

PCMDI Data Management and the Earth System Grid

PCMDI Software Team

PCMDI Advisory Committee Meeting

Livermore, California
April 8, 2008



LLNL-PRES-402704

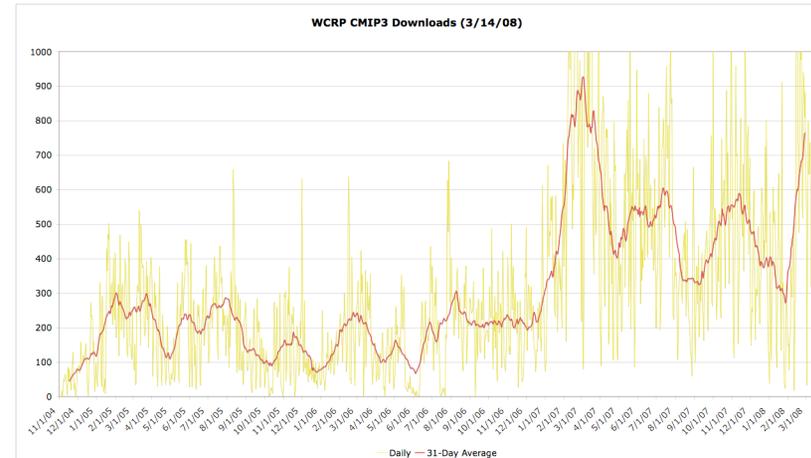


Outline

- **PCMDI data management**
- **ESG-CET**
- **ESG-CET from user and publisher perspectives**
- **Standards**
- **Collaborations**

PCMDI leadership in climate data management

- PCMDI has led the climate community in organizing and providing access to model intercomparison data
- The CMIP3/AR4 has had a significant impact
 - Currently archiving 35TB from 25 models
 - 300+ peer-reviewed publications
 - Demand for data has increased since release of the IPCC WG1 SPM (Feb. '07)
 - Data downloads of 400-600GB/day



“A major advance of this assessment of climate change projections compared with the TAR is the large number of simulations available from a broader range of models. Taken together with additional information from observations, these provide a quantitative basis for estimating likelihoods for many aspects of future climate change.”

IPCC WG1 Summary for Policymakers

- New archives are emerging:
 - Cloud Feedback Model Intercomparison Project (CFMIP)
 - North American Regional Climate Change Assessment Program (NARCCAP)
 - CCSM Carbon-Land Model Intercomparison Project (C-LAMP)
- Current software based on Earth System Grid-II

Experience from CMIP3, ESG-II

- **Standardization is key:**
 - ‘Homogenization of data’: transposition of data from raw history format to one variable per file;
 - Consistent format with standard APIs (netCDF)
 - Consistent metadata: CF-1
- Targeting a specific community (WG1) helped to bound the problem.
- Provide **multiple paths to the data.**
 - ESG Web portal for search and discovery
 - FTP for bulk downloads, scriptable with wget, ease of client access, familiarity
 - OPeNDAP allows subsetting, richer client access
- Registering users gives important information on data usage.
 - Helps demonstrate the value of the service to our sponsors.
- **Model data is not static.**
 - >100 entries on the CMIP3 errata pages
- Goal: Retain what worked in the next iteration: ESG-CET

Earth System Grid Center for Enabling Technologies (ESG-CET)



