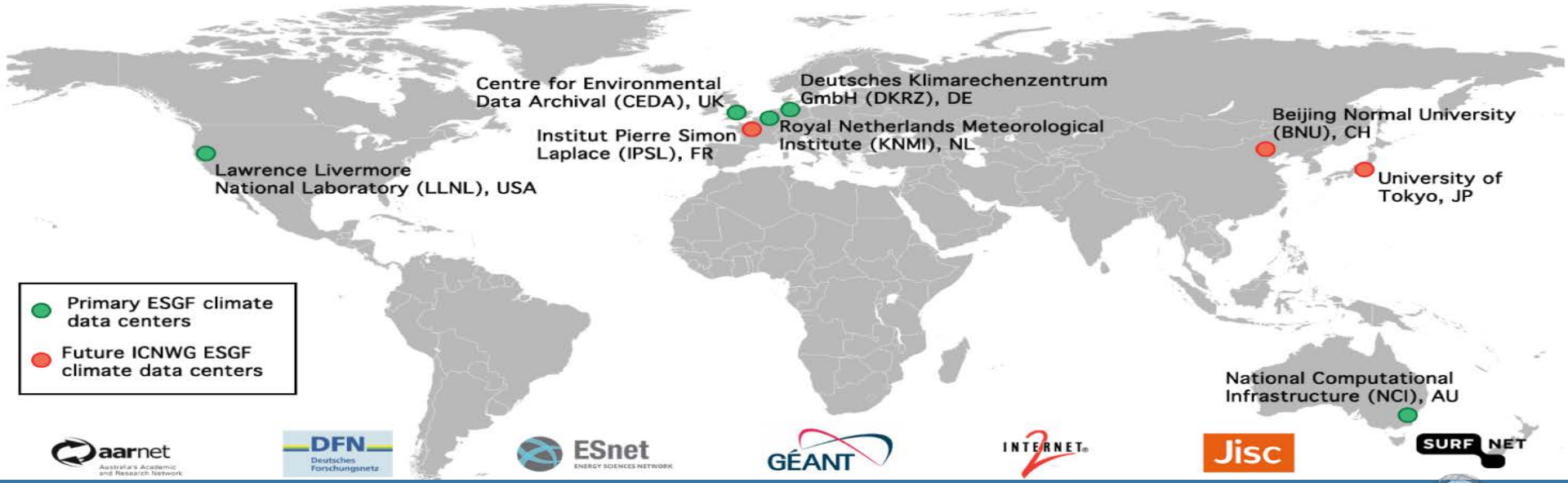
The Earth System Grid Federation (ESGF)

http://esgf.llnl.gov

STREAM 2016: Streaming Requirements, Experience, Applications and Middleware Workshop

March 22, 2016

Dean N. Williams (ESGF Executive Committee Chair)
On behalf of the ESGF Executive Committee and Development Teams

