### CMIP archive infrastructure support

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Work Description</th>
<th>Effort (FTE’s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>riba metadata convention support, operational website support, CMOR data transport, quality control, statistical quality control, publication, replication, network, hardware, documentation, data access metrics Modeling group support, support for user community Multi-model operational outreach, modernization, operational coordination</td>
<td>Software Dev. Operations</td>
</tr>
<tr>
<td>Data Preparation</td>
<td>CF metadata convention support, operational website support, CMOR data transport, quality control, statistical quality control, publication, replication, network, hardware, documentation, data access metrics</td>
<td>3.7 0.75 1.5 0.4</td>
</tr>
<tr>
<td>Data Handling</td>
<td>data transport, quality control, statistical quality control, publication, replication, network, hardware, documentation, data access metrics</td>
<td>0.25 0.55</td>
</tr>
<tr>
<td>Help and Support</td>
<td>Modeling group support, support for user community</td>
<td>0.1 1.3</td>
</tr>
<tr>
<td>Administration &amp; Modernization</td>
<td>Multi-model operational outreach, modernization, operational coordination</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1.95 0.75 3.35 0.75</td>
</tr>
</tbody>
</table>
## Operations and Support Activities

### ESGF software system support

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Work Description</th>
<th>Effort (FTE’s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Software Dev.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Higher Priority</td>
</tr>
<tr>
<td>**Security ***</td>
<td>Support configuration, security integration, data transfer tools (WGET, Globus Online, OPeNDAP, GridFTP, etc.)</td>
<td>0.2</td>
</tr>
<tr>
<td>**Deep Storage Access and Transfers ***</td>
<td>Manage multiple concurrent transfers, securely replicating data via disparate networks</td>
<td>0.1</td>
</tr>
<tr>
<td><strong>LLNL</strong></td>
<td>leadership, data provider, publisher, replication, UI, QC, software stack, installation, product server, certificates</td>
<td>3.1</td>
</tr>
<tr>
<td>**User Interface ***</td>
<td>support for ESGF web environment: registration, virtual organization, browsing and searching</td>
<td>0.3</td>
</tr>
<tr>
<td><strong>Product Service</strong></td>
<td>Support for ESGF analysis and product services, dependence on the publication component</td>
<td>0.1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>0.4</td>
</tr>
</tbody>
</table>
Outline

- Background and History
- Operations and Support
- Usage Statistics
- Lessons Learned and Future Operational Support
Background, History, and the Relationship of CMIP and ESG
Background and History

International context of climate research: Coupled Model Intercomparison Projects (CMIPs)

- The Intergovernmental Panel on Climate Change (IPCC) prepares an **assessment report** about climate science approximately **every 6 years**

- This climate modeling research requires enormous **scientific and computational resources**
  - Involves over **two dozen modeling research groups worldwide**; and
  - Encourages **cooperative community international research activities**

- The **World Climate Research Program (WCRP)** serves as the primary coordinating body for this research activity.
  - The WCRP’s **Working Group on Coupled Modelling (WGCM)** relies on the **Lawrence Livermore National Laboratory (LLNL)** to support these activities by coordinating and maintaining the distributed petabyte data archive.
Background and History

### Coupled Model Intercomparison Project (CMIP) data history

- **CMIP = Coupled Model Intercomparison Project**
  - Phase 1: Idealized simulations of present-day climate (early 90s) (~1 Gigabyte (GB))
  - Phase 2: Idealized simulations of future climate changes (late 90s) (~500 GB: CMIP2/CMIP1=500)

- **CMIP 5 multi-model archive expected to include (2010 – present) (3.5 Petabytes (PB): CMIP5/CMIP3 = 100):**
  - 3 suites of experiments (“Near-Term” decadal prediction, “Long-Term century & longer), and “Atmosphere-Only”)
  - 25+ modeling centers in 21 countries
  - 60+ models
  - Total data, ~3.5 PB; Replica 1.2 – 2 PB; Derived data ~1 PB
  - Driver for scale of data, global distribution
  - Timeline fixed by IPCC (September 2013)

- **LLNL organizes, manages and distributes the CMIP database**

- **CMIP6 (2016) (350 PB – 3 EB ?)**
Coupled Model Intercomparison Project, phase 3 (CMIP3) impact

- More than 600 journal articles published
- Nearly all the new, model-based conclusions in the IPCC’s AR4
  - ~75% of the more than 100 figures in 3 major chapters of the report are based on CMIP3 results.
  - 4 of the 7 figures appearing in the “Summary for Policy Makers” are based on CMIP3
  - The multi-model perspective provided the foundation for many of the IPCC conclusions.

