

CMIP, CSSEF, ESGF, and BER Related Data Projects: Climate Model Analysis, Visualization, and Test Bed Efforts for Ultra-Large Data Sets

Dean N. Williams and Dave Bader on behalf of Multiple Earth System Communities and Projects
Federated and Integrated Data from Multiple Sources

CESD Data Meeting – Infrastructure and Framework Presentation ◆ June 26, 2012



Outline

- CMIP5 – Redefined requirements for data federation
- ESGF concept
- ESGF and CMIP
- ESGF pilot projects with ARM and CDIAC
- CSSEF
- Associated BER efforts
- Current status and near term directions
- A possible long-term vision for BER’s “Climate K-Base”

CMIP: Experiment Design

CMIP5: 47 models **available** from 21 centers

- CMIP = **Coupled Model Intercomparison Project**
 - Phase 1: Idealized simulations of present-day climate
 - Phase 2: Idealized simulations of future climate changes
 - Phase 3: More realistic simulations (2004 – present)
- **CMIP 5** multi-model archive expected to include
 - **3** suites of experiments
 - **24** modeling centers in **19** countries
 - **58** models
 - Total data, **~3.5 PB**
 - Replica **1 – 2 PB**
 - Derived data **~1 PB**
- **Global distribution**
- **Timeline fixed** by IPCC (2012 - 2013)
- **LLNL organizes, manages and distributes** the CMIP/IPCC (Intergovernmental Panel on Climate Change) database of climate model output

CMIP3 archive vs. CMIP5 archive

CMIP3 Modeling Centers		volume (GB)
NCAR	USA	9,173
MIROC3	Japan	3,975
GFDL	USA	3,843
IAP	China	2,868
MPI	Germany	2,700
CSIRO	Australia	2,088
CCCma	Canada	2,071
INGV	Italy	1,472
GISS	USA	1,097
MRI	Japan	1,025
CNRM	France	999
IPSL	France	998
UKMO	UK	973
BCCR	Norway	862
MIUB	Germany/Korea	477
INMCM3	Russia	368
Totals		34,989 (TB)

35 TB

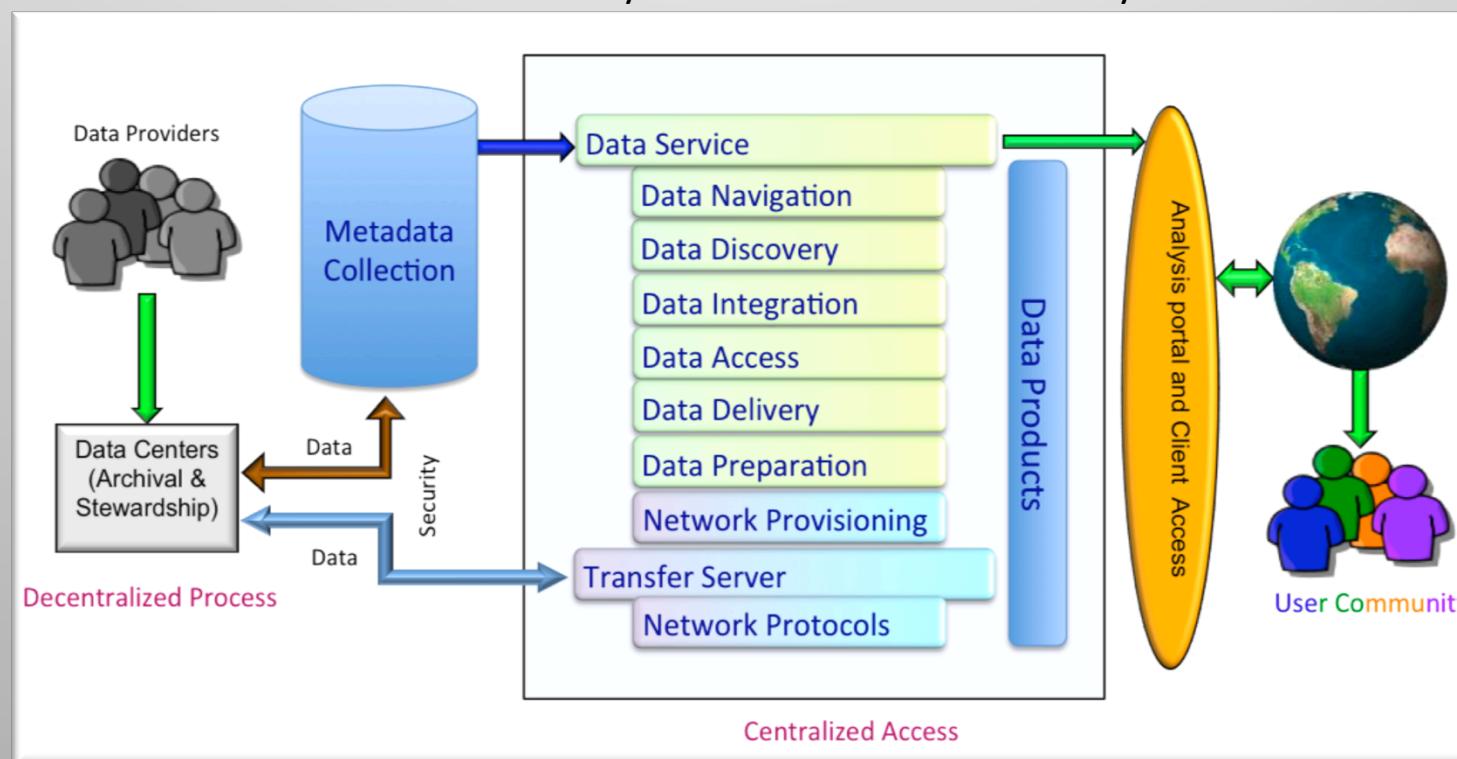
CMIP5 Modeling Centers		volume (TB)
BCC	China	51
CCCma	Canada	51
CMCC	Europe (Italy)	158
CNRM	France	71
CSIRO	Australia	81
EC-EARTH	Europe (Netherlands)	97
GCESS	China	24
INM	Russia	30
IPSL	France	121
LASG	China	100
MIROC	Japan	350
MOHC	UK	195
MPI	Germany	166
MRI	Japan	269
NASA	USA	375
NCAR	USA	739
NCC	Norway	32
NCEP	USA	26
NIMR/KMA	Korea	14
NOAA GFDL	USA	158
Totals		3,108 (PB)

3.1 PB
Currently 1.2 PB

CMIP5/CMIP3 = 10²

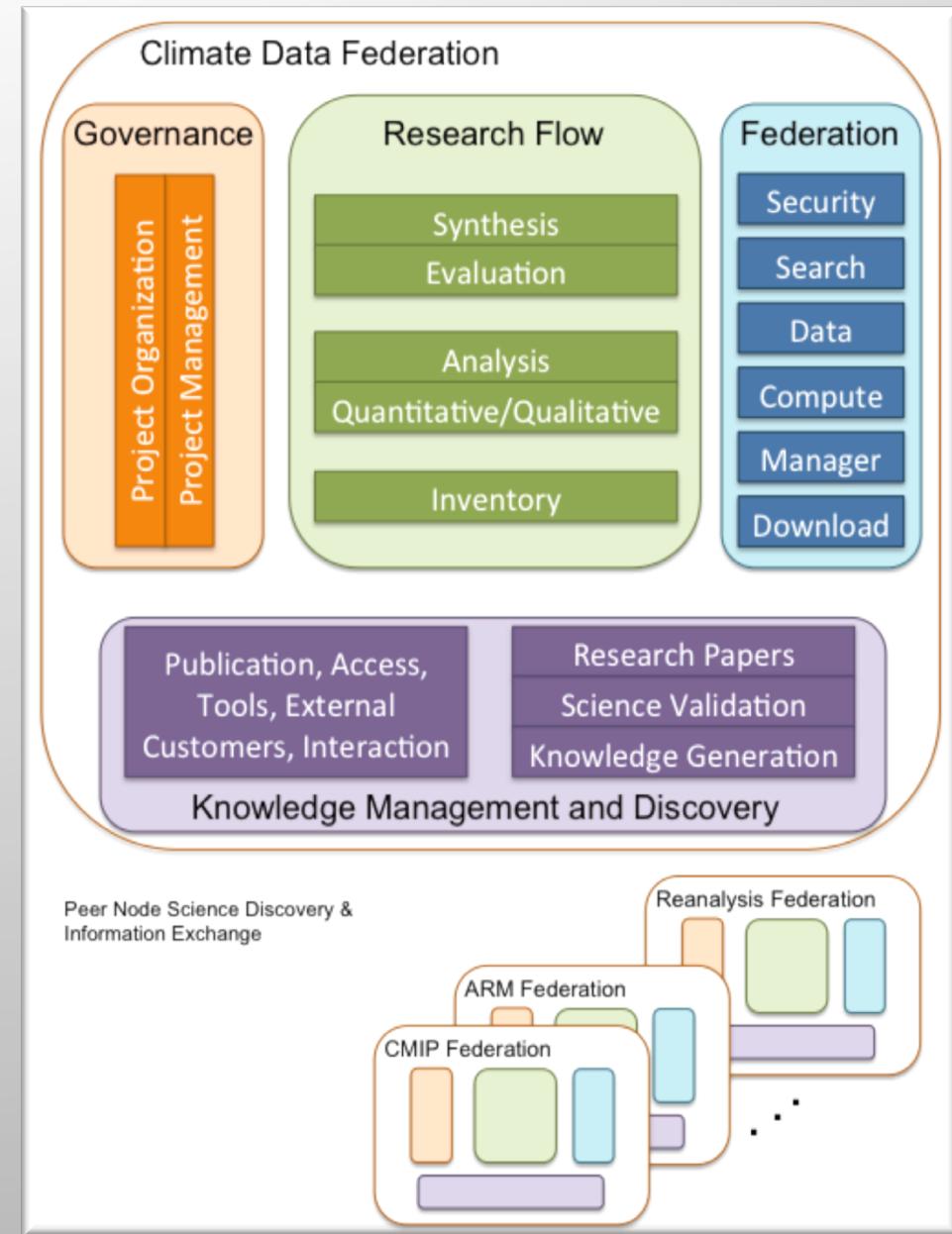
Big Data Challenges

- **Centralized process** for CMIP3, data was **shipped to LLNL via disk**, then made available to users
 - Several weeks delay between generation of files and availability to users
 - Correction errors in data was too difficult
 - LLNL became a single point of failure (a catastrophic crash of our RAID system meant data was inaccessible for several weeks)
 - Backup was incredibly difficult, which made restoration equally as hard
- **Decentralized process** for CMIP5, a new (and **more complicated approach**) was necessary
 - Distributed data Archive
 - No single point of failure
 - Replication of heavily-used data sets
 - Data can be made available without delay and can be corrected locally