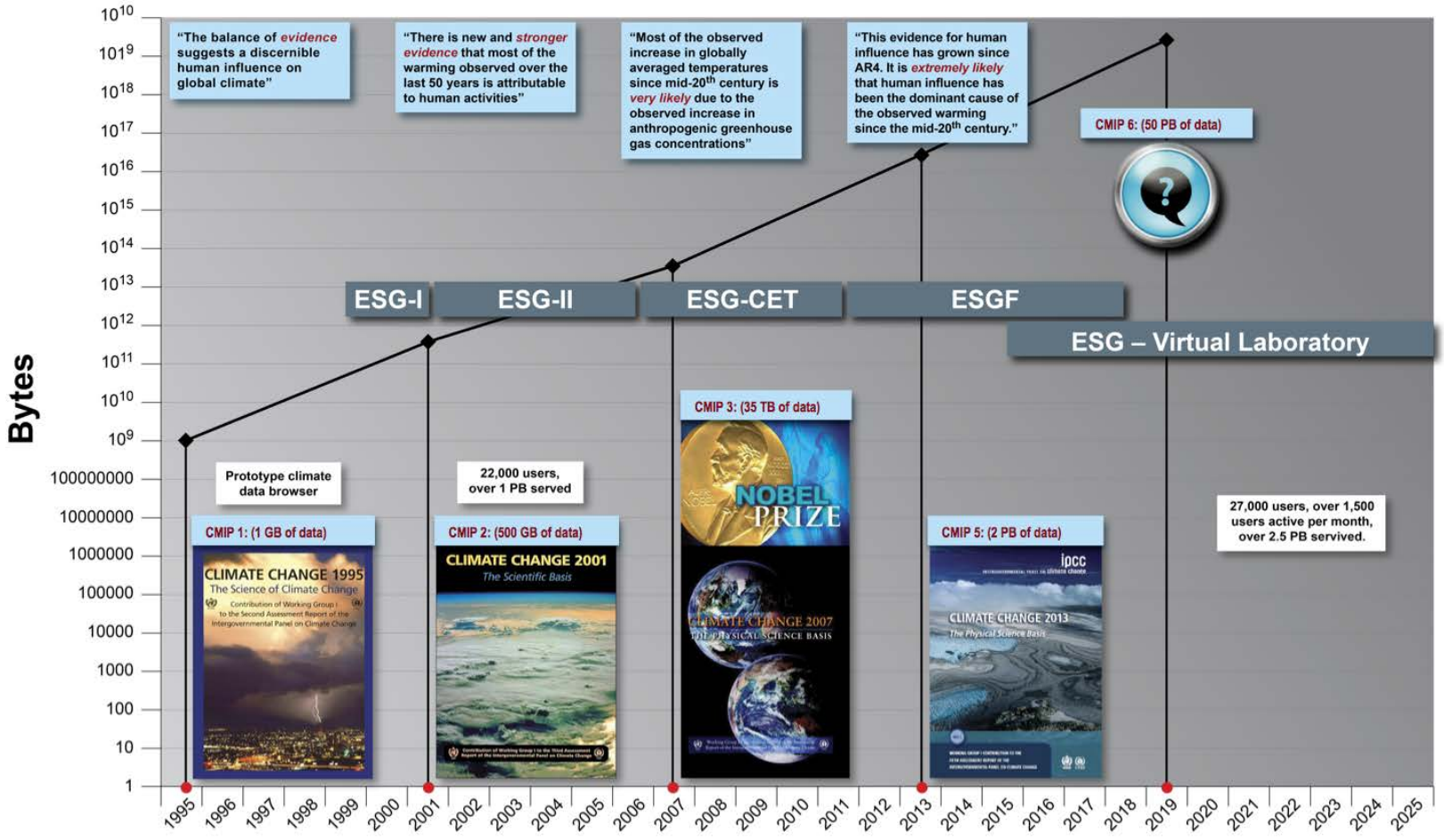


ESGF Earth System Grid Federation

ESGF[1] is a coordinated multiagency, international collaboration of institutions that continually develop, deploy, and maintain software needed to facilitate and empower the study of climate

1. Dean N. Williams, V. Balaji, Luca Cinquini, Sébastien Denvil, Daniel Duffy, Ben Evans, Robert Ferraro, Rose Hansen, Michael Lautenschlager, and Claire Trenham, "A Global Repository for Planet-Sized Experiments and Observations", Bulletin of the American Meteorological Society, early release, 2016, doi: http://dx.doi.org/10.1175/BAMS-D-15-00132.1.

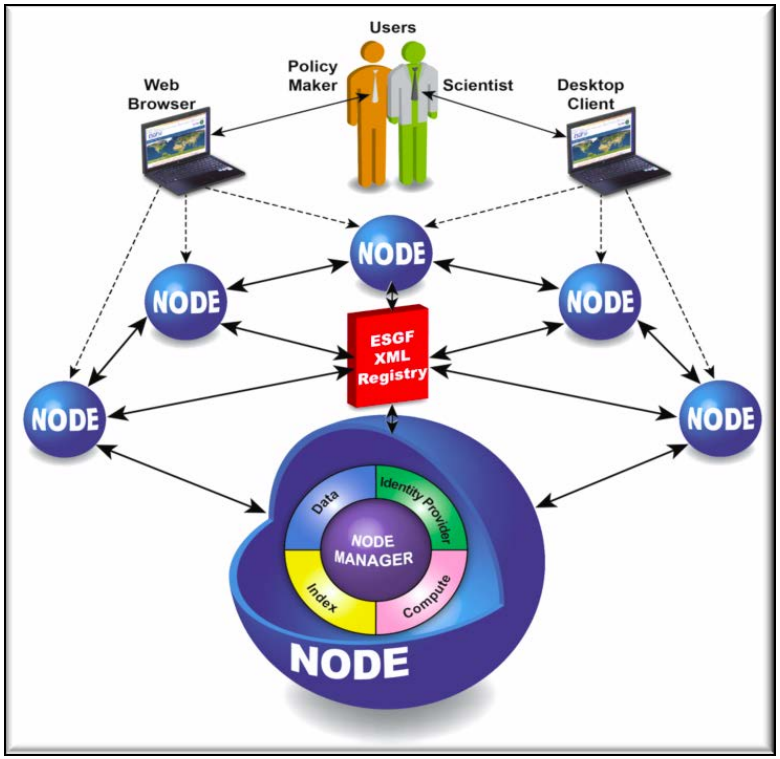
CMIP and ESGF history: scientific challenges and motivation use case