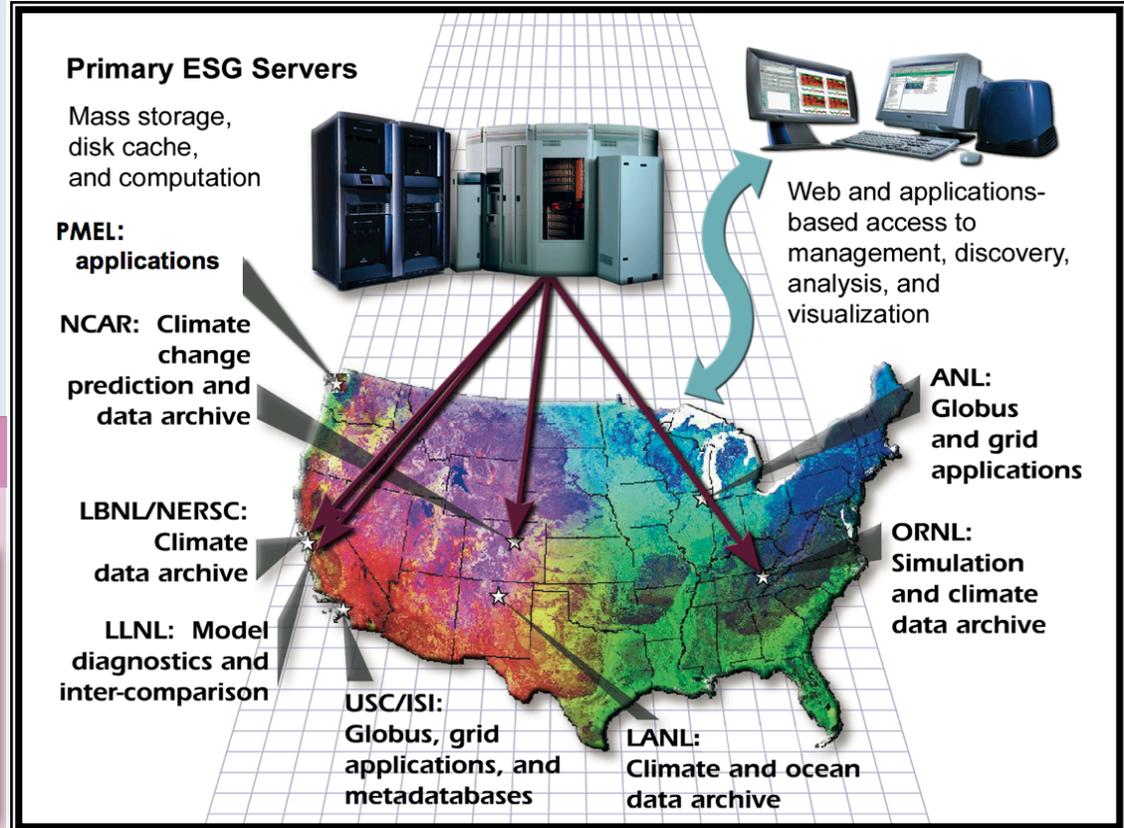
ESG Goals

- Petabyte-scale data volumes
- Federation of sites
- “Virtual Datasets” created through subsetting and aggregation
- Metadata-based search and discovery
- Bulk data access
- Web-based and analysis tool access
- Increased flexibility and robustness
- PI: Dean Williams (PCMDI)

<http://www-pcmdi.llnl.gov>



Current ESG Sites



CMIP5/AR5 requirements are driving ESG-CET

AR4

- **Data Volume**
 - 10s of terabytes (10^{12} bytes)
 - Downloads: ~500GB/day
- **Models**
 - 25 models
- **Metadata**
 - CF-1 + IPCC-specific
- **User Community**
 - Thousands of users
 - WG1, domain knowledge

AR5

- **Data Volume**
 - 1-10 petabytes (10^{15} bytes)
 - Downloads: 10s of TB/day
- **Models**
 - ~35 models
 - Increased resolution
 - More experiments
 - Increased complexity (ex: biogeochemistry)
- **Metadata**
 - CF-1 + IPCC-specific
 - Richer set of search criteria
 - Model configuration
 - Grid specification from CF (support for native grids)
- **User Community**
 - 10s of thousands of users
 - Wider range of user groups will require better descriptions of data, attention to ease-of-use

Improvements over ESG-II

- **'Faceted' search** capability guides the user toward datasets of interest
 - At a given point in the search, only those options which produce non-empty result sets are shown
 - Avoids 'deadend' searches
 - Flexible browsing hierarchy
- Automated, GUI-based publication tools
- **Single sign-on**
- 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.
- **Client access** to subsetting, visualization services.
- Server-generated visualization products
- Fine-grained access to datasets based on user groups and roles.
- User notification service
 - Users can choose to be notified when a dataset has been modified.
- **Pre-computed products** (e.g., global averages)

ESG-CET timeline

- **2008: Design and implement core functionality:**
 - Browse and search
 - Registration
 - Single-sign on / security
 - Publication
 - Distributed metadata
 - Server-side processing
- **Testbed in early 2009**
 - Plan to include at least seven centers in the US, Europe, and Japan:
PCMDI, NCAR, GFDL, ORNL, BADC, MPI, CCSR
- **PCMDI will archive a core set of runs, as in AR4**
 - Will accept data from AR5 centers that choose not to join ESG
- **2009: Deal with system integration issues, develop production system.**
- **2010: Modeling centers publish data.**
- **2011: Research completed and journal articles submitted.**

Gateways and nodes

- Federated architecture

Federation is a virtual trust relationship among independent management domains that have their own set of services. Users authenticate once to gain access to data across multiple systems and organizations.