Data Federation Services

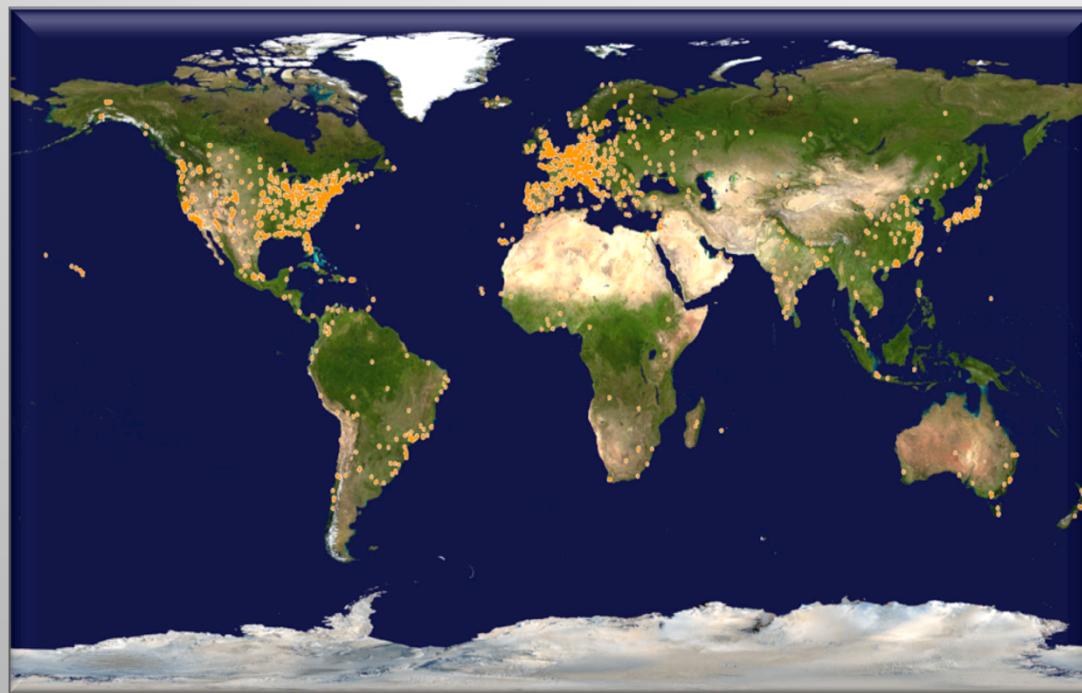
- NetCDF Climate and Forecast (CF) Metadata Convention
 - (LibCF)
 - Mosaic
- Climate Model Output Rewriter 2 (CMOR-2)
- GRIDSPEC & SCRIP
- Publishing
- Search & Discovery
- Replication and Transport
 - GridFTP, OPeNDAP, DML, Globus Online, ftp, BeSTMan (HPSS)
 - Networks
- Data Reference Syntax (DRS)
- 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
- Product Services
 - Live Access Server, UV-CDAT
- Security



Earth System Grid Federation (ESGF)

Approach

Free and open consortium of institutions, laboratories and centers around the world that are dedicated to **supporting climate change research** and its environmental and societal impact.

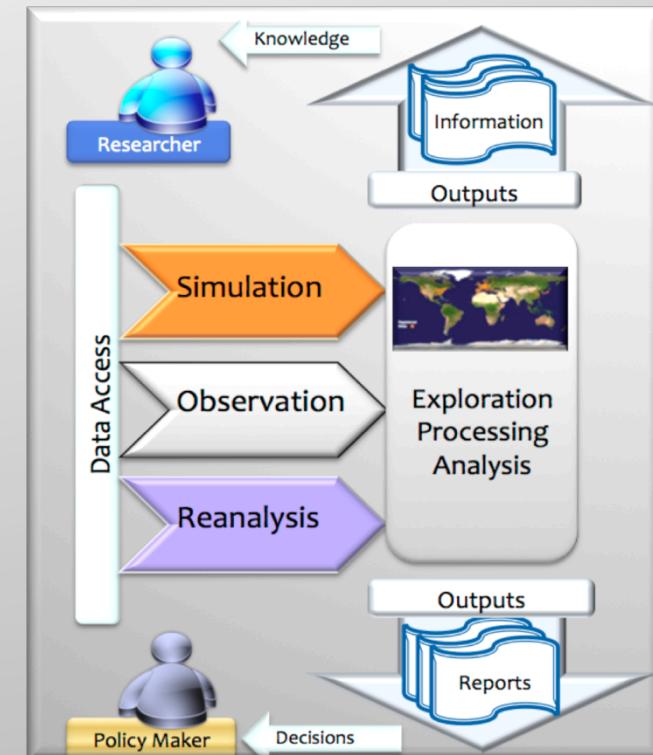


Contributions

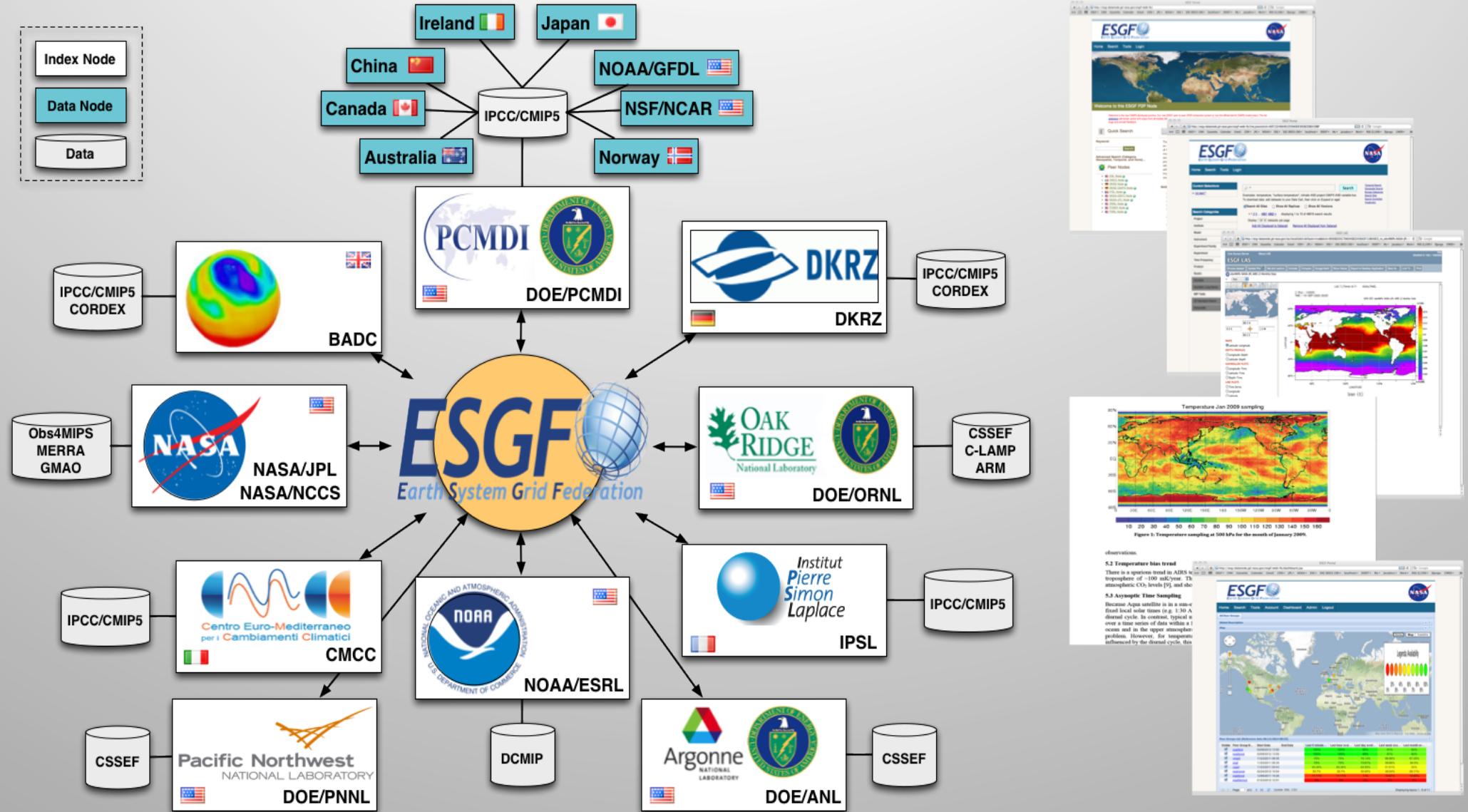
Worldwide organizations contributing to the ESGF efforts include: NASA's Observation Intercomparison Project (obs4MIPs); NOAA's Earth System Curator and Global Interoperability Program; EU's MetaFor and Infrastructure for the European Network for Earth System Modeling, and many more.