Interest in LLNL’s CMIP3 model output archive continues unabated
~600 GB/day downloaded

Daily Download Rate

- 1 TB
- IPCC deadline for paper submissions
- End of ESG2 SciDAC Project
- IPCC AR4 released

2004  2010
## Background and History

### Earth System Grid (ESG): ESG-I, ESG-II, ESG-CET

- **ESG-I** funded under DOE’s Next Generation Internet (NGI), “address the emerging challenge of climate data”, 1999 – 2001 (ANL, LANL, LBNL, LLNL, NCAR, USC/ISI)
  - Data movement and replication; Prototype climate “data browser”; Hottest Infrastructure” Award at SC’ 2000.

- **ESG-II** funded under DOE’s Scientific Discovery through Advanced Computing (SciDAC), “turning climate data sets into community resources”, 2001-2006 (ORNL addition)
  - Web-based portal, metadata, access to archival storage, security, operational services, 2004 first operational portal CCSM (NCAR), IPCC CMIP3/AR4 (LLNL); 200 TB of data, 4,000 users, 130 TB served.

- **ESG-CET** funded under DOE’s Scientific Discovery through Advanced Computing II (SciDAC-2), “providing climate researchers worldwide with access to: data, information, models, analysis tools, and computational resources required to make sense of enormous climate simulation and observational data sets”, 2006 – 2011 (PMEL addition)
  - 2010 Awarded by American Meteorological Society (AMS) for leadership which led to a new era in climate system analysis and understanding.
  - CMIP3, CMIP5, CCSM, POP, NARCCAP, C-LAMP, AIRS, MLS, Cloudsat, etc.
  - 22,000 users, 500-800 users active per month, over 1 PB served
## Background and History

### Earth System Grid Federation (ESGF): 2011 - present

- **ESGF multiple funding sources**, “an open consortium of institutions, laboratories and centers around the world that are dedicated to supporting research of climate change, and its environmental and societal impact” (2011 – present)
- Historically originated from Earth System Grid project, expanded beyond its constituency and mission to include many other partners in U.S., Europe, Asia, and Australia
- Integration with many external projects: ESC, Metafor, GIP, etc.
- **U.S. funding sources**: DOE, NASA, NOAA, NSF
- **U.S. Participants**: PMEL/NOAA, LLNL/DOE, LBNL/DOE, NCAR/NSF, LANL/DOE, ORNL/DOE, ANL/DOE, JPL/NASA, GFDL/NOAA, ESRL/NOAA
- **Europe**: IS-ENES, BADC, UK-MetOffice, DKRZ, MPIM, IPSL, LSCE
- **Asia**: Univ. of Tokyo, Japanese Centre for Global Environmental Research, Jamstec, Korea Meteorological Administration
- **U.S. Universities**: ISI/USC, Purdue University, University of Michigan, UC Davis, …
- … and more …
Background and History

Coupled Model Intercomparison Project, phase 5 (CMIP5)

- Working with over **30 key national and international partners** to establish a federated data archival and retrieval system
- LLNL is in cooperation with the **IPCC Data Distribution Centre (DDC)** for supporting access of impacts researchers to impacts-related variables coming out of CMIP5 model runs
Operations and Support of CMIP
The ESGF distributed data archival and retrieval system

- Distributed and **federated architecture**
- Support discipline specific **portals**
- Support **browser-based** and direct client access
- **Single Sign-on**
- Automated script and GUI-based publication tools
- Full support for **data aggregations**
  - A collection of files, usually ordered by simulation time, that can be treated as a single file for purposes of data access, computation, and visualization
- **User notification service**
  - Users can choose to be notified when a data set has been modified

http://esgf.llnl.gov
CMIP5 data processing stream

- **Publishing data** to an ESGF portal performs QC Level 1 (QCL1) check
  - QCL1 data are visible to users and are identified as QCL1 on the UI
- **DKRZ (MPI) quality control code** is run on data to perform QC Level 2 (QCL2) check
  - QCL2 data are visible to users and are identified as QCL2 on the UI
- **Visual inspections** are performed for inconsistencies and metadata correctness at QC Level 3 (QCL3) check
  - QCL3 data are visible to users and are identified as QCL3 on the UI
  - **Digital Object Identifiers (DOIs)** are given to data sets that pass the QCL3 check
ESGF software system integrates data federation services (i.e., data services)