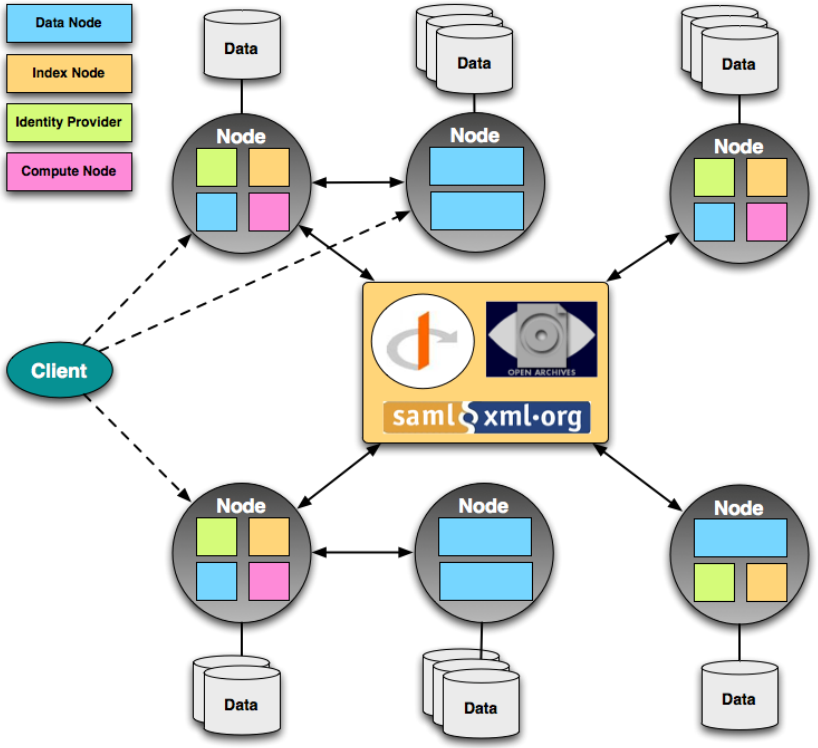


ESG recognizes that data management, stewardship and curation is an ongoing and long-lived function that requires a strategy that is resilient to continuing evolution in hardware and software.

ESGF software infrastructure

ESGF is a software infrastructure for management, dissemination, and analysis of simulation and observational data. The software utilizes hardware, networks, software for data management, access and processing.





ESGF federation nodes interact as equals. Users log onto any node using single sign-on OpenID to obtain and access data throughout the entire federation.

ESGF release version 2.0 (overhauled)

Following a security incident in June 2015, the ESGF system was brought offline and the software stack was extensively re-engineered to accomplish the following goals:

- Execute complete software scan of all modules, fix all exposed and other potential security breaches
- Major upgrade of underlying system infrastructure:
  - CentOS7, Java 1.8, Tomcat 8, Postgres 9.4, OpenSSL 1.0, Python 2.7.9
  - Switch ESGF installer to RPM-based components
  - Run Apache httpd server in front of Tomcat (better performance, flexibility)
- Major upgrade of all ESGF services:
  - Search services (Solr5), data download (TDS5), high performance data transfer (Globus-Connect-Server), computation (UV-CDAT), visualization (LAS)
  - Replace old web-front-end with new CoG user interface
- Republish ALL data collections (CMIP5, CORDEX, Obs4MIPs, ana4MIPs,…)

ESGF sub-tasks and task leaders

Sub-Task	Task Leads	Description
1. CoG User Interface Working Team	**Cecelia DeLuca (NOAA) and Luca Cinquini (NOAA)**	**Improved ESGF search and data cart management and interface**
2. Compute Working Team	**Charles Doutriaux (DOE) and Daniel Duffy (NASA)**	**Developing the capability to enable data analytics within ESGF**
3. Dashboard Working Team	**Sandro Fiore (IS-ENES)**	**Statistics related to ESGF user metrics**
4. Data Transfer Working Team	Lukasz Lacinski (DOE) and Rachana Ananthakrishnan	ESGF data transfer and enhancement of the web-based download
5. Documentation Working Team	Matthew Harris (DOE) and Sam Fries (DOE)	Document the use of the ESGF software stack
6. Identity Entitlement Access	Philp Kershaw (IS-ENES) and Rachana Ananthakrishnan (DOE)	ESGF X.509 certificate-based authentication and improved interface
7. Installation Working Team	**Nicolas Carenton and Prashanth Dwarakanath (IS-ENES)**	**Installation of the components of the ESGF software stack**
8. International Climate Network Working Group	**Eli Dart (DOE/ESnet) and Mary Hester (DOE/ESnet)**	**Increase data transfer rates between the ESGF climate data centers**
9. Metadata and Search Working Team	Luca Cinquini (NASA)	ESGF search engine based on Solr5; discoverable search metadata
10. Node Manager Working Team	Sasha Ames (DOE) and Prashanth Dwarakanath (IS-ENES)	Management of ESGF nodes and node communications
11. Provenance Capture Working Team	Bibi Raju (DOE)	ESGF provenance capture for reproducibility and repeatability
12. Publication Working Team	Sasha Ames (DOE) and Rachana Ananthakrishnan	Capability to publish data sets for CMIP and other projects to ESGF
13. Quality Control Working Team	**Martina Stockhause (IS-ENES) and Katharina Berger (IS-ENES)**	**Integration of external information into the ESGF portal**
14. Replication Working Team	Stephan Kindermann (IS-ENES) and Tobias Weigel (IS-ENES)	Replication tool for moving data from one ESGF center to another
15. Software Security Working Team	Prashanth Dwarakanath (IS-ENES) and Laura Carriere (NASA)	Security scans to identify vulnerabilities in the ESFF software
16. Tracking / Feedback Notification Working Team	Sasha Ames (DOE)	User and node notification of changed data in the ESGF ecosystem
17. User Support Working Team	Torsten Rathmann (IS-ENES) and Matthew Harris (DOE)	User frequently asked questions regarding ESGF and housed data
18. Versioning Working Team	Stephan Kindermann (IS-ENES) and Tobias Weigel (IS-ENES)	Versioning history of the ESGF published data sets