History

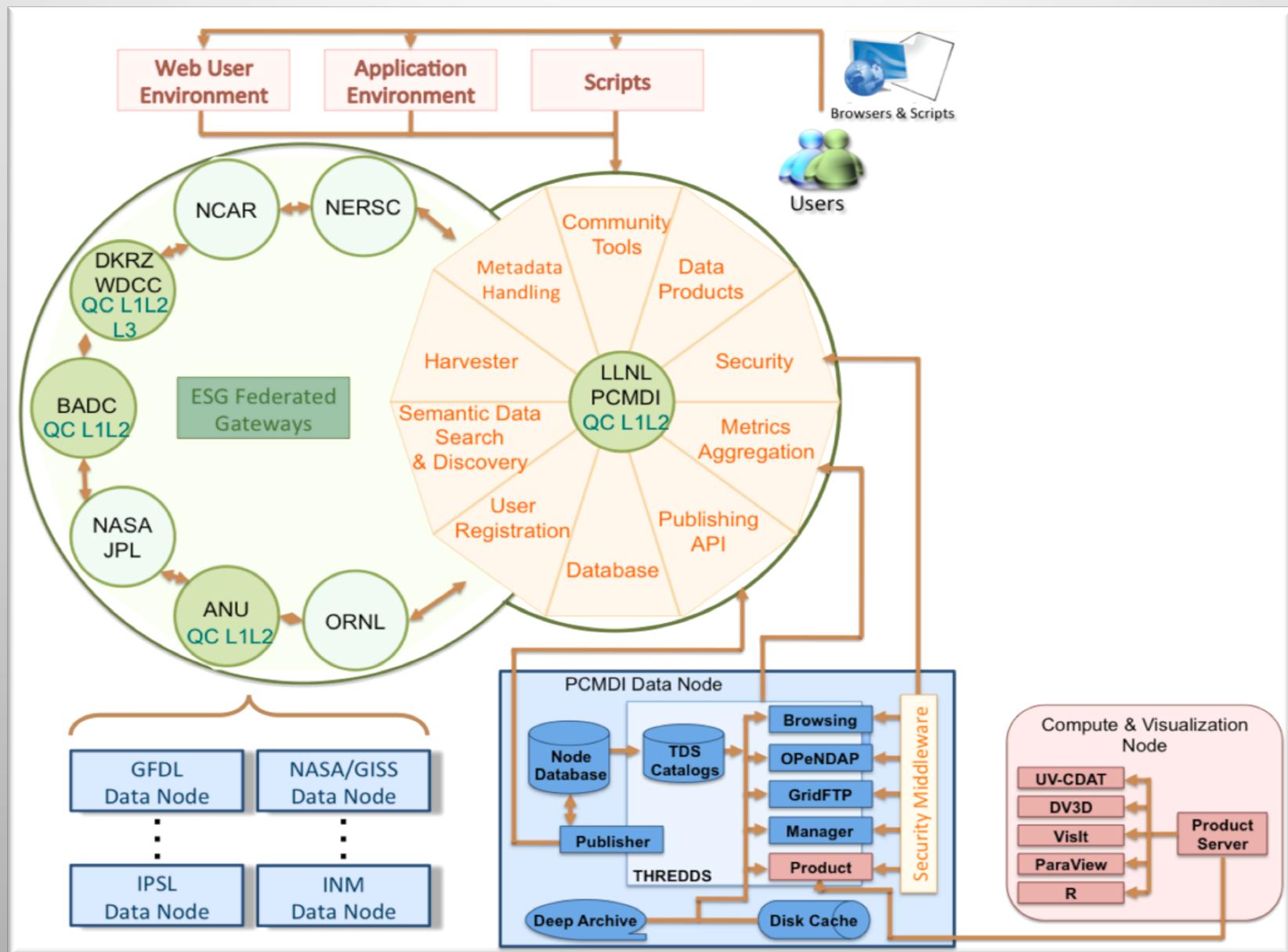
Historically originated from the Earth System Grid (ESG) project, started in 1999, expanded beyond its constituency and mission to include many other projects and **partners in the U.S., Europe, Asia, and Australia**.



Federated and Integrated Data from Multiple Sources

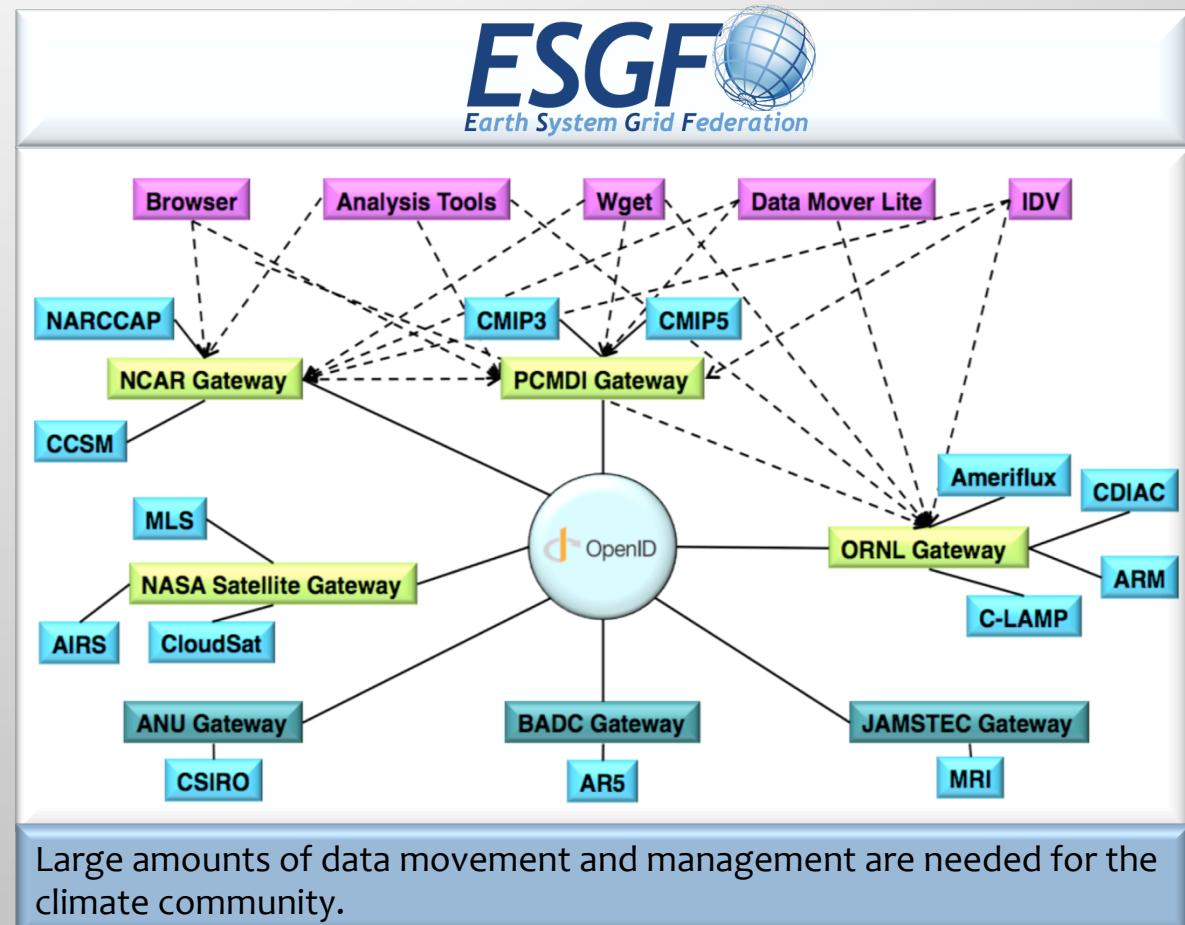


CMIP5 Managed Distributed Archive led by the Earth System Grid Federation (ESGF)



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



ARM/CDIAC Integration with the Earth System Grid Federation (ESGF)

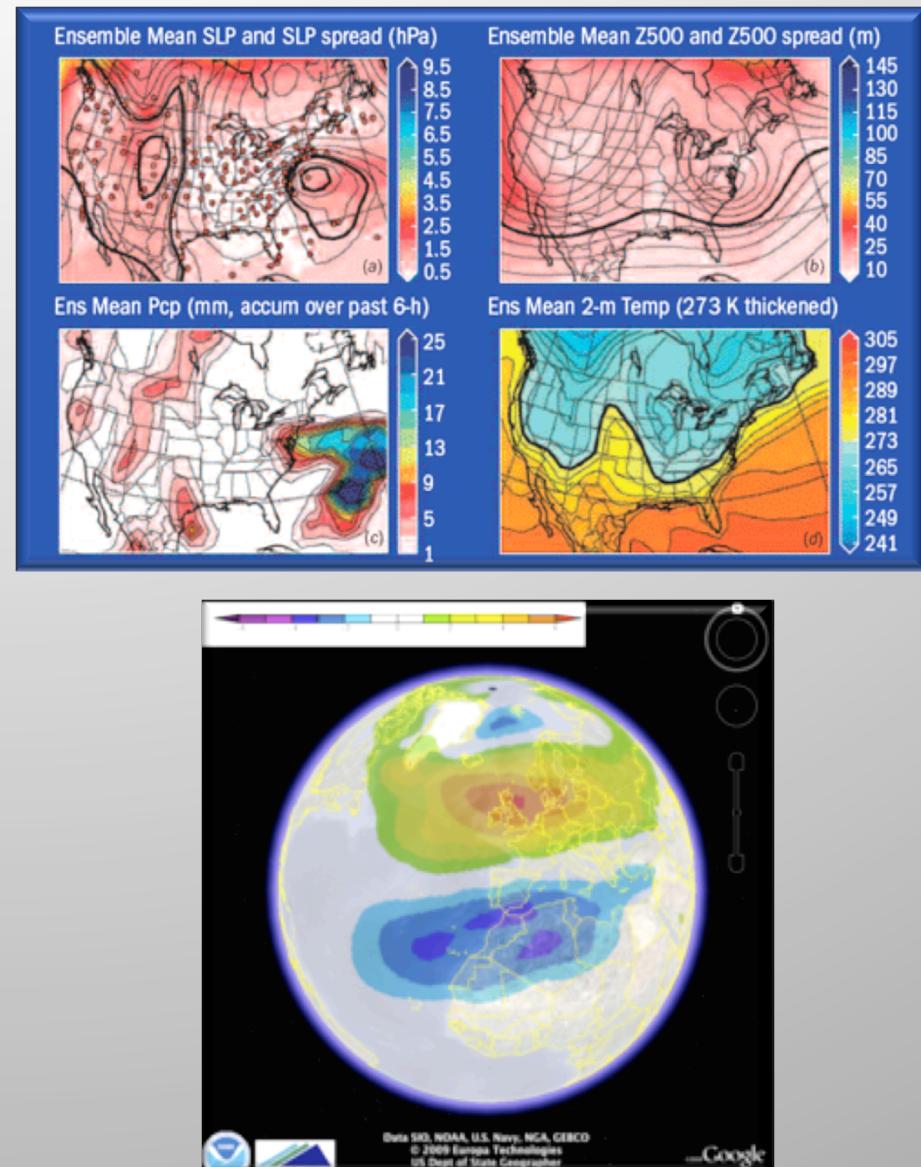
- ARM/CDIAC data set publication plan

Currently Published	To Be Published
Ameriflux	Balloon-Borne Sounding System
Fossil Fuels	Active Remotely-Sensed Cloud Locations
Obs4MIPs	Microwave Water Radiometer