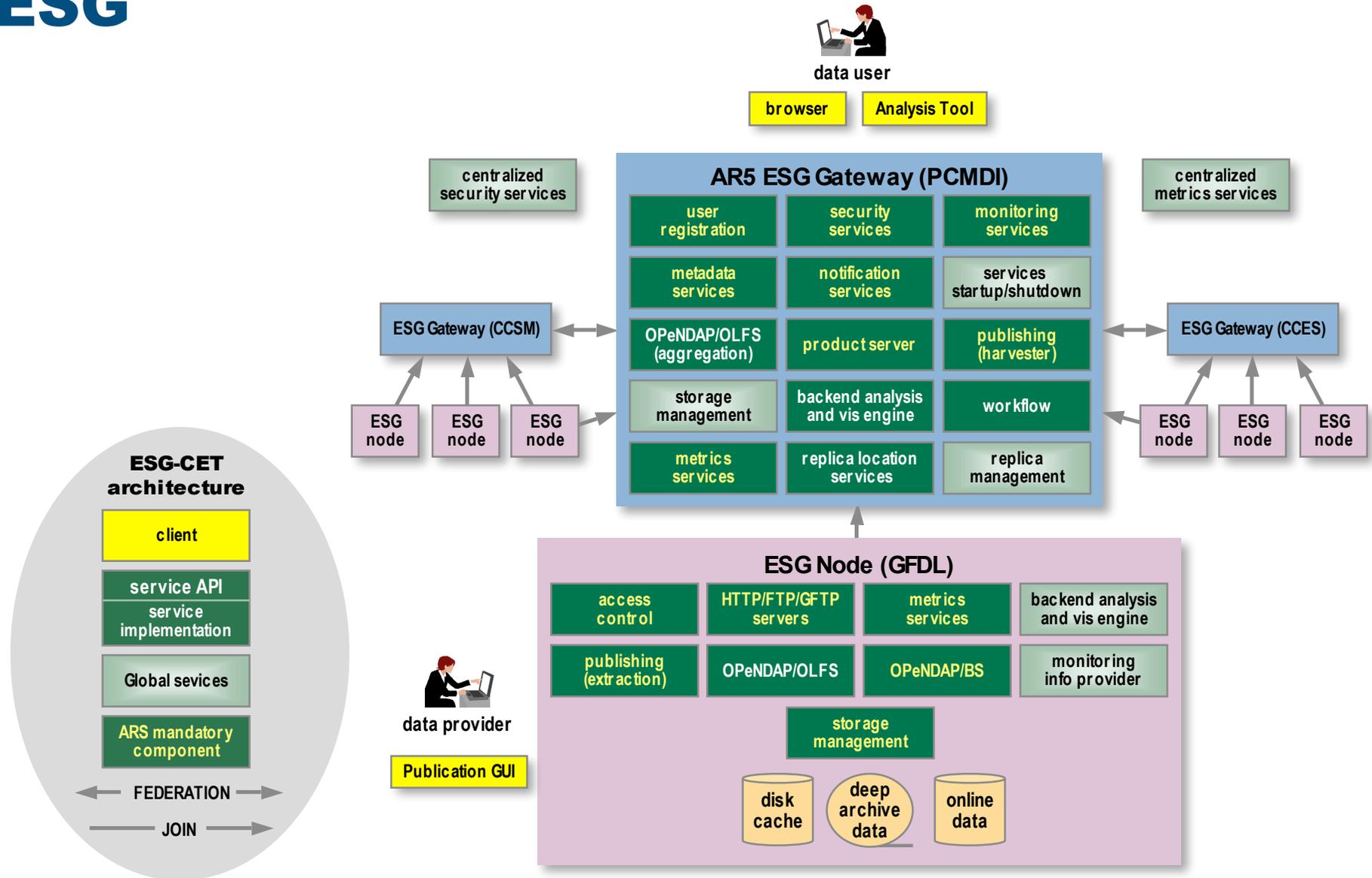
- Gateways

- Portals, search capability, distributed metadata, registration and user management
- More complex architecture than nodes, fewer sites
- Initially PCMDI, NCAR, ORNL, eventually GFDL
- May be customized to an institution's requirements

- Nodes

- Where data is stored and published
- Data may be on disk or tertiary mass store.
- Each node has a trust relationship with a specific gateway, for publication.
- Less complex architecture
- A site can be both a gateway and a node.

Architecture of the next generation of ESG



ESG-CET portal: a walkthrough from the user perspective

The screenshot shows the ESG-CET Gateway website. At the top, there is a navigation bar with 'Home', 'Data', 'About ESG', and 'Login' links. A 'Single sign-on login' button is highlighted in a green box. Below the navigation bar, a 'Welcome to ESG-CET' message is displayed. The main content area features a section titled 'The Earth System Grid' with a map of the United States and a text description. A green box annotation points to this section, stating 'Portal may be customized for a particular institution.' Below this, there are several functional areas: 'ESG Data Gateways' with links to NCAR, Climate End Station, and IPCC; 'Quick Links' for account creation and data search; a 'Quick Data Search' box with a search input and 'Go' button; a 'Browse By Project' section listing CCSM, NARCCAP, and PCM models; and a 'Spotlight: CCSM-3 Model' section featuring a temperature change map and links for more information and dataset download. A light blue box annotation points to the CCSM-3 Model section, stating 'AR5 datasets will be visible from all portals.' At the bottom, the user is identified as 'guest' and the page includes copyright information for UCAR.

ESG-CET Gateway

Earth System Grid

Home Data About ESG Login

Single sign-on login

Welcome to ESG-CET

The Earth System Grid

The Earth System Grid (ESG) integrates supercomputers with large-scale data and analysis servers located at numerous national labs and research centers to create a powerful environment for next generation climate research. Access to ESG is provided through a system of federated Data Gateways, that collectively allow access to massive data sets from Climate Global and Regional Models, IPCC research, and analysis and visualization software. The Earth System Grid - Center for Enabling Technologies (ESG-CET) is funded by the U.S. Department of Energy as part of the SciDAC (Scientific Discovery through Advanced Computing) program. [Read More](#)

Portal may be customized for a particular institution.