Further elaborations of the sub-tasks are described in the ESGF progress reports, which can be found online: http://esgf.llnl.gov/reports.html

Compute Working Team

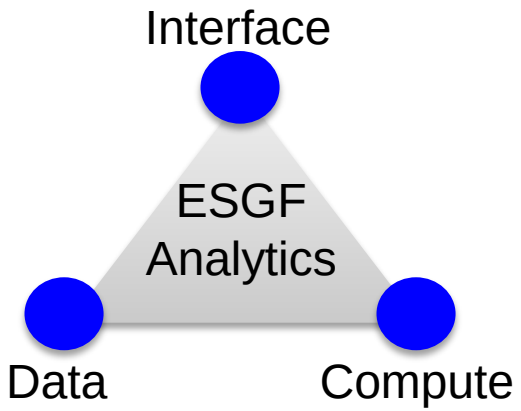
ESGF - CWT Team Leads:
Daniel Duffy (NASA/GSFC)
Charles Doutriaux (DOE/LLNL)

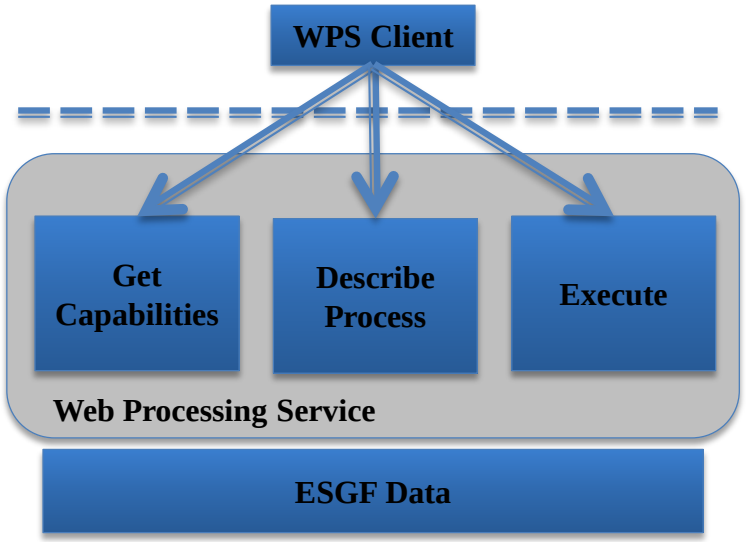
Enabling data proximal analytics

- Create ESGF Analytics capabilities by exposing compute resources through well defined interfaces
- Analytics that may require high performance computing resources (compute and memory)
- Allow ESGF users to download the outputs of analysis rather than huge data sets