- Revised workflow for seamless publication of ARM/CDIAC data sets to ESGF
 - ARM/CDIAC centers maintain a THREDDS server containing ESGF-parsable metadata catalogs of native archives
 - An ESGF “Plugin” is built containing:
 - Flexible Publication tools ([EZ-Pub](#))
 - A catalog crawler for ingestion into the ESGF search index to enable data set discovery
 - Modules for data visualization, analysis, and mining
 - Security mechanisms for user authentication and verification
 - Usage metrics and statistics collectors for policy makers and administrators

Reanalysis Integration with the Earth System Grid Federation (ESGF)

- To enhance climate science resources, NASA Goddard Space Flight Center has agreed to host a subset of the **primary reanalysis data** on their ESGF node in a similar format to the CMIP5 archive. The ESGF node is managed by the NASA Center for Climate Simulation (NCCS).
- Participating reanalysis institutions:
 - NASA/GSFC/GMAO – MERRA
 - NOAA/NCEP – CFSR
 - ECMWF – Interim
 - JMA/JMA – 25
 - NOAA – 20CR
- Data will need to be translated to the CMIP5 **NetCDF/CF format** through the CMOR application
- <https://oodt.jpl.nasa.gov/wiki/display/CLIMATE/Sharing+Observations+for+Climate+Research>



Climate Science for a Sustainable Energy Future (CSSEF)

- **Organization of model component data and scientific needs through the Earth System Grid Federation (ESGF) Peer-to-Peer (P2P) system**

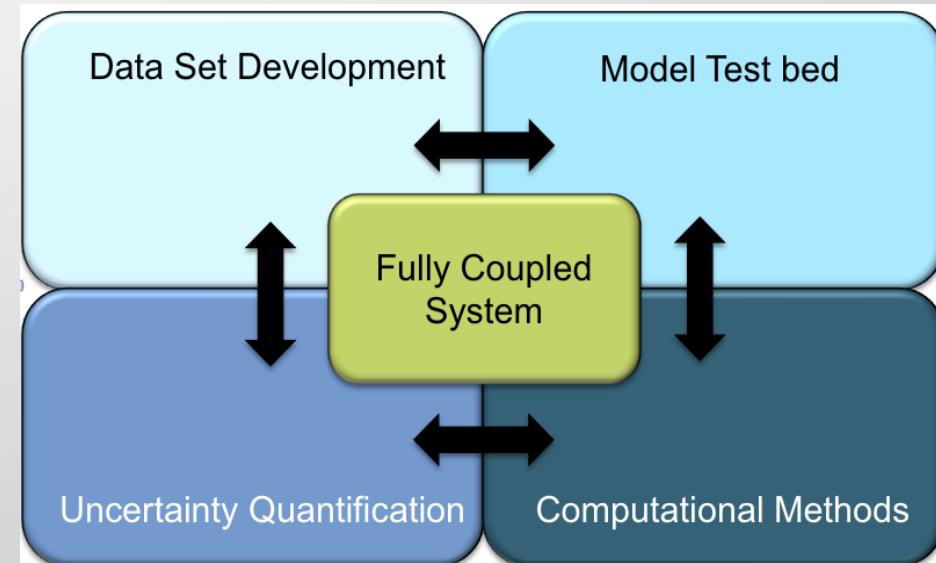
- Manage existing data and add new data to CSSEF archive
- Provenance must be documented and retained
- Stewardship and security
- Discoverability and accessibility

- **New paradigm**

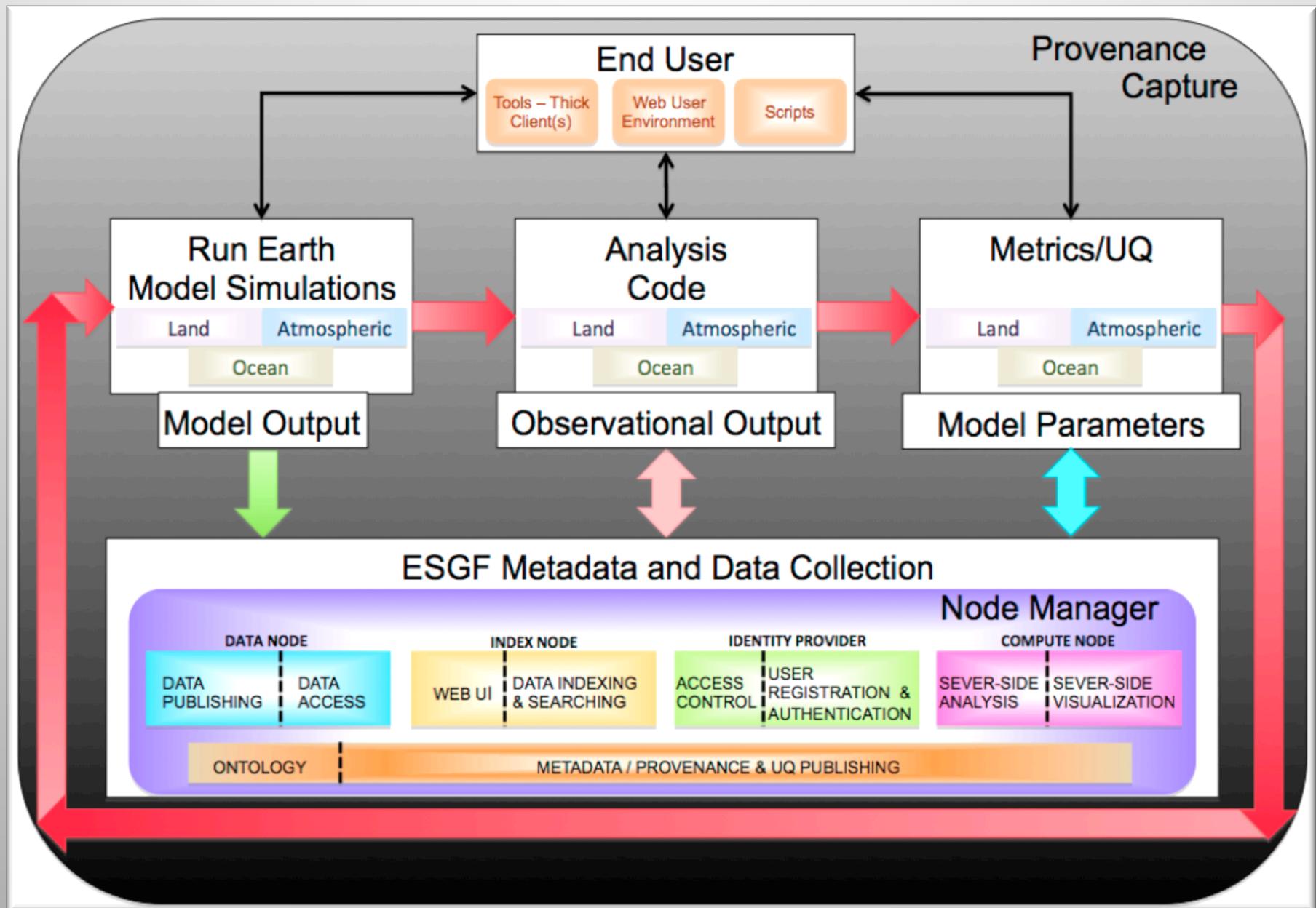
- CAM Integration
- CLM integration
- Diagnostic calculations on CAM and CLM output integration and publishing results into ESGF
- Uncertainty Quantification (UQ) calculations on CAM and CLM integration
- Distribution of workload between clients and “data engines”
- Automation and capture of workflows for reproducibility and efficiency
- Documentation of various steps; record script used for publication; note any problems encountered

- **Current projects**

- Integrate current tools and approaches for test bed and refine for efficiency
- Test and build from initial subsystem