ESG Data Gateways

- NCAR Gateway
- Climate End Station
- IPCC Gateway

Quick Links

- Create Account
- Browse Catalogs
- Search for Data
- Visualize Data

Quick Data Search

Go

Power Search

Browse By Project Browse By Experiment

- CCSM : (Collections: 2)
Community Climate System Model
- NARCCAP : (Collections: 0)
North American Regional Climate Change Assessment Program
- PCM : (Collections: 1)
Parallel Climate Model

Spotlight: CCSM-3 Model

Surface temperature change relative to 1870-1899 baseline. CCSM3 PCL A1B

AR5 datasets will be visible from all portals.

The graphic depicts the surface temperature increase (relative to the 1870-1899 period) from the average of a set of CCSM3 experiments of the IPCC AR4 SRES A1B (midrange) climate change scenario. [Learn More](#)
[Download These Datasets](#)

User: guest | [ESG Home](#) | [Contact Us](#)
Gateway Portal Software version 0.2 © UCAR, all rights reserved.

Power search



Advanced Search

Instructions: Use the categories below to create a search hierarchy. Search results will be filtered from left to right by each option selected within each category. Rearrange the categories to change the search hierarchy.

Search Categories

Model Experiment CF Standard Name Time Frequency Data Format Domain Grid

User selects search categories of interest and orders them.

Selected Option:

<p>Model</p> <p>Model > PCM 1</p>		
--------------------------------------	--	--

Free Text

User: guest | [ESG Home](#) | [Contact Us](#)
Gateway Portal Software version 0.2 © UCAR, all rights reserved.

Search is 'context sensitive'



Advanced Search

Instructions: Use the categories below to create a search hierarchy. Search results will be filtered from left to right by each option selected within each category. Rearrange the categories to change the search hierarchy.

Search Categories

Model Experiment **CF Standard Name** Time Frequency Data Format Domain Grid

Selected Option:

Model

- Model
- Model > PCM 1

Experiment

- IPCC AR4
- IPCC AR4 > IPCC AR4 20C3M_bc

CF Standard Name

- air temperature lapse rate
- atmosphere mass per unit area
- barotropic northward sea water ve
- change in atmosphere energy con
- cloud liquid water content of atm
- direction_of_swell_wave_velocity
- downwelling longwave flux in air
- freezing temperature of sea water
- lwe convective snowfall rate

At each step, only relevant options are displayed.

Free Text

Reset All Categories Reset All Options Submit Query

Total Number of Results: 1

1-1 of 1 results

- [PCM testsim_1782904354 Land History Files Data \(Daily\)](#)
Description: PCM testsim_1782904354 Land History Files Data (Daily) for simulation PCM testsim_1782904354
Authorization: Guest Users
Access: [Data Visualization \(LAS\)](#)

Search returns:
- datasets
- available products for each dataset

User: guest | [ESG Home](#) | [Contact Us](#)
Gateway Portal Software version 0.2 © UCAR, all rights reserved.

File download



Collection Browsing

Projects >> PCM >> PCM (Parallel Climate Model) run B04.10 >> PCM run B04.10 data organized by time (original model output) >> *PCM run B04.10 atmosphere data*

Summary Variables **Files** Administration Metrics

Filename Pattern: * File Count Limit (max 1000): 100 [List Files](#) [Reset File Search Criteria](#)

Selected Files: [Generate WGET script](#) [Add to Data Cart](#)

<< 1 >> 25

<input type="checkbox"/>	File Name	Size	Download
<input type="checkbox"/>	B04.10.atm.0049.nc	274 MB	Direct Download
<input type="checkbox"/>	B04.10.atm.0050.nc	274 MB	Direct Download
<input type="checkbox"/>	B04.10.atm.0051.nc	274 MB	Direct Download
<input type="checkbox"/>	B04.10.atm.0052.nc	274 MB	Direct Download