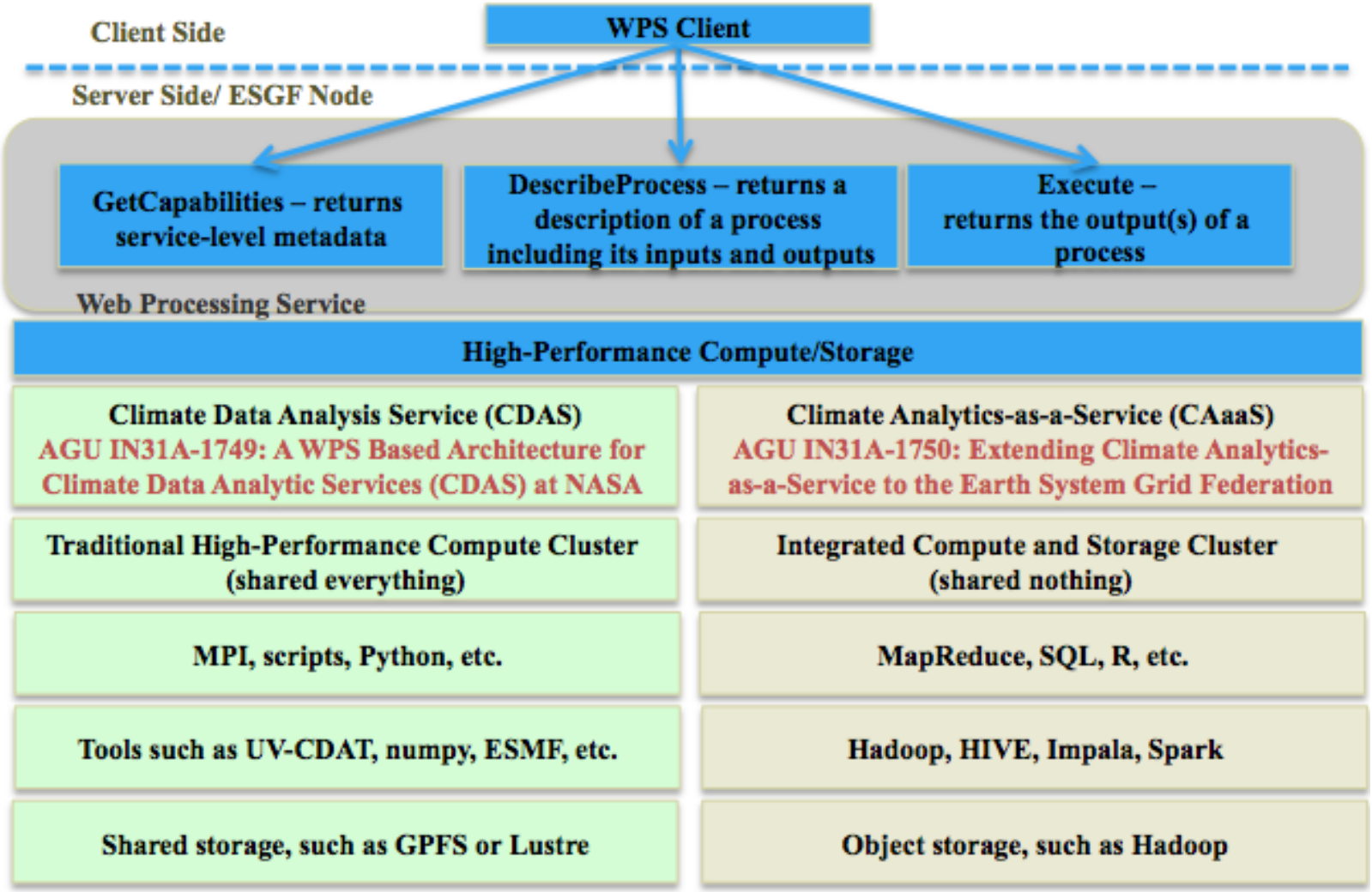
Web Processing Service Application Programming Interface (WPS API)

- Suite of analysis applications that can be executed through an API
- API fits multiple backend implementations
- Relatively simple analysis, such as averages, maximums, minimums, etc.
- More complex routines, such as regridding, anomalies, trends, etc.





Two reference back-end implementations



ON-DEMAND STREAMING

of massive climate simulation ensembles

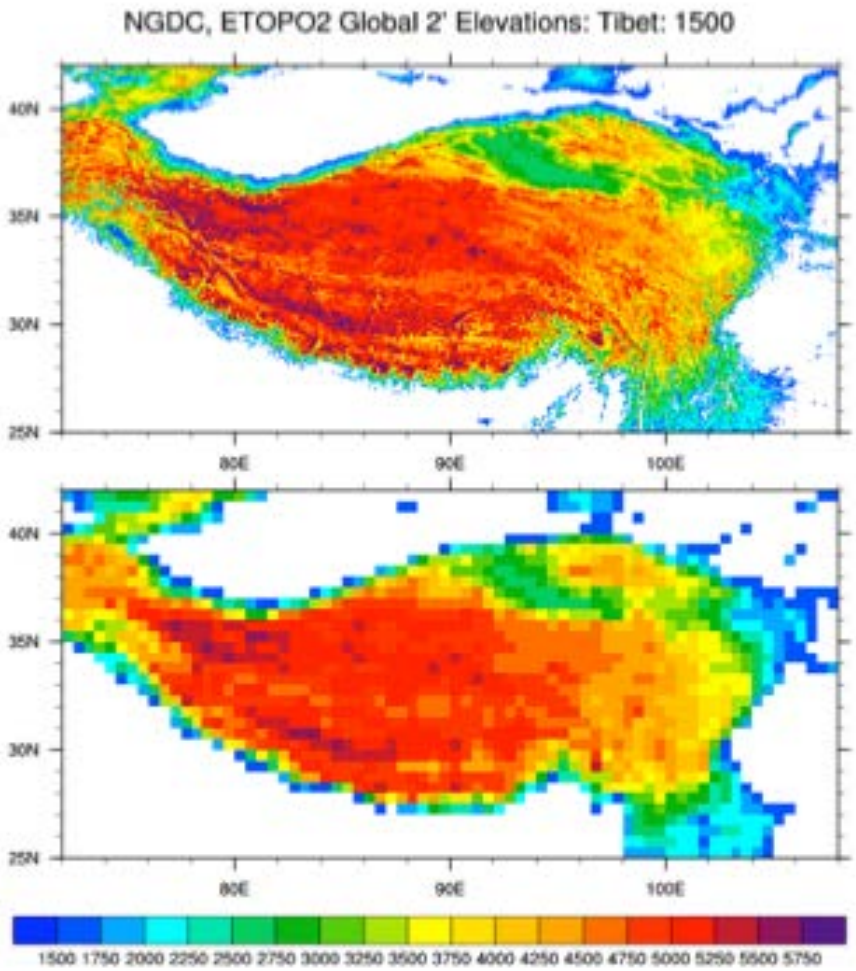
*Cameron Christensen**
*Giorgio Scorzelli**