- NetCDF Climate and Forecast (CF) Metadata Convention
  - (LibCF)
  - Mosaic
- Climate Model Output Rewriter 2 (CMOR-2)
- Regridders: GRIDSPEC, SCRIP, & ESMF
- Publishing
- Search & Discovery
- Replication and Transport
  - GridFTP, OPeNDAP, DML, Globus Online, ftp, BeSTMan (HPSS)
  - Networks
- Data Reference Syntax (DRS)
- Common Information Model (CIM)
- Quality Control
  - QC Level 1, QC Level 2, QC Level 3, Digital Object Identifiers (DOIs)
- Websites and Web Portal Development
  - Data, Metadata, Journal Publication Application
- Notifications, Monitoring, Metrics
- Security
- Product Services
  - UV-CDAT, Live Access Server
Scaling climate applications to take advantage of increasing hardware and network bandwidth

- (1) Users communicate with ESGF front-end servers on the LLNL enterprise network via HTTP
- (2) Large data sets are made available to users via GridFTP from the CSS Data Transfer Nodes (DTNs) located in the LLNL Science DMZ
- (3) ESGF may perform analysis of raw data if requested by users through the front-end servers
Summary of technologies most useful to ESGF

- Peer-to-peer
- Search Services
- Security Services
- User interface development
- Data Publisher
- Live Access Server (compute node)
- Data transfer (WGET, BDM, GridFTP, Globus Online, BeStMan)
- DAP Services (THREDDS Data Servers[TDS], OPeNDAP)
- Really Simple Syndication (RSS) feed
- UV-CDAT (client analysis tool access)
- Dashboard (system monitor service)
- Replication and versioning
- Installation Script
### ESGF climate data holdings and growing

- Phases 3 and 5 of the Coupled Model Intercomparison Project (CMIP3 and CMIP5)
- Coordinated Regional climate Downscaling Experiment (CORDEX)
- Climate Science for a Sustainable Energy Future (CSSEF)
- European Union Cloud Intercomparison, Process Study & Evaluation Project (EUCLIPSE)
- Geo-engineering Model Intercomparison Project (GeoMIP)
- Land-Use and Climate, Identification of robust impacts (LUCID)
- Paleoclimate Modeling Intercomparison Project (PMIP)
- Transpose-Atmospheric Model Intercomparison Project (TAMIP)
- Clouds and Cryosphere (cloud-cryo)
- Observational products more accessible for coupled model intercomparison (obs4MIPs)
- Reanalysis for the coupled model intercomparison (ANA4MIPs)
- Dynamical Core Model Intercomparison Project (DCMIP)
- Community Climate System Model (CCSM)
- Parallel Ocean Program (POP)
- North American Regional Climate Change Assessment Program (NARCCAP)
- Carbon Land Model Intercomparison Project (C-LAMP)
- Atmospheric Infrared Sounder (AIS)
- Microwave Limb Sounder (MLS)
Current CMIP federated portals (e.g., LLNL’s portal address: pcmdi9.llnl.gov)
Operations and Support

ESGF nodes deployment (map)
Help desk: ESGF Askbot

Help
Welcome,
Thank you for using Askbot, here is how it works.

How questions, answers and comments work
This site is for asking and answering questions, not for open-ended discussions. We encourage everyone to use "question" space for asking and "answer" for answering.
Despite that, each question and answer can be commented – the comments are good for the limited discussions.

Please search before asking your questions
Type your question in the search bar and see whether a similar question has been asked before.
Search has advanced capabilities:
• to search in title - enter [title: your text]
• to search by tags - enter [tag: sometag] or #sometag
• to search by user - enter [user: somename] or @somename or "some name"
In addition, it is possible to click on tags to add them to the search query.
Important!!! All search terms are combined with a logical "AND" expression - to narrow the search by adding new terms.

Voting
Voting in Askbot helps to select best answers and thank most helpful users.
Please vote when you find helpful information, it really helps the Askbot community.

Other topics
You can @mention users anywhere in the text to point their attention, follow users and conversations and report inappropriate content by flagging it.
Enjoy.
Operations and Support

Ultra-scale Visualization Climate Data Analysis Tools (UV-CDAT)