High-level Conceptual View of CSSEF Test Bed Architecture and Workflow



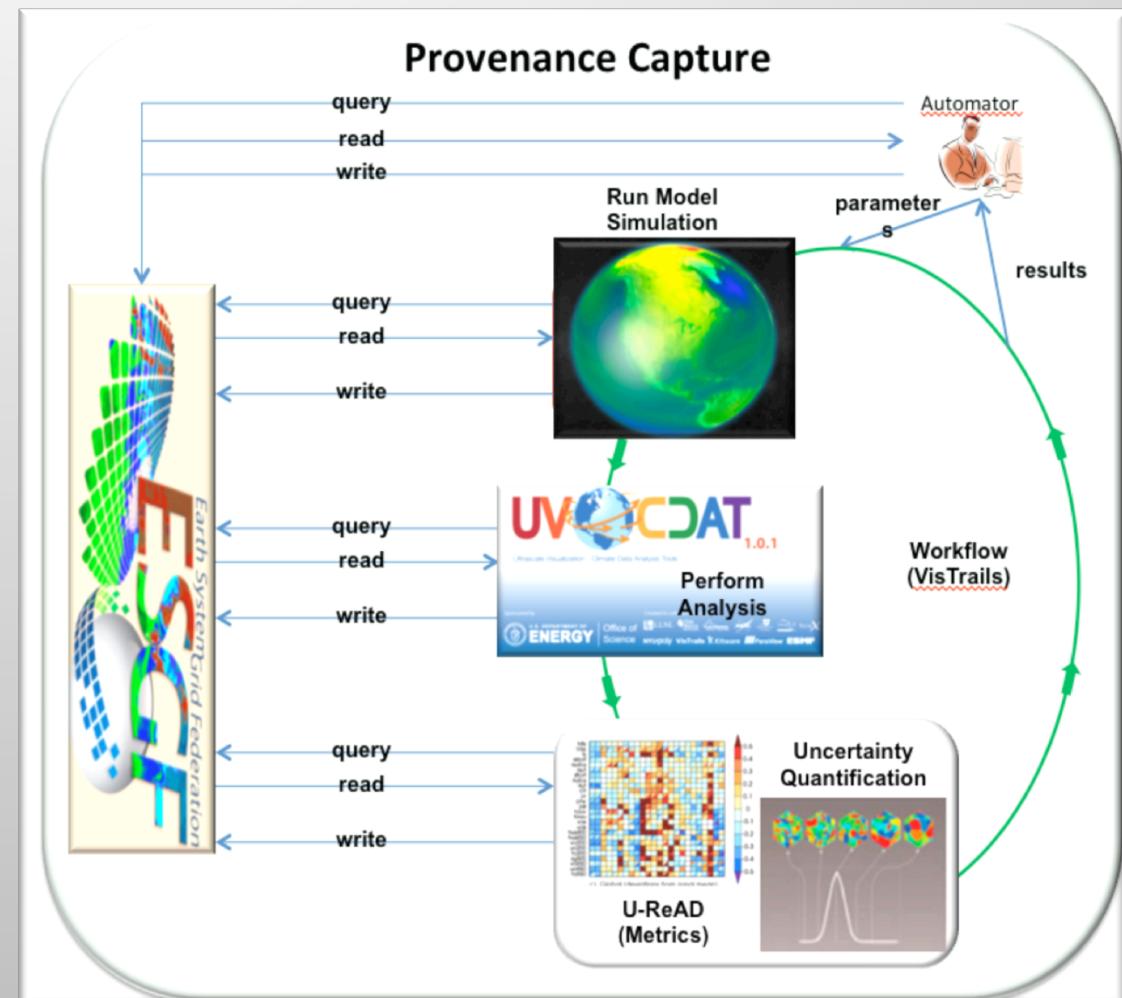
Workflows and Provenance

■ Why Use Workflows in CSSEF?

- Standardize scientific experiments, data processing, and analysis under one unified technical approach
- Workflows will work as building blocks to support more complex problems.
- Can automatically record what processes executed, what data was used, and what data resulted (Provenance)

■ Why is provenance important?

- Without knowing the origin of published data, scientists would have no way to trust the results:
 - What version of the model was used to generate it?
 - What model forcings and parameters were used to produce this result?
 - Real world atmospheric example: LLNL and PNNL both published ARM best estimate data sets. Are they redundant, interdependent or completely independent? The historical record of how they were produced would be the only way to truly differentiate the two efforts.



The Good, the Bad, and the Ugly

- The **data centers** have generally worked together well
 - with limited resources
 - to design and develop a distributed data archive
 - which has minimized the delay in getting data to users
- The monolithic original software structure has been completely **redesigned** by the ESG Federation
 - It's called "peer to peer" design (P2P)
 - Increased modularization will speed improvements
 - Increase flexibility will better meet the needs of a diverse user base
- The **new web-based** user interface substantially
 - Improves speed, accuracy, and flexibility of searches
 - Minimizes clicks
- It is now possible to **script downloads** without first accessing the web interface
 "ESGF P2P release version" is used in production now and running along side the old system until the end of July."
- **Increase speed** of downloads
 - More nodes will establish gridftp servers (~10x faster than http)
 - Data will replicated at major data centers with generally better performance
- **Reduce error rates** for downloads and simplify scripting
- Make available all model and experiment **documentation**
 - Table of forcings for each model & experiment
 - Table of parent experiments and branch times for each realization
 - All information recorded through the METAFOR questionnaire

Near Future

This should happen before the end of summer

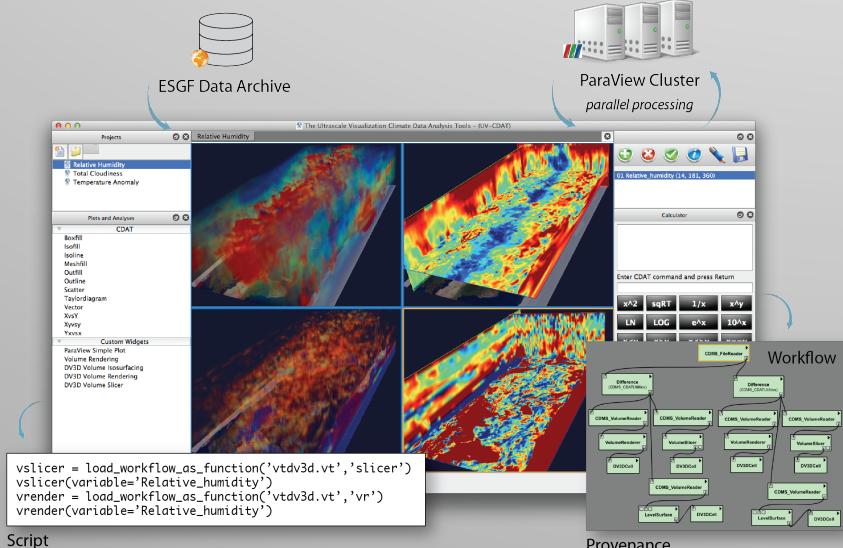
- Data error reports/notification (rudimentary pages now in place)
<http://cmip-pcmdi.llnl.gov/cmip5/errata/cmip5errata.html>
- Service to notify users when new data sets of interest become available.
- Place to record and access CMIP5 publications (web-based form ready now)
- Citable record of CMIP5 provenance (doi assignment and other options)
- Sub-setting of files before download (and other server-side calculations) – better integration with the UV-CDAT is currently underway
- **Obs4MIPs:** An effort to make observational data that can be directly compared with CMIP5 model output
 - Promoted by Duane Waliser, Robert Ferraro, and others at JPL with cooperation from PCMDI and encouragement from the WGCM
 - Data written in same structure and format as CMIP5 model output
 - Data obtainable through ESGF (product=Obs4MIPs)
 - First products from NASA and from ARM now available
 - ESA and NOAA have expressed interest in contributing
 - Wiki describing Obs4MIPs now visible at: <http://obs4mips.llnl.gov:8080/wiki>
- **ANA4MIPs:** A parallel effort is underway to make reanalysis products available (currently NASA MERRA)
 - Promoted by Jerry Potter, and others at Goddard with cooperation from PCMDI and encouragement from the WGCM
 - Data written in same structure and format as CMIP5 model output
 - Data obtainable through ESGF (product=ANA4MIPs)

BER Supported Data Activities

<http://uvcdat.llnl.gov>

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

- **Integrate** DOE's climate modeling and measurements archives
- Develop **infrastructure** for national and international model/data comparisons
- **Deploy** a wide-range of climate data visualization, diagnostic, and analysis tools with familiar interfaces for very large, high resolution climate data sets (**CDAT**, **VTK**, **R**, **VisIt**, **ParaView**, **DV3D**, ...)
- **Workflow** – data flows are directed graphs describing computational tasks
- Takes advantage of **ESGF** data management



Visual Data Exploration & Analysis of Ultra-large Climate Data

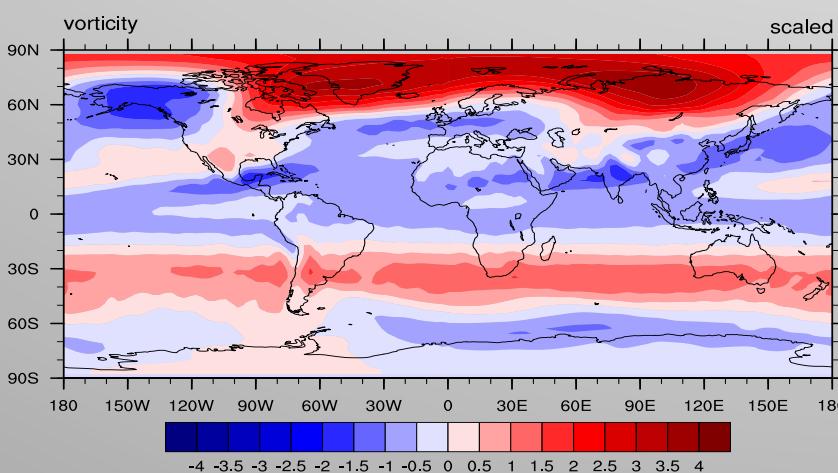
- **Climate applications**, in which specific case studies aimed at answering climate change questions will provide the science drivers for technology development;
- **Technology adaptation**, where applied research is performed to extend existing technologies to meet the needs of the science drivers; and
- **Software engineering and integration** is carried out to deliver working tools to the climate community.



BER Supported Data Activities

Parallel Analysis Tools and New Visualization Techniques for Ultra-Large Climate Data Sets

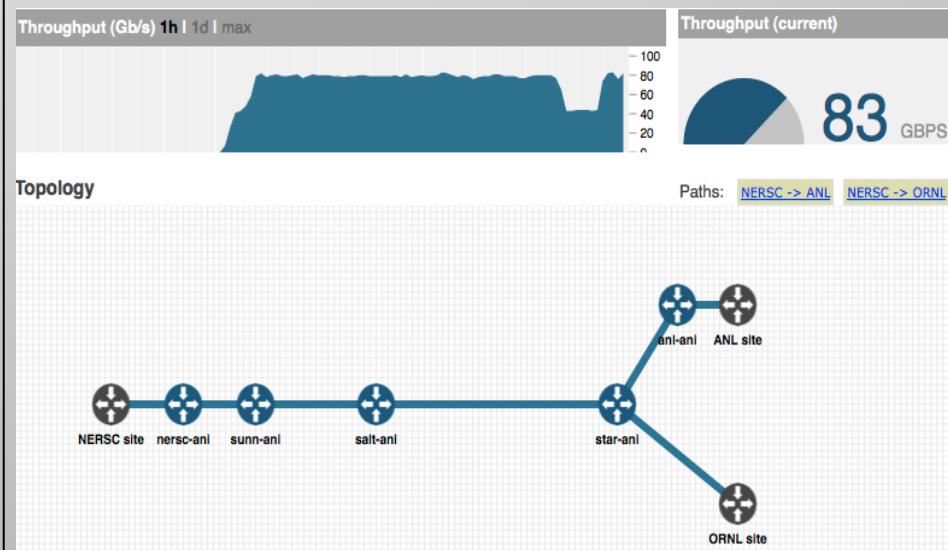
- Speed up current diagnostics (e.g the CESM-CAM atmospheric model diagnostics) with **task parallelism**.
- Create a **data-parallel version** of the NCAR Command Language (NCL) analysis and visualization package.
- Build a new library: ParCAL – Parallel Climate Analysis Library.
- Use existing software technology (MOAB, PnetCDF, Intrepid).
- **ParNCL** (built with ParCAL) will allow users to run their NCL scripts unaltered.
- Explore new ways of doing 3D visualization of climate data