Peer-Timo Bremer°
*Valerio Pascucci**

* www.sci.utah.edu

°

Visualization streaming



ANALYSIS AND VISUALIZATION

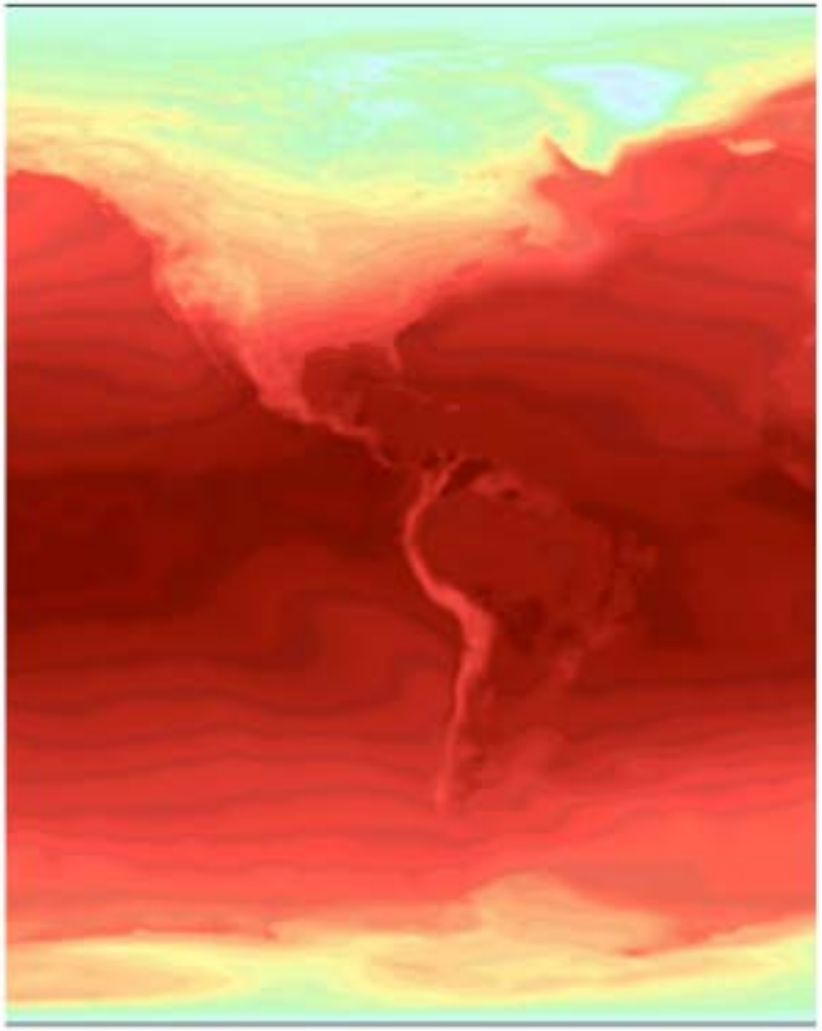
Click to add text

- Retrieve datasets to common location
- Regrid data to common domain
- Process using out-of-core or parallel supercomputer
- Downsample large domains or visualize using distributed parallel cluster application
- Store results
- Share

