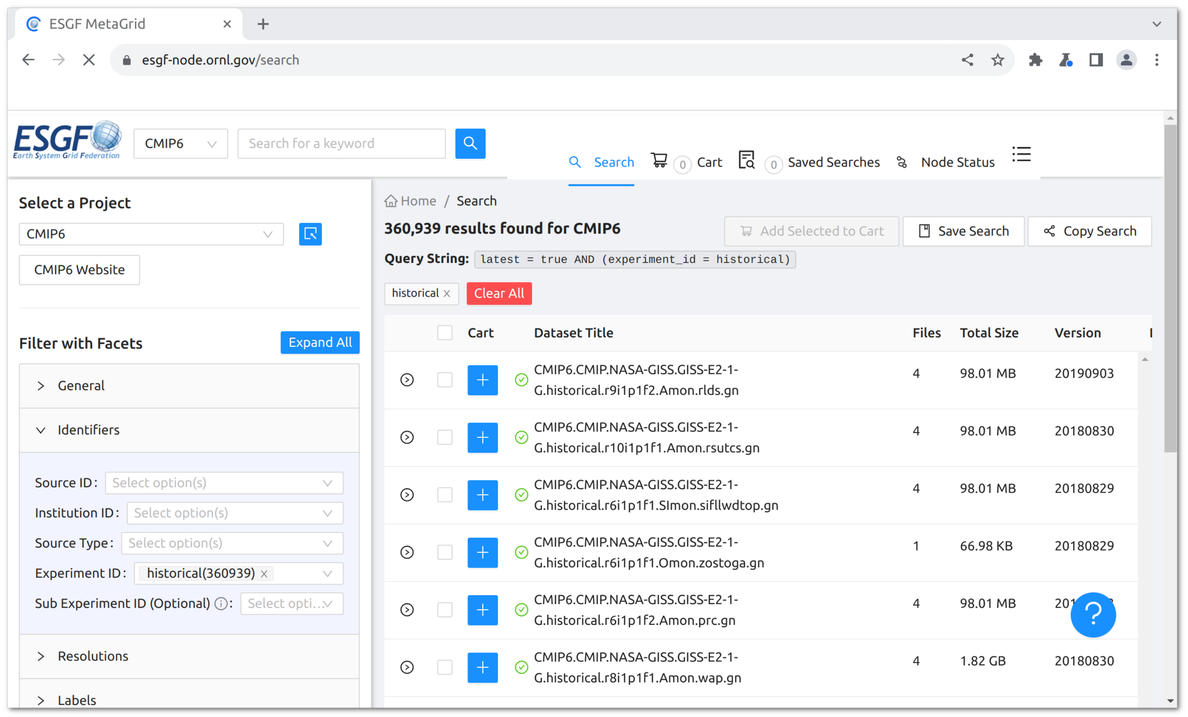
*Contributions to CMIP6 came from widely distributed modeling centers located in more than 15 countries. These modeling centers published the output from their CMIP6 simulation experiments to ESGF, following a carefully designed controlled vocabulary developed by the **Program for Climate Model Diagnosis & Intercomparison (PCMDI)** at Lawrence Livermore National Laboratory with support from the US Department of Energy.*

CMIP6	14,785,119 total datasets 27,446.61 TB	7,590,309 distinct datasets 15,953.41 TB	7,194,810 replica datasets 11,493.2 TB
CORDEX	187,785 total datasets 1,473.33 TB	187,513 distinct datasets 1,472.77 TB	272 replica datasets 0.56 TB
CMIP5	201,130 total datasets 5,293.61 TB	52,163 distinct datasets 1,525.07 TB	148,967 replica datasets 3,768.55 TB
INPUT4MIPS	5,871 total datasets 10.84 TB	21 distinct datasets 0.9 TB	5,850 replica datasets 9.95 TB
OBS4MIPS	126 total datasets 0.2 TB	108 distinct datasets 0.2 TB	18 replica datasets 0.01 TB