<< 1 >> 25

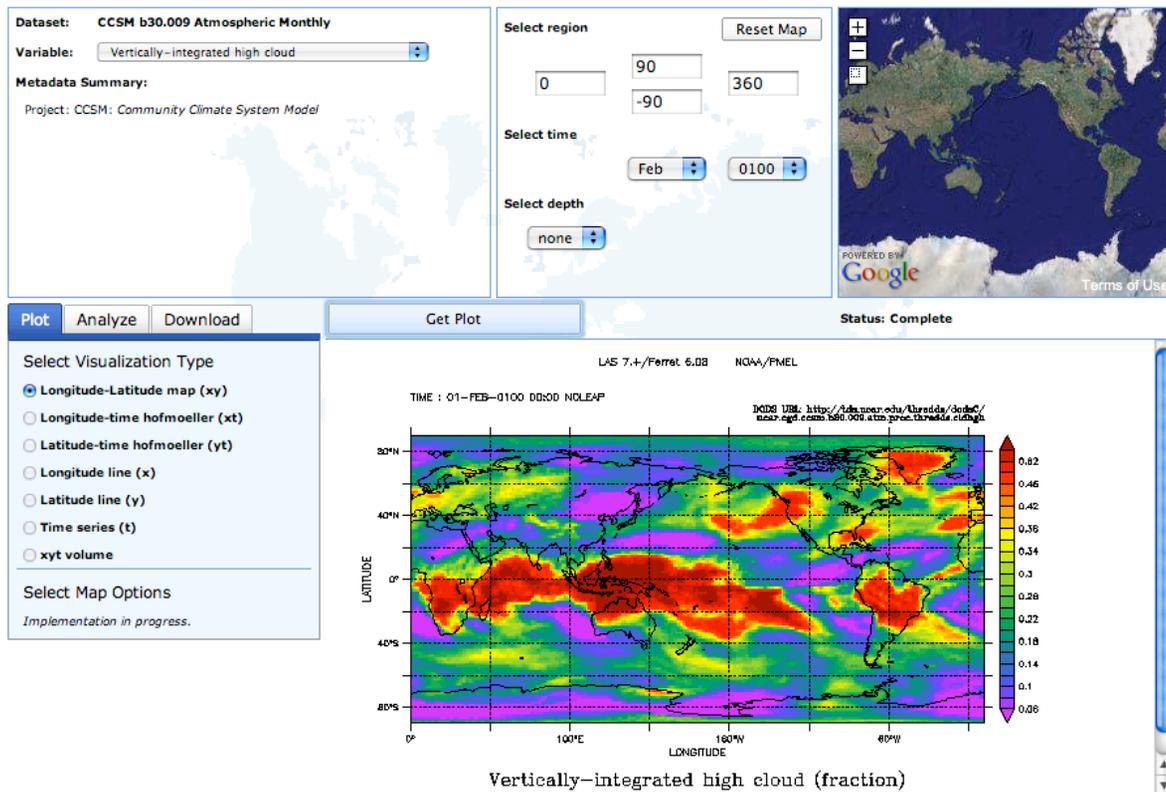
To download files:
- direct download for online datasets
- scripts generated for wget, bulk download

User: testUser | [ESG Home](#) | [Contact Us](#)
Gateway Portal Software version 0.2 © UCAR, all rights reserved.