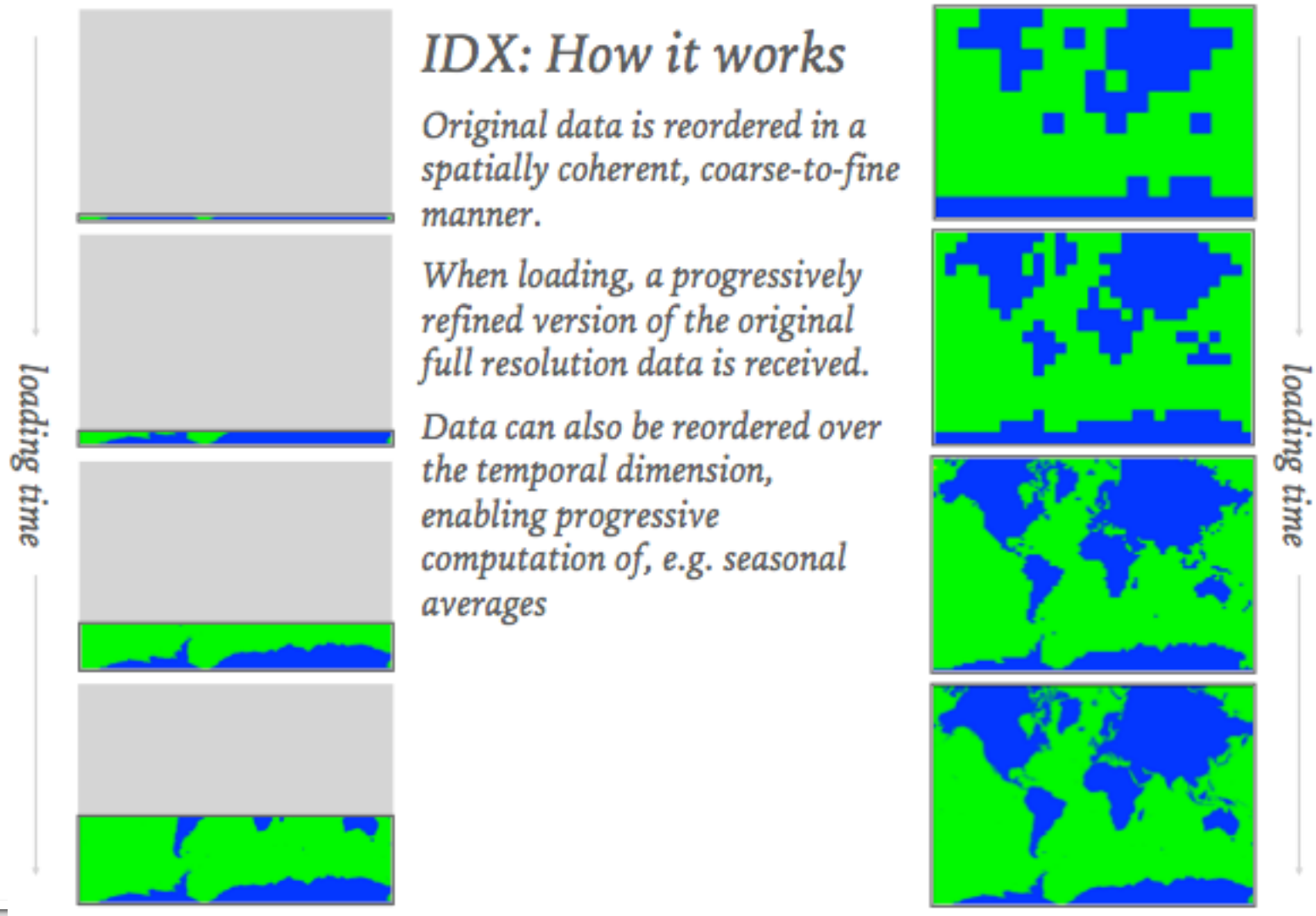
Stream multi-resolution

STREAMING MULTIRESOLUTION

- Interactive view-dependent data loading ➡permits flexible exploration outside fixed regions
- Fast coarse resolution results for visualization and analysis
- High resolution results through user-directed refinement
- Low requirements
  - low bandwidth: ➡download less
  - low storage and computation ➡process and cache only what you download
  - fast
- *you already use it every day!*



IDX format



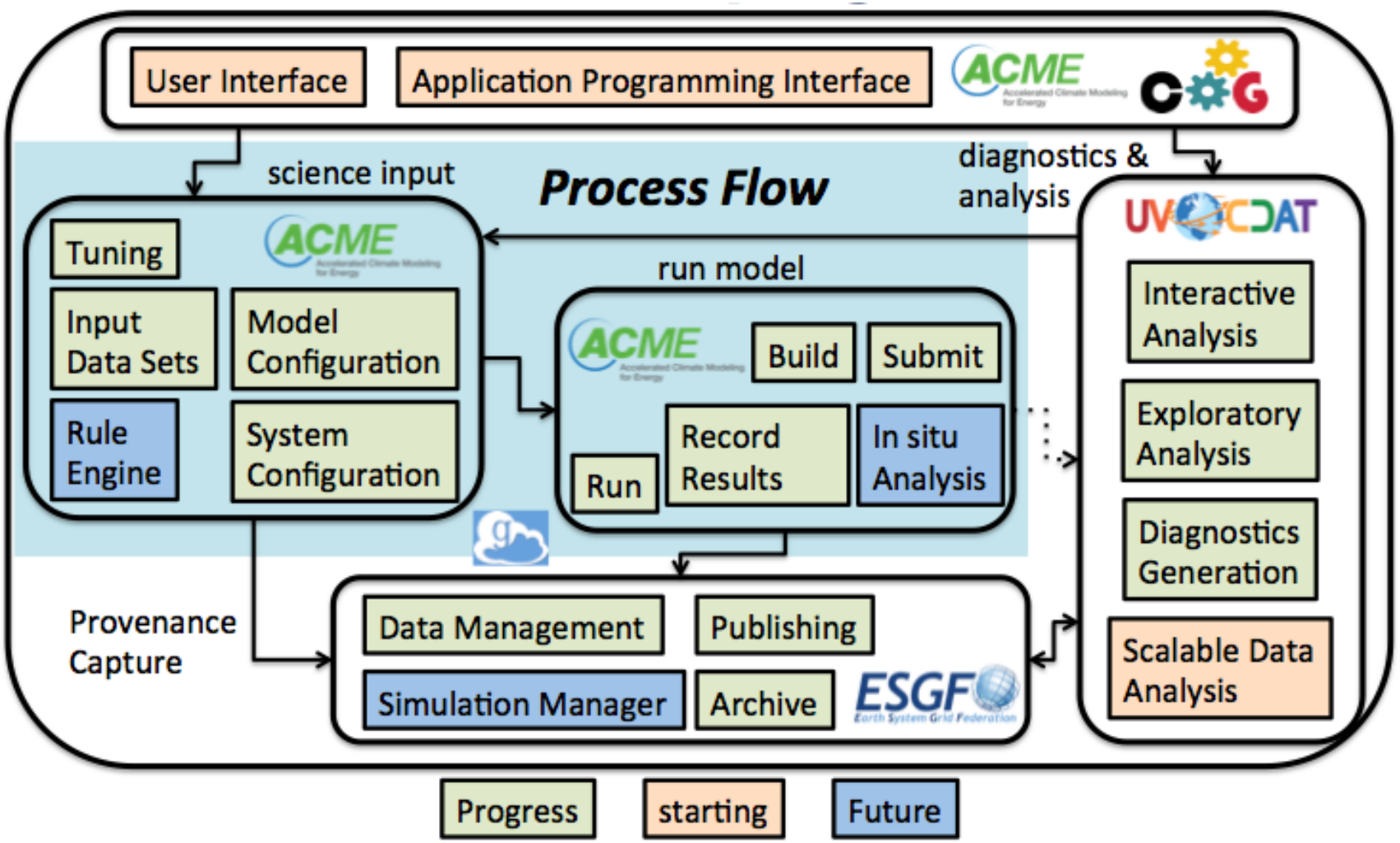
*coarse data is original, not a down-sampled copy