Climate100: Scaling Climate Applications to 100 Gbps Network

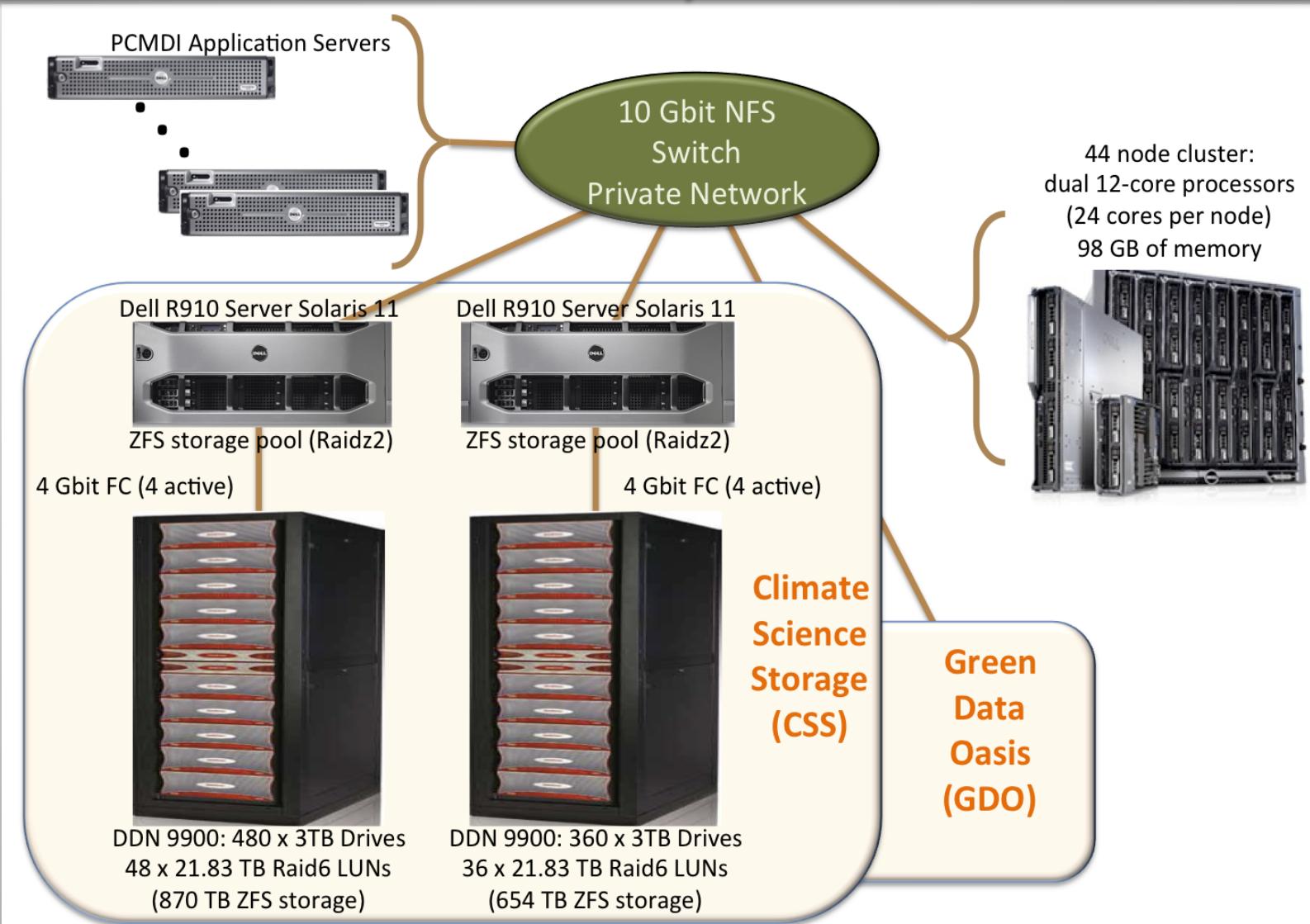
- Total size of data is **increasing**.
- There are **many files**, relatively small files, in climate data sets.
- It requires **efficient methods** to fully utilize the underlying network infrastructure with limited resources.
- **Averaging 83Gbps** on average over TCP connections.
- The 100Gbps network is in the testing phase.
(Expected to be in production by the end of 2012.)

<https://sdm.lbl.gov/climate100>



BER Funded Climate-science Storage System with Scalable Compute Clusters

ESnet 100 Gbps Internet



BER “Climate K-Base”: Bring Together Large Volumes of Diverse Data to Generate New Insights

Data integrating enterprise system

Insight into big data reveals three very significant challenges:

- **Variety:** managing **complex data**, including storage and retrieval, from multiple regional and non-regional data indices, types and schemas
- **Velocity:** distributing live data streams and **large volume data** movement quickly and efficiently
- **Volume:** analyzing large-volume data (from terabytes to zettabytes) in-place for **big data analytics**

BER invests in:

- **Accessing Global Information:** Accessing climate data and content information from everywhere via the web, **sensors**, and applications in an integrated and federated environment
- **Flexible Infrastructure:** Flexible automated administration, easy-to-use analytics, and virtualization at every level
- **Scalable Framework:** Big data analytics in a scalable environment with efficient parallelism, workload-optimization, and **real-time streaming process**

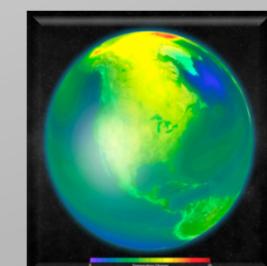
Simulation



Observation



Reanalysis





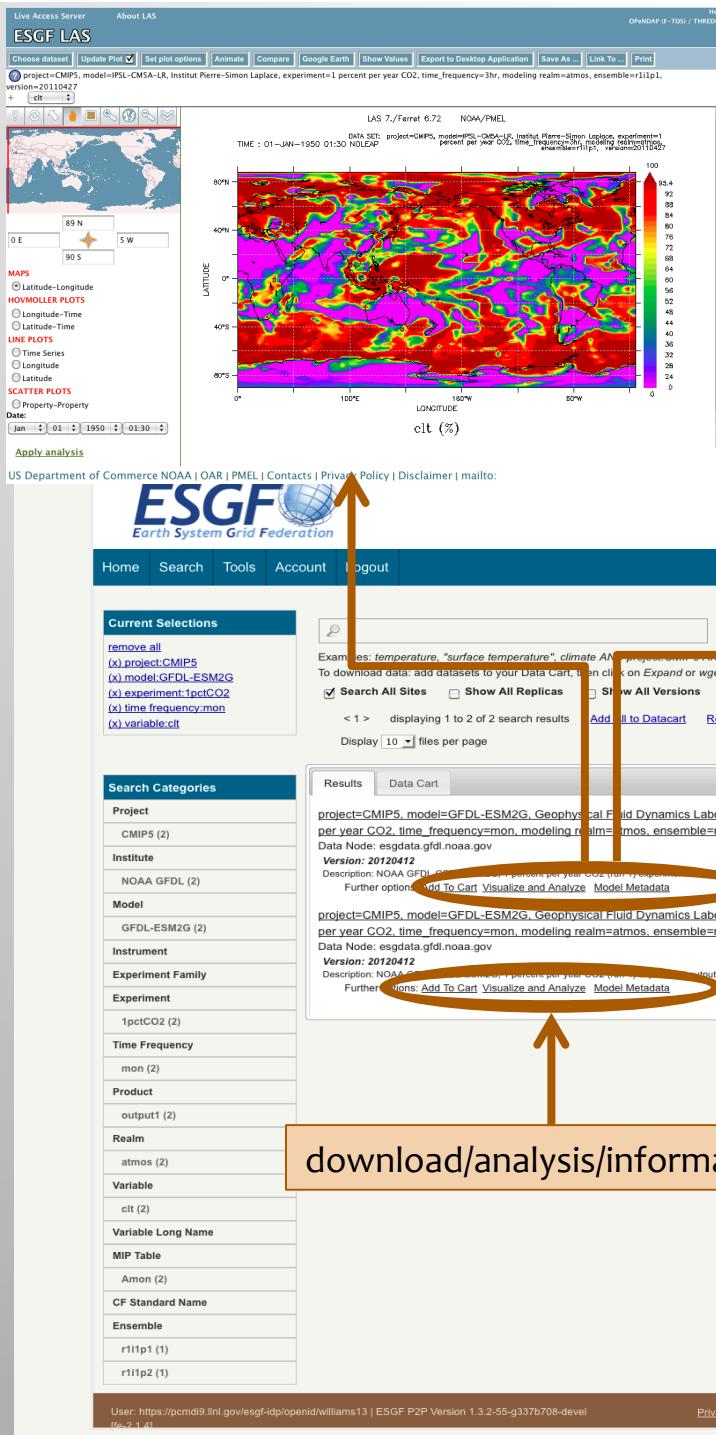
BACKUP SLIDES

ESGF Web Front-End

<http://pcmdi9.llnl.gov>

The screenshot shows the ESGF Portal homepage. At the top, there are tabs for "ESGF Portal", "Climate & Carbon Science", and "UV-CDAT". Below the tabs is a navigation bar with links for "Most Visited", "CSP Authentication", "SharePoint GS-C...", "Getting Started", "Latest Headlines", "Citrix Online", and "Bookmarks". The main content area features the ESGF logo and the PCMDI logo. A large world map is displayed. Below the map, a green banner reads "Welcome to this ESGF P2P Node". A message in red text states: "Welcome to the new CMIP5 distributed archive. Our new ESGF peer-to-peer (P2P) enterprise system is now the official site for CMIP5 model output. The old gateways will remain active and output from all models will continue to be available until the end of July 2012. Please send e-mail to esgf-user@lists.llnl.gov to report bugs and provide feedback." On the left, there is a "Quick Search" section with a keyword input field and a "Search" button. Below it is a link to "Advanced Search (Category, Geospatial, Temporal, and more...)". Under "Peer Nodes", there is a list of nodes with icons: ANL Node, CMCC Node, DKRZ Node, DKRZ CMIP5 Node, IPSL Node, NASA-GSFC Node, NASA-JPL Node, ORNL Node, PCMDI Node, and PNNL Node. At the bottom, there is a note about supported browsers and links to "Guest User | ESGF P2P Version 1.3.2-55-g337b708-devel [fe-2.1.4]" and "Privacy Policy & Legal Notice | Contact ESGF".