Server-generated products



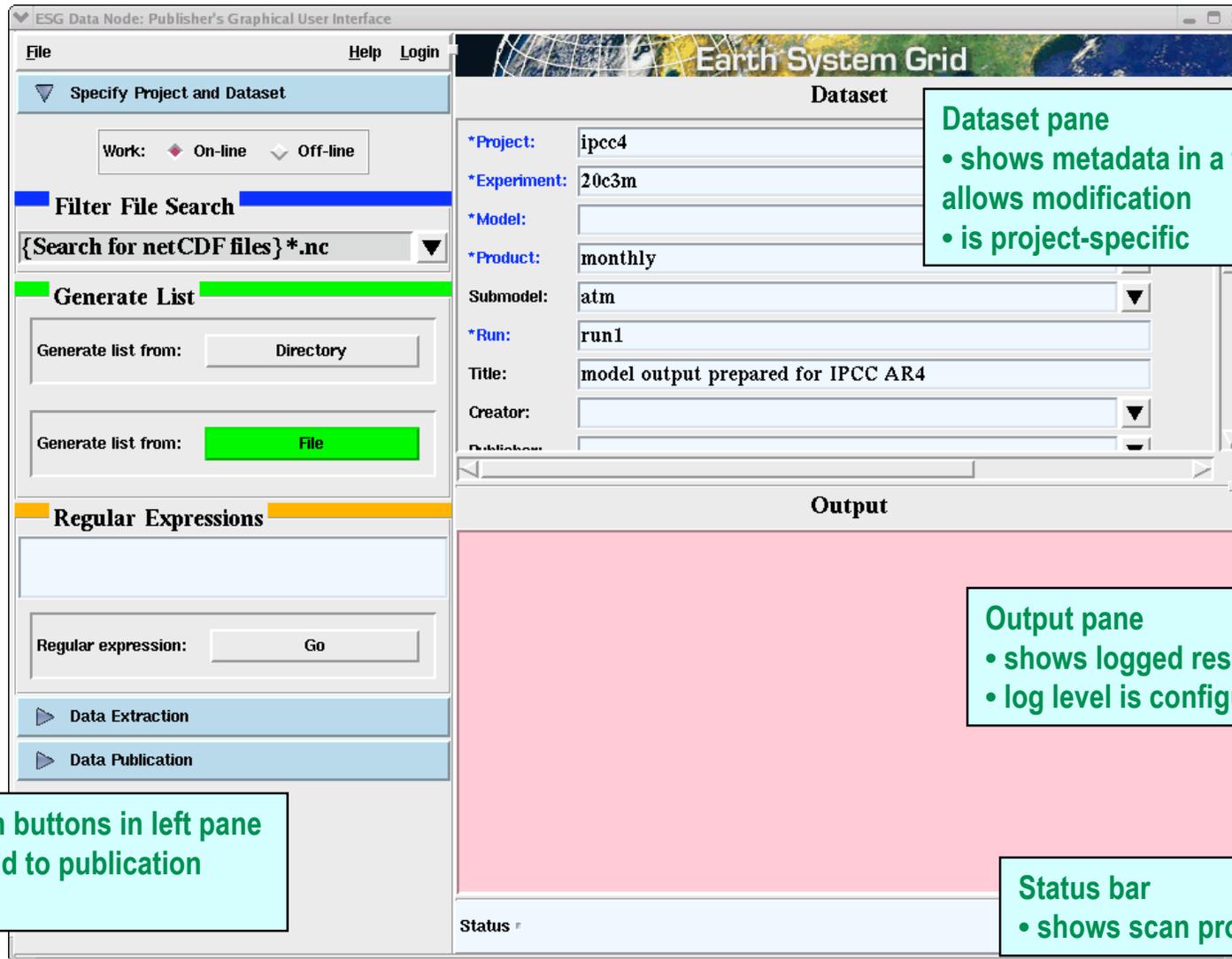
Visualization Prototype - ESG



User: testUser | [ESG Home](#) | [Contact Us](#)

Gateway Portal Software version 0.2 © UCAR, all rights reserved.

Dataset publishing on an ESG node: Metadata specification



Dataset pane

- shows metadata in a file,
- allows modification
- is project-specific

Output pane

- shows logged results
- log level is configurable

Expansion buttons in left pane correspond to publication steps.

Status bar

- shows scan progress

Data scan

The screenshot displays the 'ESG Data Node: Publisher's Graphical User Interface'. The main window is titled 'Earth System Grid' and 'Dataset'. The left sidebar contains navigation options: 'Specify Project and Dataset', 'Data Extraction', 'Data Publication', and 'Button Controls'. The 'Data Extraction' section has two buttons: 'Create/Replace' and 'Append/Update'. The 'Dataset' section contains the following fields:

- *Project: ipcc4
- *Experiment: 20c3m
- *Model: ncar_ccsm3_0
- *Product: monthly
- Submodel: atm
- *Run: run1
- Title: model output prepared for IPCC AR4
- Creator: [empty]
- Publication: [empty]

The 'Output' section shows a log of scan progress:

```
.atm.1920-01_cat.1929-12.nc
INFO 2008-03-26 14:28:27,256 Scanning /ipcc/20c3m/atm/mo/clwvi/ncar_ccsm3_0/run1/clwvi_A1.20C3M_1.CCSM
.atm.1930-01_cat.1939-12.nc
INFO 2008-03-26 14:28:27,282 Scanning /ipcc/20c3m/atm/mo/clwvi/ncar_ccsm3_0/run1/clwvi_A1.20C3M_1.CCSM
.atm.1940-01_cat.1949-12.nc
INFO 2008-03-26 14:28:27,304 Scanning /ipcc/20c3m/atm/mo/clwvi/ncar_ccsm3_0/run1/clwvi_A1.20C3M_1.CCSM
.atm.1950-01_cat.1959-12.nc
INFO 2008-03-26 14:28:27,327 Scanning /ipcc/20c3m/atm/mo/clwvi/ncar_ccsm3_0/run1/clwvi_A1.20C3M_1.CCSM
.atm.1960-01_cat.1969-12.nc
INFO 2008-03-26 14:28:27,348 Scanning /ipcc/20c3m/atm/mo/clwvi/ncar_ccsm3_0/run1/clwvi_A1.20C3M_1.CCSM
.atm.1970-01_cat.1979-12.nc
INFO 2008-03-26 14:28:27,371 Scanning /ipcc/20c3m/atm/mo/clwvi/ncar_ccsm3_0/run1/clwvi_A1.20C3M_1.CCSM
.atm.1980-01_cat.1989-12.nc
INFO 2008-03-26 14:28:27,394 Scanning /ipcc/20c3m/atm/mo/clwvi/ncar_ccsm3_0/run1/clwvi_A1.20C3M_1.CCSM
.atm.1990-01_cat.1999-12.nc
INFO 2008-03-26 14:28:27,415 Adding file info to database
INFO 2008-03-26 14:28:31,921 Aggregating variables
INFO 2008-03-26 14:28:35,299 Creating dimensions
INFO 2008-03-26 14:28:35,343 Setting aggregate dimension ra
INFO 2008-03-26 14:28:35,431 Adding variable info to databa
```

The status bar at the bottom indicates 'Status 100.00 %'.