ESGF Data Archive
ParaView Cluster parallel processing

UV-CDAT/ESGF UI access

Script

http://uv-cdat.llnl.gov
Usage Statistics

Cumulative CMIP5 downloaded data from LLNL (TBs)
Usage Statistics

Cumulative CMIP5 number of distinct downloaded files

<table>
<thead>
<tr>
<th>Time</th>
<th># of distinct downloaded files</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011/09</td>
<td>5</td>
</tr>
<tr>
<td>2011/10</td>
<td>9</td>
</tr>
<tr>
<td>2011/11</td>
<td>13</td>
</tr>
<tr>
<td>2011/12</td>
<td>18</td>
</tr>
<tr>
<td>2012/01</td>
<td>231</td>
</tr>
<tr>
<td>2012/02</td>
<td>417</td>
</tr>
<tr>
<td>2012/03</td>
<td>1526</td>
</tr>
<tr>
<td>2012/04</td>
<td>1628</td>
</tr>
<tr>
<td>2012/05</td>
<td>2316</td>
</tr>
<tr>
<td>2012/06</td>
<td>41562</td>
</tr>
<tr>
<td>2012/07</td>
<td>52708</td>
</tr>
<tr>
<td>2012/08</td>
<td>83177</td>
</tr>
<tr>
<td>2012/09</td>
<td>126452</td>
</tr>
<tr>
<td>2012/10</td>
<td>218350</td>
</tr>
<tr>
<td>2012/11</td>
<td>255161</td>
</tr>
<tr>
<td>2012/12</td>
<td>275117</td>
</tr>
<tr>
<td>2013/01</td>
<td>310228</td>
</tr>
<tr>
<td>2013/02</td>
<td>360213</td>
</tr>
<tr>
<td>2013/03</td>
<td>386754</td>
</tr>
<tr>
<td>2013/04</td>
<td>538886</td>
</tr>
<tr>
<td>2013/05</td>
<td>609835</td>
</tr>
<tr>
<td>2013/06</td>
<td>609835</td>
</tr>
</tbody>
</table>
Geo-distribution related to the users that have downloaded CMIP5 data sets from pcmdi9 (~2000 IPs in total)
Usage Statistics

User distribution percentage by country (~2000 distinct IPs in total)

Client distribution (%) by continents (2000 distinct IPs in total)

- America 47.44%
- Europe 20.50%
- Asia 27.89%
- Australia 2.00%
- Africa 0.61%
- No-resolv 1.56%

© 2013 Analytics and Informatics Management Systems
## ESGF registered CMIP5 users distributed (table)

<table>
<thead>
<tr>
<th>Host Name</th>
<th>City</th>
<th>Registered Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>pcm9.ind1.llnl.gov</td>
<td>Livermore</td>
<td>5977</td>
</tr>
<tr>
<td>esgf-data.dkz.de</td>
<td>Hamburg</td>
<td>769</td>
</tr>
<tr>
<td>esgf-data.jpl.nasa.gov</td>
<td>Sylmar</td>
<td>128</td>
</tr>
<tr>
<td>dev.esg.anl.gov</td>
<td>Lemont</td>
<td>62</td>
</tr>
<tr>
<td>esgf-node.ipsl.fr</td>
<td>Paris</td>
<td>58</td>
</tr>
<tr>
<td>esgf-index1.ceda.ac.uk</td>
<td>Appleton</td>
<td>30</td>
</tr>
<tr>
<td>esgf.nccs.nasa.gov</td>
<td>Huntsville</td>
<td>27</td>
</tr>
<tr>
<td>esgdata.gfdl.noaa.gov</td>
<td>Princeton</td>
<td>25</td>
</tr>
<tr>
<td>eclipse1.dkz.de</td>
<td>Hamburg</td>
<td>20</td>
</tr>
<tr>
<td>adn07.cmcc.it</td>
<td>Salento</td>
<td>18</td>
</tr>
<tr>
<td>pcm911.ind1.llnl.gov</td>
<td>Livermore</td>
<td>17</td>
</tr>
<tr>
<td>esgf.bnu.edu.cn</td>
<td>Beijing</td>
<td>14</td>
</tr>
<tr>
<td>esgf1.nersc.gov</td>
<td>Berkeley</td>
<td>11</td>
</tr>
<tr>
<td>esgf.oms.orl.gov</td>
<td>Oak Ridge</td>
<td>6</td>
</tr>
<tr>
<td>cmip5.bnu.edu.cn</td>
<td>Beijing</td>
<td>2</td>
</tr>
<tr>
<td>esgf-cn1.nsc.liu.se</td>
<td>Linköping</td>
<td>2</td>
</tr>
<tr>
<td>wavona.ca.sandie.gov</td>
<td>Livermore</td>
<td>1</td>
</tr>
<tr>
<td>esgnode2.nci.org.au</td>
<td>Canberra</td>
<td>1</td>
</tr>
<tr>
<td>esgfvm3.ind1.llnl.gov</td>
<td>Livermore</td>
<td>1</td>
</tr>
<tr>
<td>esgf-data1.ceda.ac.uk</td>
<td>Appleton</td>
<td>1</td>
</tr>
<tr>
<td>vsgf.ipsl.polytechnique.fr</td>
<td>Palaiseau</td>
<td>0</td>
</tr>
<tr>
<td>esg.lasg.ac.cn</td>
<td>Beijing</td>
<td>0</td>
</tr>
<tr>
<td>cmip3.dkz.de</td>
<td>Hamburg</td>
<td>0</td>
</tr>
<tr>
<td>esgf-dev.dkz.de</td>
<td>Hamburg</td>
<td>0</td>
</tr>
<tr>
<td>vsgf.ipsl.fr</td>
<td>Paris</td>
<td>0</td>
</tr>
<tr>
<td>esgf-data1.nccs.nasa.gov</td>
<td>Huntsville</td>
<td>0</td>
</tr>
<tr>
<td>bmbf-ipcc-ar5.dkz.de</td>
<td>Hamburg</td>
<td>0</td>
</tr>
</tbody>
</table>
Usage Statistics