The screenshot shows the ESGF search results page. At the top, there are tabs for "ESGF Portal", "Climate & Carbon Science", and "UV-CDAT". Below the tabs is a navigation bar with links for "Most Visited", "CSP Authentication", "SharePoint GS-C...", "Getting Started", "Latest Headlines", "Citrix Online", "ESGF_Transition ...", and "Bookmarks". The main content area features the ESGF logo and the PCMDI logo. On the left, there is a "Current Selections" section with a message: "No search criteria selected". Below it is a "Search Categories" section with a list of categories and their counts: Project (2793), CMIP5 (47748), CORDEX (?), CSSEF (4), GeoMIP (263), LUCID (124), NARCCAP (109), PCM (489), PMIP3 (30), TAMIP (1152), TEST (15), cloud-cryo (10), cordex (57), euclipse (1), geomip (23). To the right of the search categories is a search bar with a "Search" button and a link to "Temporal Search". Below the search bar is a message: "Examples: temperature, \"surface temperature\", climate AND project:CMIP5 AND variable:hus". Underneath the search bar are checkboxes for "Search All Sites", "Show All Replicas", and "Show All Versions", along with buttons for "Add All to Datacart" and "Remove All from Datacart". A "Display" dropdown set to "10 files per page" is also present. At the bottom, there are "Results" and "Data Cart" tabs.



ESGF Web Front-End

This screenshot shows the 'CIM Viewer v0.7.1 - CMIP5 Experiment - 1pctCO2' interface. It has tabs for 'Model' and 'Experiment'. Under 'Experiment', there are sections for 'Overview', 'Requirements', and 'CIM Info'. The 'Overview' section contains details about the 'Short Name' (1pctCO2), 'Long Name' (1 percent per year CO2), 'Description' (Idealized 1pct per year increase in atmospheric CO2 to quadrupling), and 'Rationale' (Diagnostic experiment for understanding the long-term simulations. Evaluate model response without complications of aerosols, land-use change etc.). Below this is a 'Documents : 2' section and a note 'Current Document : CMIP5 Experiment - 1pctCO2'.

This screenshot shows the ESGF web front-end. It features a search bar and a 'Current Selections' sidebar with the same list of selected items as the previous interface. The main area shows search results for 'project:CMIP5', 'model:GFDL-ESM2G', 'experiment:1pctCO2', 'time frequency:mon', and 'variable:cit'. The results list includes entries for 'cmip5.output1.NOAA-GFDL.GFDL-ESM2G.1pctCO2.mon.atmos.Amon.r1i1p1.v20120412' and 'cmip5.output1.NOAA-GFDL.GFDL-ESM2G.1pctCO2.mon.atmos.Amon.r1i1p2.v20120412', each with 'Expand | WGET | Globus Online | Remove' buttons. An orange box labeled 'Download apps' points to these buttons.

ESGF Web Front-End

Geospatial Search

Search Type: Overlaps Encloses

Enter address:

Clear Markers

[1] lat 43.38, lon -87.98
[2] lat 34.59, lon -104.76

Define Area: Square Circle

Submit Geospatial Constraints

Current Selections: (x) overlaps bounding [43.38-104.76,34.59-104.76]

Search Categories: project, model, experiment, frequency, realm, instrument, variable, cf_variable

SURFACE METEOROLOGY DATA: NCDC/FIFE

ESGF Dashboard

ESGF Portal ESGFNodeDesktop UV-CDAT

ESGF Earth System Grid Federation

Dashboard Admin Logout

Elapsed Time

Measure: Number Downloads Dim: Year/Month/Day/Hour Order: Dimension ASC

Filter: All Downloads Source: All Projects Display

ESGF Download Statistics

Dimension

Hosts List [Reference date 06/07/2012 07:59]

Visible Alias Host Name

Last 5 minutes ... Last hour available... Last day available... Last week available... Last month available...

Map data ©2012 MapLink, Tele Atlas - Terms of Use

Geospatial Search

Surface Meteorology Data: NCDC/FIFE

ESGF LAS

Choose dataset Update Plot Set plot options Animate Compare Google Earth Show Values Export to Desktop Application Save As ... Link To ... Print

obs4MIPs NASA-GSFC MODIS L3 Monthly Data

+ clt

LATITUDE

MAPS Latitude-Longitude HOMMOELLER PLOTS Longitude-Time Latitude-Time LINE PLOTS Time Series Longitude Latitude SCATTER PLOTS Property-Property Date: Mar 2000

TIME : 16-MAR-2000 12:00

DATA SET: obs4MIPs NASA-GSFC MODIS L3 Monthly Data

LAS 7./Ferret 6.71 NOAA/PMEL

clt (%)

Latitude: 88.5 N, 0.5 E, 89.5 S; Longitude: 1.5 W

Latitude: 80°N, 40°N, 0°, 40°S, 80°S; Longitude: 50°E, 150°E, 110°W, 10°W

Color scale: 0.05 to 99.99

Server-side Analysis and Visualization

US Department of Commerce NOAA | OAR | PMEL | Contacts | Privacy Policy | Disclaimer | mailto:

ESGF Portal ESGFNodeDesktop UV-CDAT

ESGF Earth System Grid Federation

Dashboard Admin Logout

Elapsed Time

Measure: Number Downloads (x300) Dim: Experiment Order: Dimension ASC

Filter: Successful downloads Source: CNPS Display

Hosts List [Reference date 07/12/2012 07:42]

Visible Alias Host Name

Last 5 minutes ... Last hour available... Last day available... Last week available... Last month available...

Peer Group Maps

Globe Online Setup for Use in the Earth System Grid Federation

Using Globe Online in the Earth System Grid Federation

The ESG POP will request the use of Globe Online to download data. The Globe Online option can be used to download data to the user's machine or to another server.

The following are instructions for the ESG POP code to get data with Globe Online.

Setup Globe Online Account

This page is intended for users who want to use the Globe Online account. This is a one-time setup that only needs to be completed once.

1. Make sure you have an account already. If you do not have an account already, click "Create New Account".
2. Click "My Profile" option from the page and click drop down menu in the bar.
3. Click "Edit Profile" button.
4. The next step would be to determine what your ESG POP key is to be used to link your Globe Online account. You can find this information in the "Edit Profile" section.

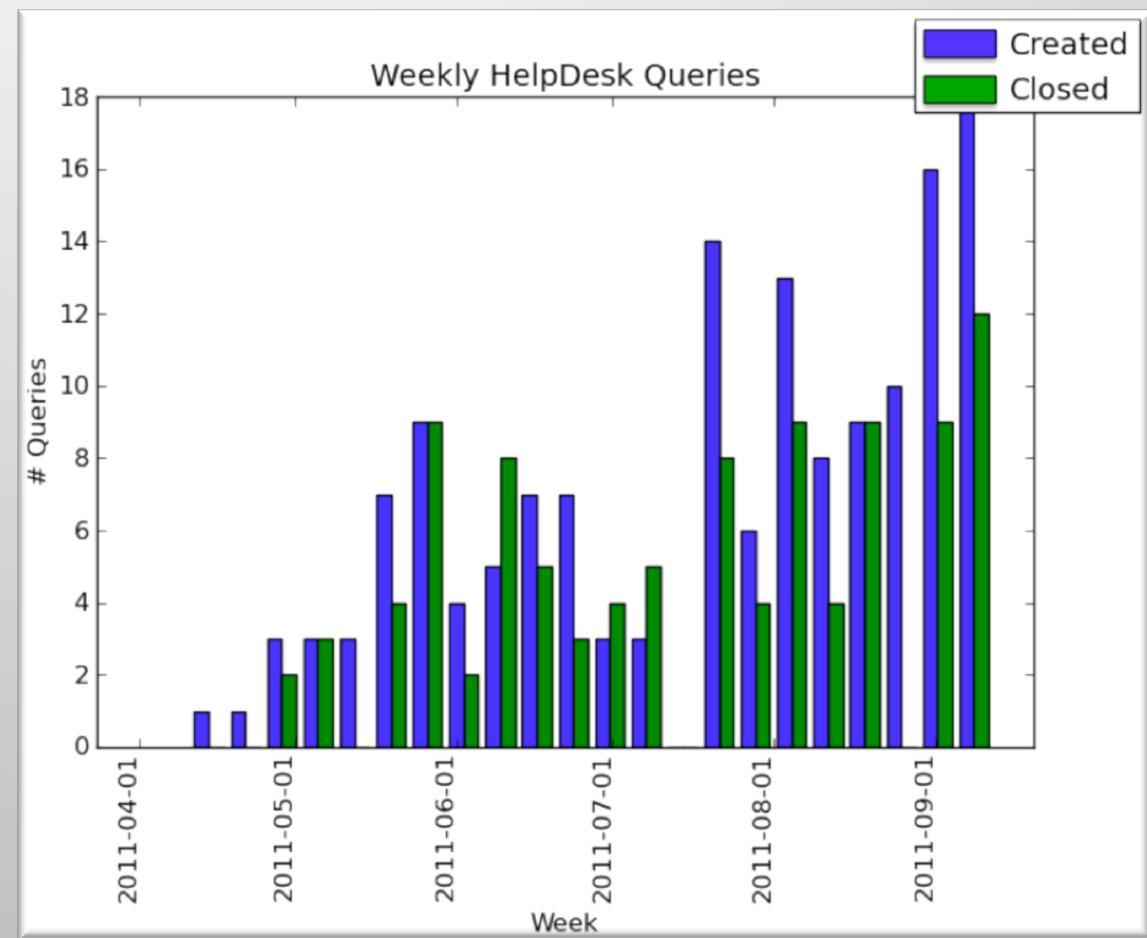
Connecting Globe Online and ESG identities

NOTE: This page is intended for an expert level of difficulty. We are working on the Globe Online team to ways to make these steps for future ease of use. For now, please look at us and contact Globe@mitre.org if you question anything.