ESGF sets networking best practices into place to effectively transport tens of petabytes of climate data

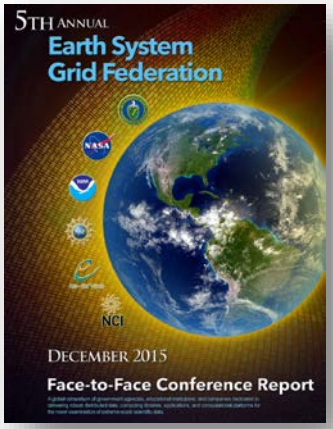
Immediate goal: 4 Gbps (1 PB/month) of sustained disk-to-disk data transfer between ESGF primary data centers

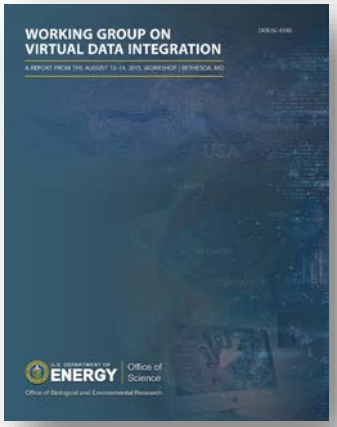
Stretch goal: 16 Gbps (1 PB/week) of sustained disk-to-disk data transfer between ESGF primary data centers

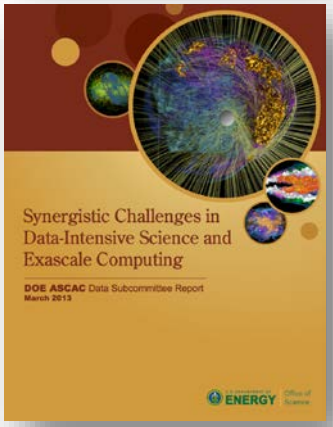
Accelerated Climate Modeling for Energy (ACME) end-to-end workflow

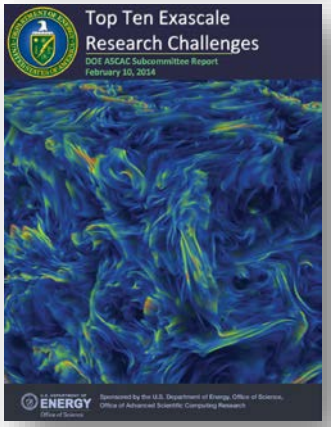


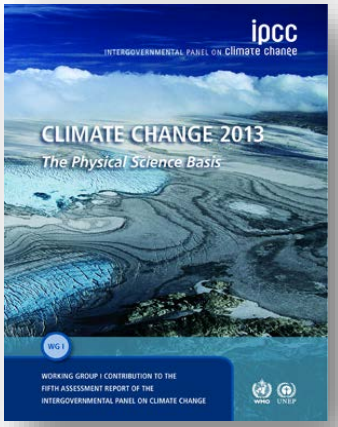
Data workshop and conferences reports: community involvement and outreach











DOE ESGF workshop and conference reports can be found at:
http://esgf.llnl.gov/reports.html

- esgf.llnl.gov; ESGF public website
- esgf.llnl.gov/reports.html; ESGF reports
- uvcdat.llnl.gov; UV-CDAT public website
- icnwg.llnl.gov; international network website
- github.com/esgf; ESGF software repository website
- github.com/uv-cdat; UV-CDAT software repository website