Users distribution by identity provider

- cmip5
- cmip3
- euflipe1.dkrz.de
- esg-dn1.nsc.liu.se
- ceda.ac.uk
- dev.esg.anl.gov
- esg-datanode.jpl.nasa.gov
- esg-gateway.jpl.nasa.gov
- esg-vm-demo02.ccs.ornl.gov
- esg.ccs.ornl.gov
- esg.nci.org.au
- esg1-gw.pnl.gov
- esg2.nci.org.au
- esgf-data.dkrz.de
- esgf-node.ipsl.fr
- ipcc-ar5.dkrz.de
- pcmi3.llnl.gov
- pcmi9.llnl.gov
- www.earthsystemgrid.org
Lessons Learned and Future Operational Support
Lessons learned

- ESGF infrastructure is under constant requirements to improve and adapt
- ESGF must continue to rely on careful integration of already proven technologies and applications that have been developed by teams over the course of many years (e.g., Solr, TDS, UV-CDAT, HTTP, OPeNDAP, SSL, …)
- Promote participation and involvement by a large community of stakeholders, managers, and engineers, through an open source, meritocracy based system (not dissimilar to the principles promoted by the Apache Software Foundation, for example)
- Establish a governance model from the very beginning, in order to represent the interests of all stakeholders, prioritize requirements, and guide the overall system development
- Avoid single points-of-failure in the engineering workforce
- Large infrastructure like ESGF should consider scalability as one of its major requirement (e.g., data discovery, movement, processing, etc. testing should be scaled to 10 to 100 times the current amount of data)
- Funding is always a struggle (U.S. and EU agencies tend to fund innovative research and new ideas and less prone to support ongoing successful projects such as ESGF
### Governance of ESGF

- **Steering Committee:** Funding agencies and stakeholders responsible for providing resources.
- **Executive Committee:** Overall responsibility for meeting sponsors, stakeholders, and community needs and prioritizing work.
- **Technical Committee:** Responsible for the development of the system architecture, the management of the development lifecycle and scheduling releases.
- The executive and technical committees are responsible for setting release contents and reporting what was actually delivered in the releases to the community.
### Future Operations and Support

#### Distributive technologies that need more research for ESGF use

- Machine learning for pattern discovery and prediction
- Decision analytics based on the quantification of uncertainties
- Streaming analysis, visualization and sensors simulation
- Full suite of server-side analysis and visualization
- Hadoop and distributed computing
- Cloud computing installation and VMs
- Streaming clients for data download
- Research possible alternative search technologies such as NoSQL and MongoDB
- Research alternative communication protocols to P2P
- Scalable semantic technologies
- Research quality control checking on the data
- Digital Object Identifiers (DOIs)
- Server-side data reduction and calculations
- Workflow and provenance
Funding of ongoing operational and software support for CMIP and climate science data capabilities is a critical issue

- Software and Model Archive Federated Data System Components have been developed entirely as 3-5 year R&D projects through support from BER and ASCR.
- Over the last few years, we were generously given additional support by BER.
- The current budget for AIMES under the SFA provides only one FTE of support.
- From the charts estimate that multiple FTE’s are needed:
  - CMIP community outreach and engagement
    - participation metadata standards definition committees
    - Collaboration and infrastructure design and deployment with other BER and USGCRP climate data centers
  - Operational support of the ESGF-based CMIP model data distribution system
    - Ongoing ESGF software support
    - Assistance to modeling centers for software deployment and maintenance
  - Maintenance of CMIP replicated data archive housed at LLNL
  - Servicing requests driven by the DOE, CMIP, and the research community
  - …..
Questions and Discussion