1. Dataset is created or updated based on input metadata.

Selecting an extraction option starts the dataset scan

2. Files are scanned and internal database tables populated.

3. If an 'aggregation dimension' is specified, variables are aggregated.

Data aggregation and publication

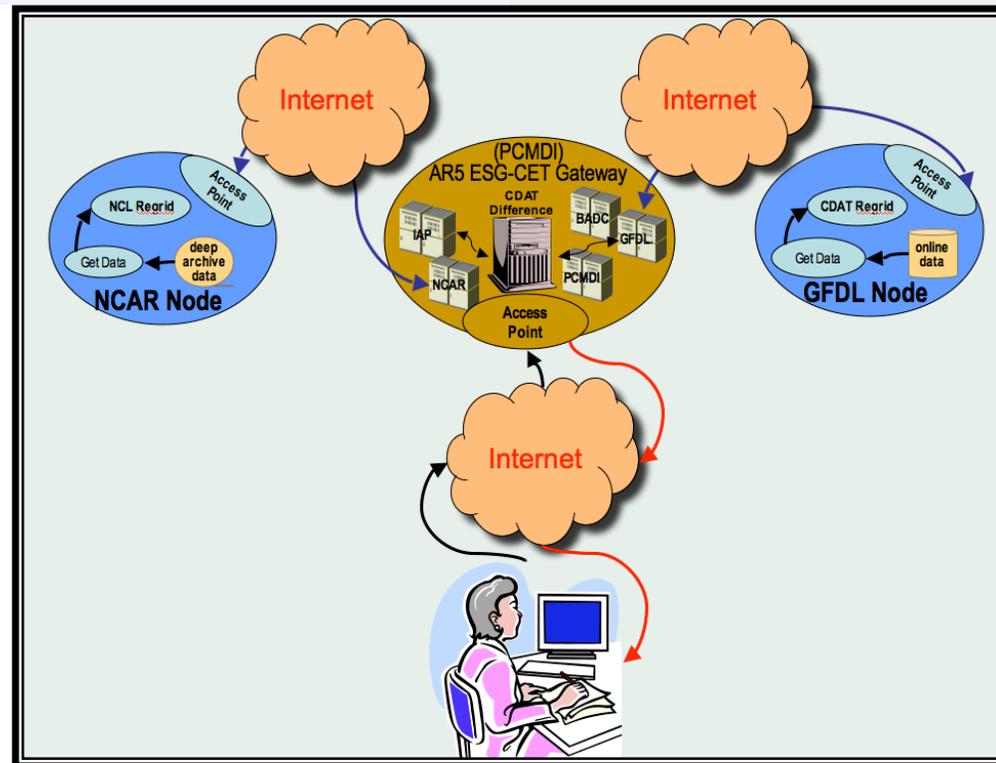
The screenshot displays the 'ESG Data Node: Publisher's Graphical User Interface'. The main window is titled 'Earth System Grid' and features a 'Dataset' configuration panel on the right and a 'Button Controls' panel on the left. The 'Dataset' panel includes fields for Project (ipcc4), Experiment (20c3m), Model (near_ccsm3_0), Product (monthly), Submodel (atm), Run (run1), Title (model output prepared for IPCC AR4), and Creator. The 'Button Controls' panel has a 'Release data:' section with a 'Publish' button and a 'Generate:' section with a 'THREDDS' button. An 'Authentication Required' dialog box is open in the foreground, prompting for a Username and Password. The 'Output' panel at the bottom shows a list of files being scanned, including paths like '/ipcc/20c3m/atm/mo/clw...'. The 'Status' bar at the bottom indicates '100.00 %' completion.

Publication step:

- Generate THREDDS catalog for harvesting, server-side configuration
- Release data for harvesting

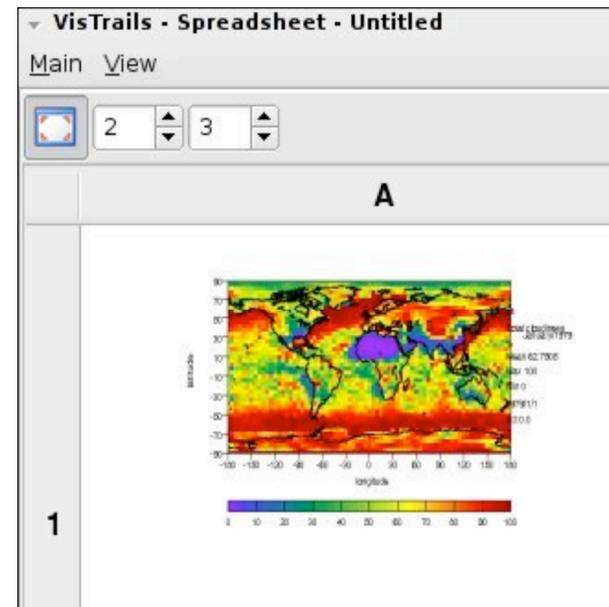
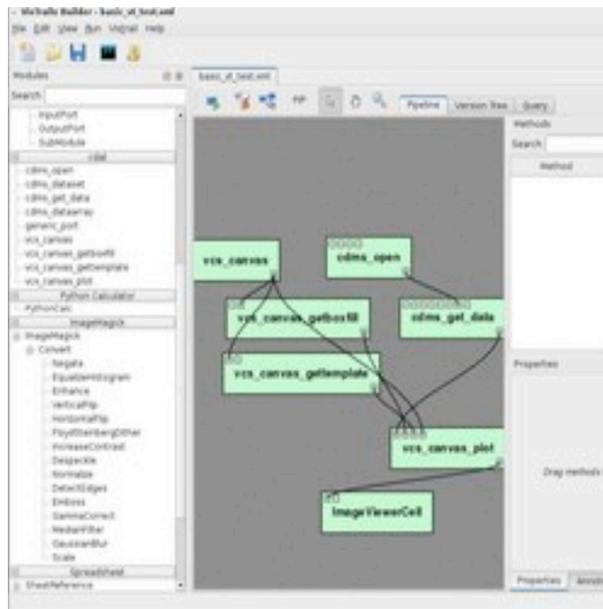
Intercomparison example