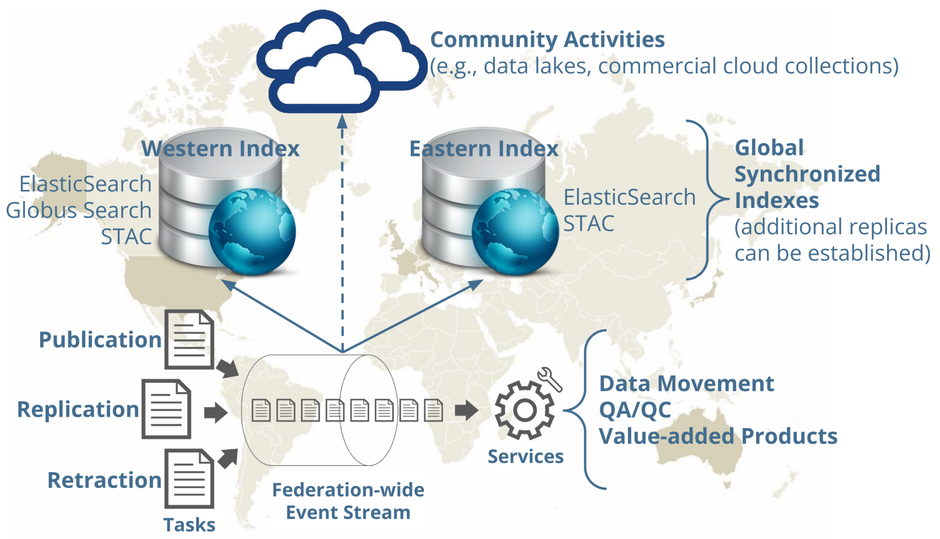
An important goal of CMIP is to make the multi-model output publicly available in a standardized format. In coordination with CMIP and the WIP, ESGF hosts and distributes model output, related forcing, reanalysis, downscaled, and observational data. CMIP6 output constitutes the largest proportion of data holdings by volume in CMIP6, exceeding 15 out of 22 PB of unique datasets. Including replica copies of datasets, total global ESGF holdings exceed 37 PB. (Updated 8 December 2024)

Modernizing ESGF Infrastructure

- The international ESGF consortium is **modernizing the data system architecture and building tools and platforms for the research community** in preparation for the follow-on CMIP6+ activity and for CMIP7.
- Employing newer computing technologies, we are designing, developing, and deploying new capabilities, including
  - **container-based data and index node software** based on Docker and Kubernetes to simplify maintenance and operation of ESGF nodes;
  - **institutional-based authentication** to eliminate the need to maintain separate credentials only for ESGF;
  - **index nodes operating in the cloud** to reduce the total number of index nodes and improve the scalability and performance of data searches;
  - **managed automation of data publishing** to ease ingest of model output from modeling centers;
  - **data access interfaces and data discovery tools** to simplify finding and accessing desired data;
  - **server-side computing capabilities** to provide subsets and to compute summaries and value-added products;
  - **data transfer tools and protocols**, like Globus and S3-compatible interfaces, to enable high-speed, unattended data transfers and interactive on-demand streaming of data only as needed during analysis; and
  - **user computing platforms** based on Kubernetes and JupyterHub to offer analysis capabilities where the data are stored and avoid data transfers.
- A new, redesigned **web-based data search user interface, called Metagrid**, is available for testing. It
  - Was developed on the popular React / JavaScript framework;
  - Offers new features, including a shopping cart, ability to save & share searches, and a page tour & support dialog;
  - Provides user experience enhancements that make it faster to discover published data;
  - Will soon provide Globus integration for authentication and transfer to offer faster and reliable unattended data movement; and
  - Will be deployed across ESGF nodes by the end of 2024.



*The new **Metagrid** user interface offers a redesigned faceted search capability with a variety of new features.*



A new catalog indexing strategy would provide two global synchronized indexes, one in the USA and one in Europe, that would provide comprehensive federated search across all data nodes. A Federation-wide Event Stream would be used to ensure synchronization of the two indexes and to initiate data movement and production of value-added data products.

ESGF Outreach Activities

To support the community, ESGF will organize or host

- **Webinars, tutorials, and bootcamps** for
  - Sharing data management lessons learned,
  - Instructing modeling centers on best practices, and
  - Fostering more efficient data discovery & access;
- **Hackathons and workshops** on
  - Data standards and controlled vocabularies,
  - Data node and user computing deployment,
  - Developing machine learning tools using Earth system model data;
- **ESGF Developer and User Conferences** to
  - Foster collaborative development of software tools and interface standards,
  - Provide a venue for early career researchers to learn data analysis approaches, and
  - Enhance integration with other data centers and Earth science data resources.

ESGF Webinar series recordings are currently available at https://www.youtube.com/@esgf2432

Acknowledgments

ESGF is supported by climate and Earth system model data infrastructure activities in multiple nations. Contributions to ESGF mondernization development in the United States is supported by the Earth System Grid Federation 2-US (ESGF2-US) Project, which is sponsored by the Data Management Program in the Earth and Environmental Systems Sciences Division (EESSD) of the Office of Biological and Environmental Research (BER) in the US Department of Energy Office of Science. This research uses resources of the Oak Ridge Leadership Computing Facility (OLCF) at Oak Ridge National Laboratory (ORNL), which is managed by UT-Battelle, LLC, for the US Department of Energy under Contract No. DE-AC05-00OR22725. We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP. For CMIP the US Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals.