Data Management Linkages with RGMA

Justin Hnilo
Program Manager, Data Management

October 14, 2020

Virtual RGMA Meeting

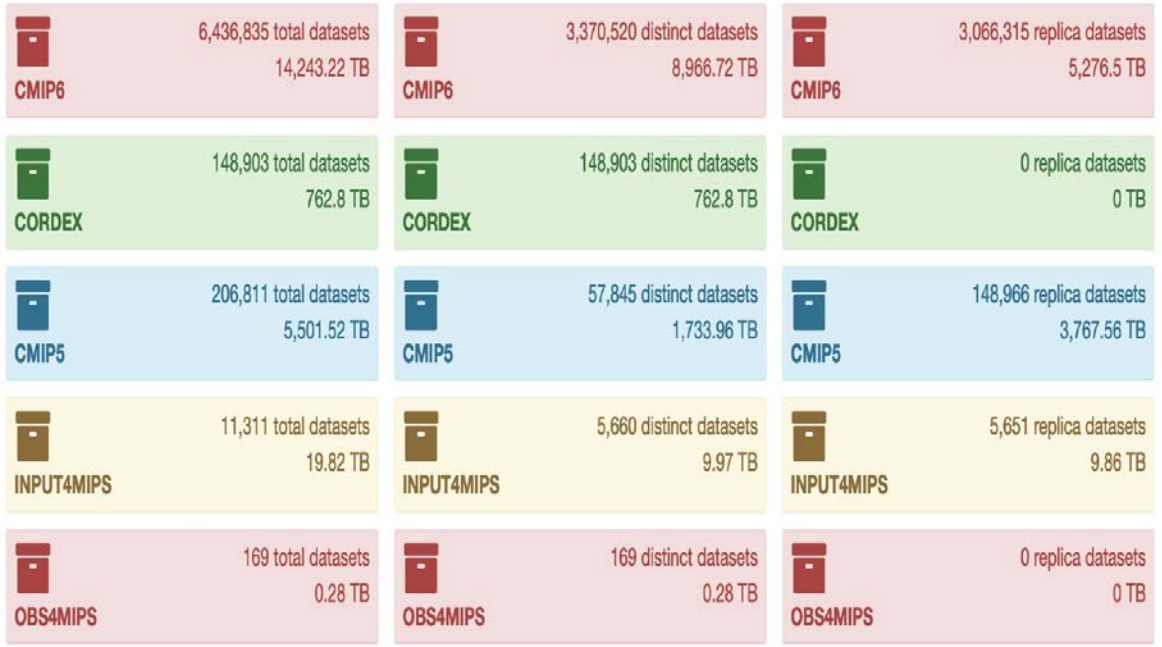
The Earth System Grid Federation (ESGF)

Program Description: The Earth System Grid Federation (ESGF) is an international collaboration for the software that powers most global climate change research, notably assessments by the Intergovernmental Panel on Climate Change (IPCC).

- ESGF manages the first-ever decentralized database for handling climate science data, with multiple petabytes of data at dozens of federated sites worldwide. It is recognized as the leading infrastructure for the management and access of large distributed data volumes for climate change research. It supports the Coupled Model Intercomparison Project (CMIP), whose protocols enable the periodic assessments carried out by the IPCC.
- Using a system of geographically distributed peer nodes—independently administered yet united by common protocols and interfaces—the ESGF community holds the premier collection of simulations and observational and reanalysis data for climate change research.
- Scientific users: >2500/year
- Scientific publications: ~ 750/year
- Location: Nodes exist both domestically and internationally, there are in excess of 70 nodes today.

ESGF uses a federated “data cloud” to manage and distribute Earth and Environmental Science Division model and other observational data to the climate community

- Current data archives managed by ESGF:
  - The IPCC CMIP3 and CMIP5 data archives used in the IPCC AR5, and the CMIP6 data archive being used in the upcoming IPCC AR6;
  - Multi-model data sets managed and distributed by the BER-funded Program for Climate Model Diagnosis and Intercomparison (PCMDI);
  - The DOE E3SM project, which currently distributes 330 TB (and growing rapidly) of model output publicly to researchers;
  - The input4MIPs and obs4MIPs data sets managed and distributed by PCMDI.

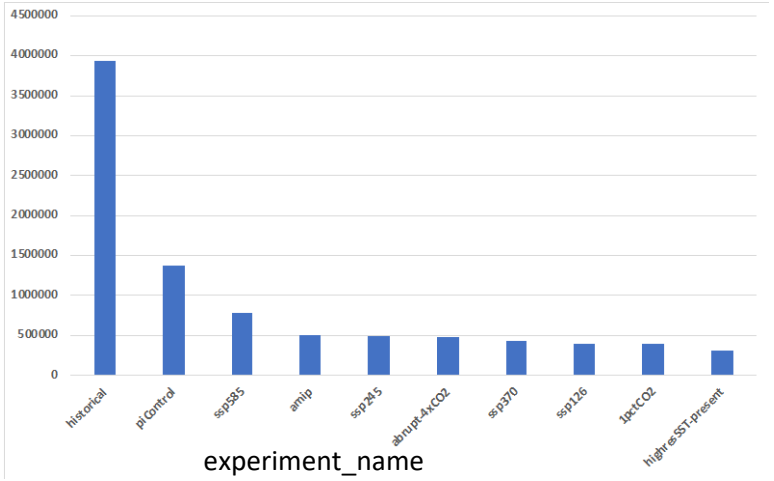


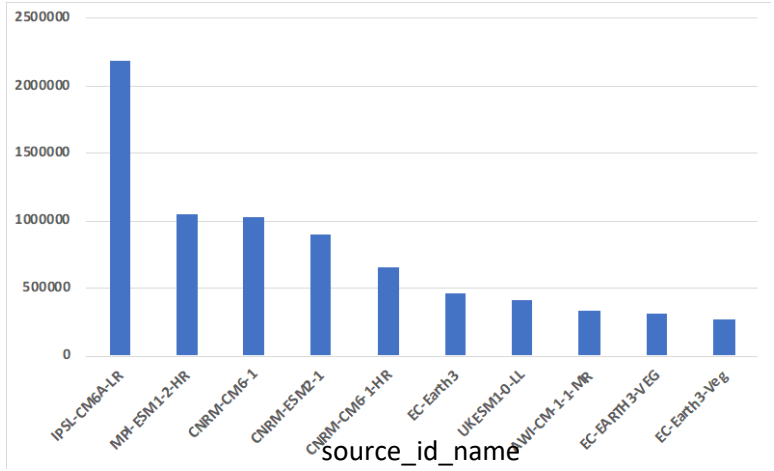
Current total datasets and their sizes for the top ESGF-hosted international projects. Over the past three years, the ESGF data collection has more than doubled; CMIP6 is the top project in terms of number of total data sets and data sizes.

Between September 2019 and September 2020 users downloaded 135,637,177 files with a total size of 9.5 PB

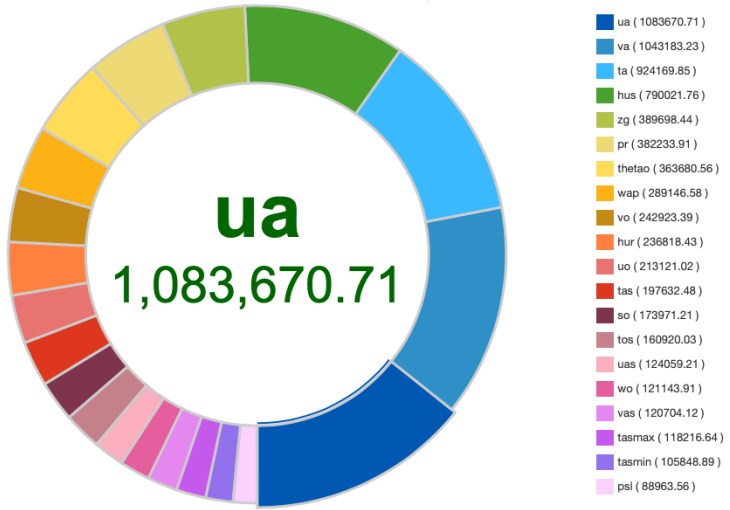
Data volumes for the top CMIP downloaded Experiments, sources, and variables

- Data volumes in GB for the top CMIP downloaded Experiments, sources, and variables.
  - These insights can help serve the community better
  - Replication , data caching, and, data subsetting algorithms can utilize this information

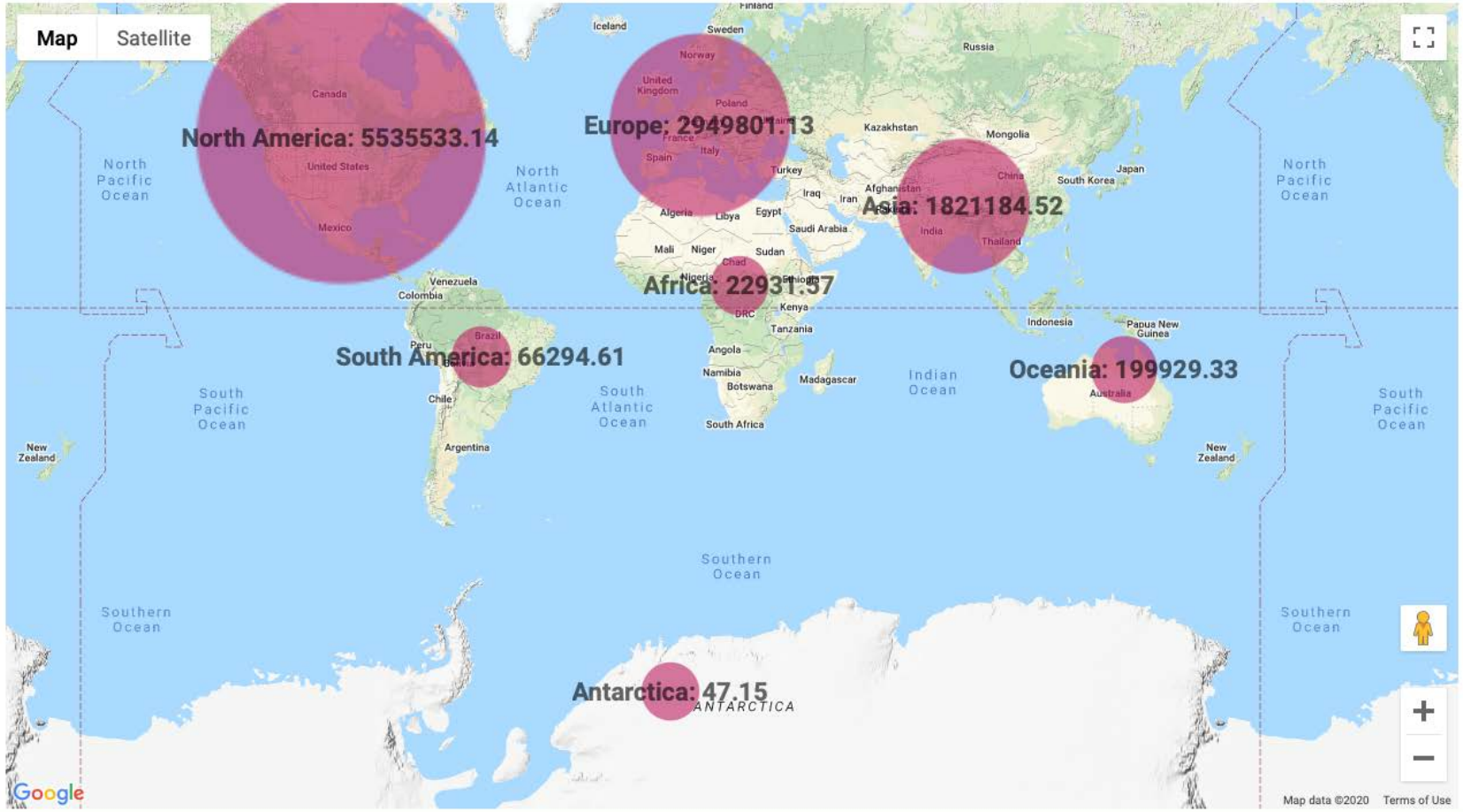




Top CMIP downloaded variables



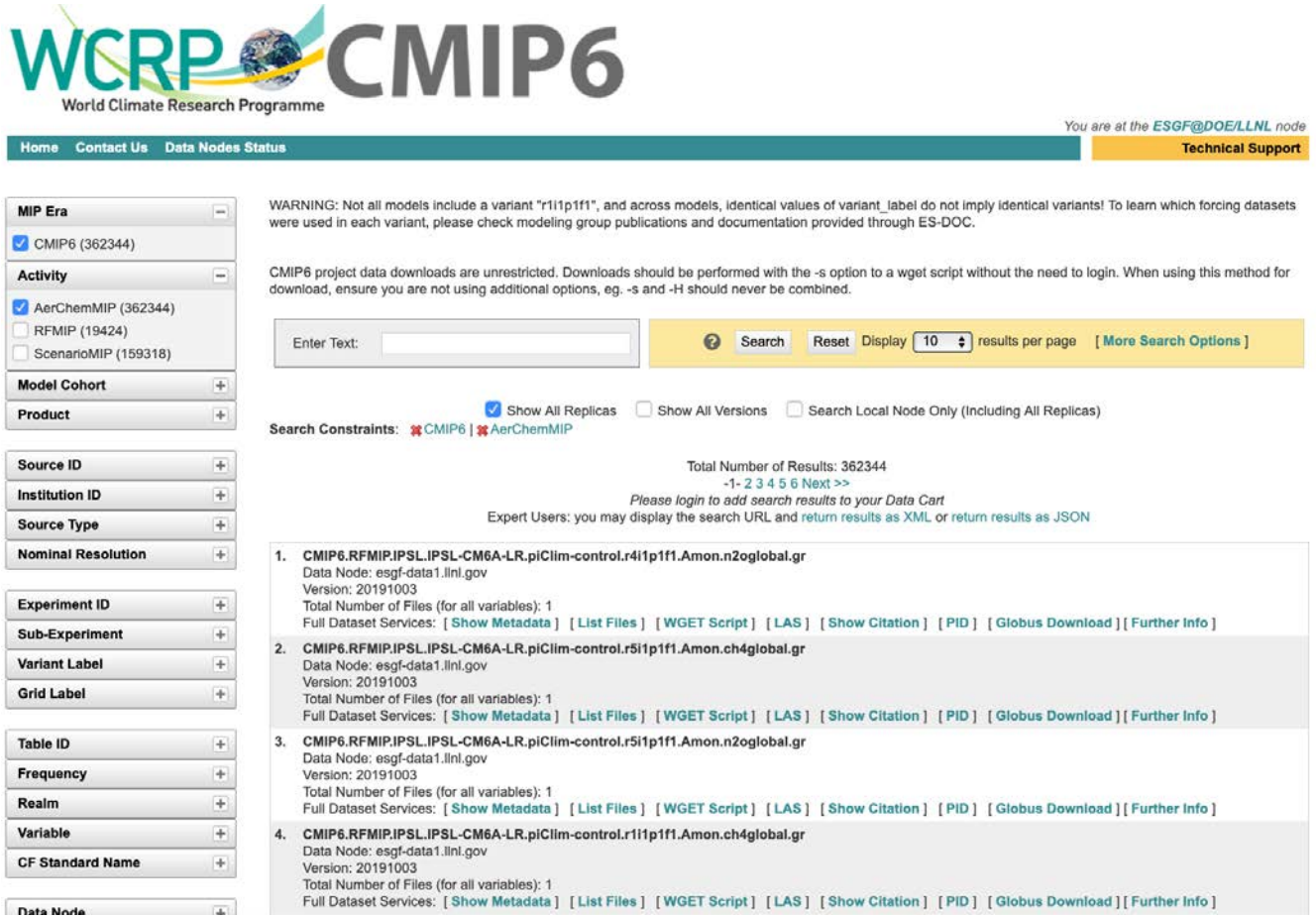
CMIP6 downloaded data volume by Continent



- Very active community across the world
- China is responsible for the majority of activity in Asia

ESGF CMIP6 search portal

- The screenshot shows results from a simple search using the faceted search
- Users can alternatively perform text search by entering the terms in the text box



ESS-DIVE: an Accessible Archive of Environmental Data

A new digital archive enables community use of terrestrial and subsurface ecosystem data sets

Challenge

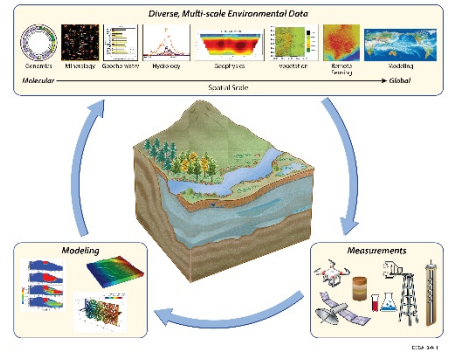
- Earth scientists need access to long-term, spatially dense, high-quality observational data sets coupled with simulations to understand and predict ecosystem behavior over timescales spanning decades to centuries.