Current Usage	Future Usage
<ul style="list-style-type: none"> • Browse PCMDI's centralized database • Download data • Organize data on local site • Regrid data at local site • Perform diagnostics • Produces results 	<ul style="list-style-type: none"> • Search, browse and discover distributed data • Remote site <ul style="list-style-type: none"> ➢ Request data ➢ Regrids ➢ Data system reduction • ESG returns user defined products



Future application: workflow

- **VisTrails** is a new scientific workflow management system. While originally (and solely) developed by researchers at the University of Utah to provide support for data exploration and visualization, VisTrails now is being applied to climate data analysis and visualization as part of the SciDAC-2 Visualization and Analytics Center for Enabling Technology (VACET) collaboration. The image below shows the use of the visual workflow interface to connect CDAT module boxes to perform calculations and a related plot.



The result of the CDAT run viewed in VisTrails showing results in a spreadsheet application
Work on this new GUI application interface for climate data analysis and exploration continues in collaboration with the VACET team.

Client access

- **ESG-CET will expose application programming interfaces (APIs) that allow (non-browser) clients to access data in read-only mode.**
- **Key design challenges:**
 - Authorization and authentication in a distributed environment
 - Client access to files or data aggregations, or both?
- **Eventually want to move the computation closer to the data.**
 - Minimize the amount of data movement
 - Workflow, computational grids
- **Clients such as CDAT (PCMDI Climate Data Analysis Tools) will have ESG access built in.**
- **Analysis and visualization applications will include CDAT, NCL, Ferret, OPeNDAP clients (e.g., Matlab, NCO)**

PCMDI is leading climate community data standards development.

- **CF: The Climate and Forecast Metadata Convention**
 - Designed to **enhance utility of data files** created with the NetCDF application programmer's interface.
 - Standard name table: enables users of data from different sources to decide which quantities are comparable
 - **Facilitates building applications** with powerful extraction, regridding, and display capabilities.
 - PCMDI hosts the CF web site and mailing lists
 - **CMOR: Climate Model Output Rewriter**
 - Used to **produce CF-compliant netCDF files** that fulfill the requirements of many of the climate community's standard model experiments (such as CMIP, CFMIP, NARCCAP, etc.).
 - CMOR-2 is under development for CMIP5 use.
- **GO-ESSP: Global Organization for Earth System Science Portal**
 - A collaboration of developers and institutions from the climate and NWP communities. Focus is on development, information sharing, standards, and interoperability of portals and climate data management software.



Collaborations and publications

- **NOAA GFDL** is an active contributor to AR5 and ESG-CET, CF and GO-ESSP
- **SciDac Scientific Data Management Center (LBNL)**
 - **DataMover Lite** - efficient bulk transfer of data in a secure grid environment.
 - <http://sdm.lbl.gov/sdmcenter/index.html>
- **SciDac Visualization and Analytics Center (VACET)**
 - University of Utah, LLNL, LBNL, ORNL; <http://www.vacet.org>
 - Integration of **VisTrails** visual analysis tool with CDAT.
- **Earth System Curator**
 - Developing database schemas and interfaces for model configuration.
 - <http://www.earthsystemcurator.org/>
- **Cyberinfrastructure Technology Watch (CTWatch) Quarterly**
 - D. N. Williams, D. E. Bernholdt, I. T. Foster, and D. E. Middleton, 2007: The Earth System Grid Center for Enabling Technologies: Enabling community access to petascale climate datasets. CTWatch, Vol. 3, number 4.
- **Bulletin of the American Meteorological Society (BAMS)**
 - D. N. Williams et. al.: The Earth System Grid: Enabling access to multi-model climate simulation data. BAMS (in review).

Recap

- **PCMDI is continuing to play a lead role in standardization and dissemination of model intercomparison datasets.**
- **CMIP5/AR5 is the driving application for ESG-CET**
- **ESG-CET is leveraging the success of ESG-II, to:**
 - support 2-3 orders of magnitude more data,
 - with richer functionality,
 - for a larger group of users.
- **CDAT is an integral part of an emerging global climate data and analysis infrastructure.**