YouTube video player

ESGF Help Desk Weekly Traffic

- About half of these queries come directly to **cmip5-helpdesk@stfc.ac.uk** and the other half to the esg-support mailing list **esgf-user@lists.llnl.gov**
- Questions that are resolved are place on the **ESGF FAQ** list
- **Scientists** at PCMDI/LLNL, BADC, and DKRZ are charged with addressing CMIP5 data questions
- **Technical staff** at LLNL, BADC, DKRZ, NCAR, and JPL are charged with addressing ESGF system questions



Publication Web Application

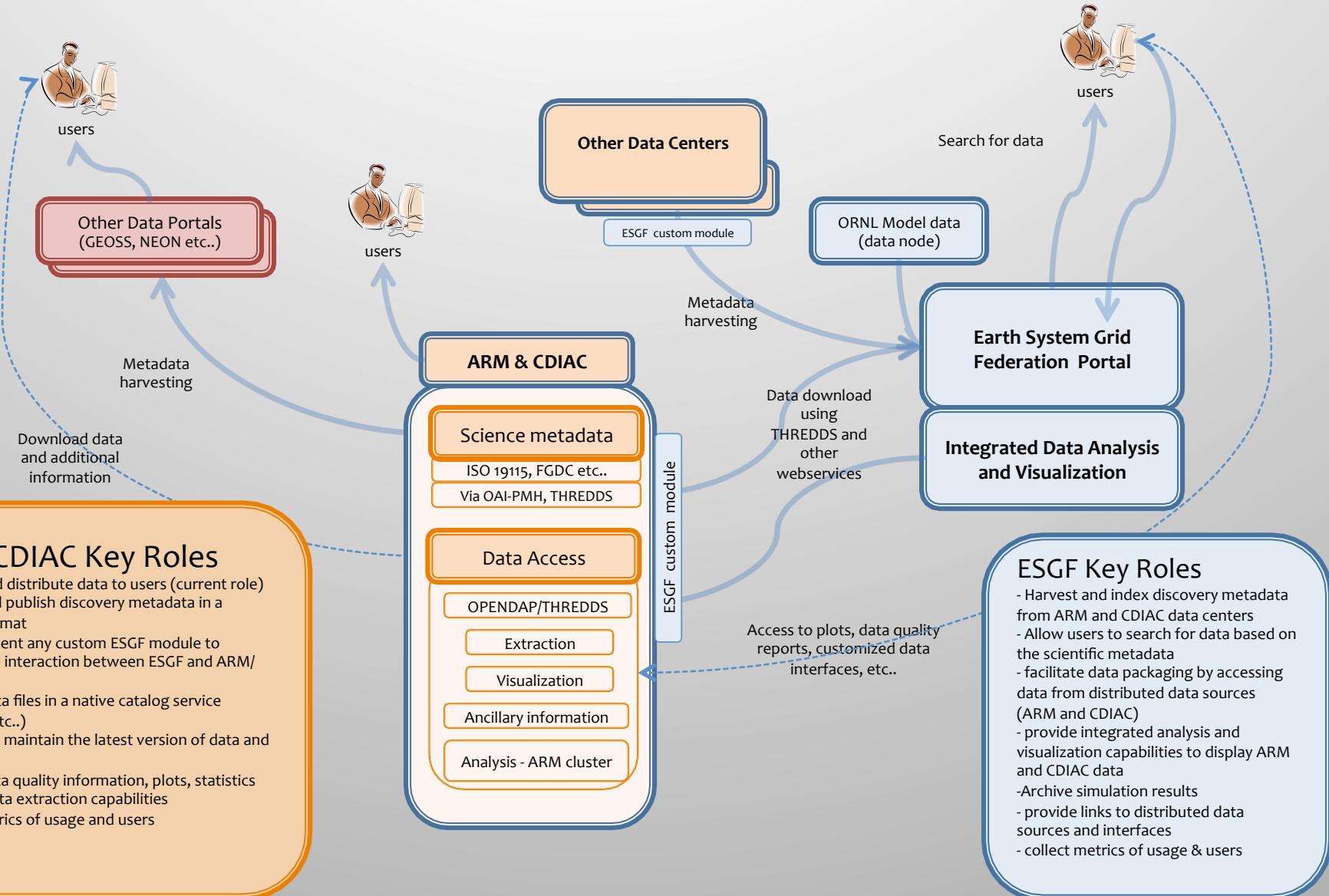
- **Web based submission** of citations to database
- **Capture tags** to enable contextual search, the tags include
 - Keywords
 - Funding
 - Data used (model, experiment, variable, time frequency)
- Submission page differentiate with the **journal type**
 - The types are: journal, book, proceeding, presentation, technical report
- **Editable database** entry
 - Enable search of all submitted publications, update, change or add keywords
- Admin page for **editing** or removing publications

Submit Pages

List

Search

ESGF – ARM and CDIAC Publication Architecture



ARM/CDIAC Key Roles

- Archive and distribute data to users (current role)
- Prepare and publish discovery metadata in a standard format
- will implement any custom ESGF module to facilitate the interaction between ESGF and ARM/CDIAC
- Publish data files in a native catalog service (THREDDS etc..)
- update and maintain the latest version of data and metadata
- provide data quality information, plots, statistics
- provide data extraction capabilities
- collect metrics of usage and users

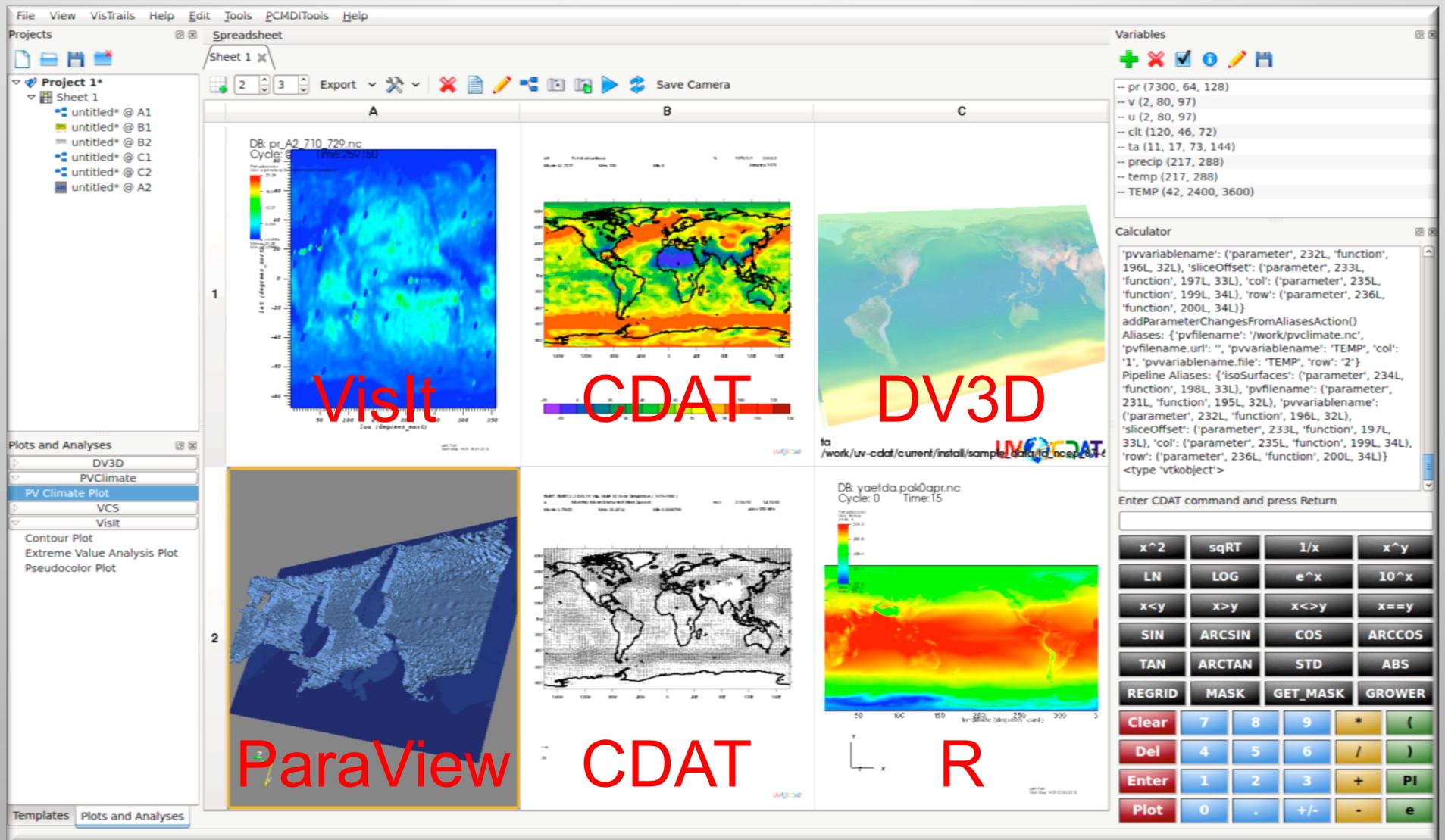
ESGF Key Roles

- Harvest and index discovery metadata from ARM and CDIAC data centers
- Allow users to search for data based on the scientific metadata
- facilitate data packaging by accessing data from distributed data sources (ARM and CDIAC)
- provide integrated analysis and visualization capabilities to display ARM and CDIAC data
- Archive simulation results
- provide links to distributed data sources and interfaces
- collect metrics of usage & users

Ultra-large Climate Data Analysis and Visualization

UV-CDAT Displaying CDAT, DV3D, ParaView, VisIt, and R

<http://uvcdat.llnl.gov>



Layered Data Distributed Open Architecture