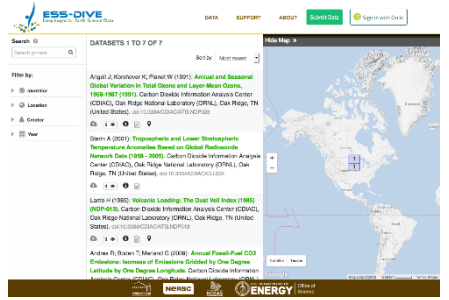
Approach and Results

- Data services launched on 1 April 2018.
- Established to provide long-term stewardship and enable broad usage of data from research in the DOE's Environmental System Science (ESS) domain.
- Proactively engaging with the ESS scientific research community to understand their needs and to adopt or develop standards.
- Designed using Findable, Accessible, Interoperable, and Reusable (FAIR) principles.

Significance and Impact

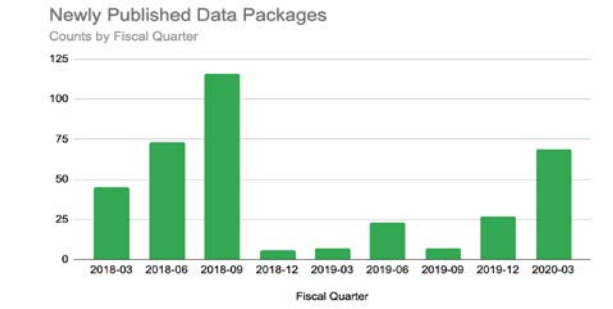
- The ESS-DIVE data portal allows members of the public to search published data and download files (249 datasets available).
- Engaging with the ESS Community to define and develop features needed.
- Includes all relevant data from previous ESS archive.

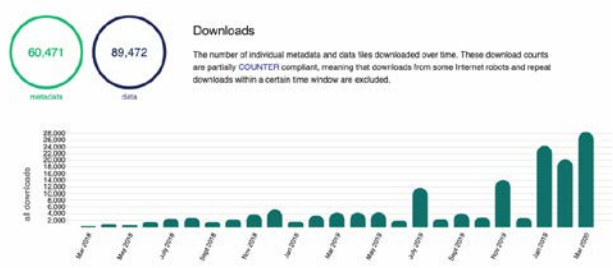


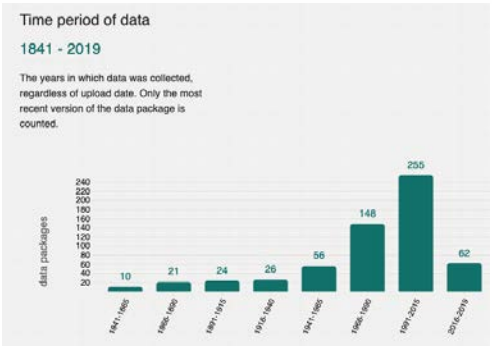


Reference: Varadharajan, C., *et al*., (2019), Launching an accessible archive of environmental data, *Eos, 100,* https://doi.org/10.1029/2019EO111263. Published on 08 January 2019.

ESS-DIVE - ESS Public Data Reaching A Broad Audience







- Data available at https://data.ess-dive.lbl.gov
- ESS-DIVE: Environmental Systems Science Data Infrastructure for a Virtual Ecosystem; DOI: https://doi.org/10.25504/FAIRsharing.d6Pe1f;
- Contact: ess-dive-support@lbl.gov

Technical Advancement

- Web and programmatic storage of data packages in ESS-DIVE deployed along with semi-automated curation
- ESS-DIVE data searchable via a broad range of cross-agency data search engines

Significance and Impact

- Over 28,000 downloads of data and metadata in March 2020
- Total of 373 data packages published in ESS-DIVE since April 2018 with 76 published in FY20 Q2
- Standardized metadata enables cross data package search
- Data from ESS projects publicly available from a single location

Research Details

- Data package upload and edit provide web services for data package storage and updates (in use by large-scale projects)
- Automated data quality reports available on web site
- Funding 6 ESS community groups to develop standards for ESS data and file types
- Incorporating best-practices (e.g. FAIR data) from digital libraries community. Attending meetings to provide input on standards
- Built on Metacat and Metacat UI developed by NCEAS/DataONE

Environmental Data are Highly Diverse Consisting of Multi-Disciplinary, Multi-scale, Multi-resolution, Heterogeneous Formats

ESS Community Cyberinfrastructure The Challenges ESS-DIVE Must face.