The WGCM Infrastructure Panel (WIP)
Update on CMIP6 Infrastructure

Karl E. Taylor & V. Balaji

Presented at the Twentieth session of the
WCRP Working Group on Coupled Modeling (WGCM

Princeton, NJ
1-2 November 2016

CMIP infrastructure coordination

- The WGCM Infrastructure Panel (WIP) manages and coordinates infrastructure development, implementation, and operations.

- The WIP maintains a website where "Position papers" and specifications for CMIP6 should be examined.
  - https://www.earthsystemcog.org/projects/wip/

Roadmap for contributing model results to CMIP6

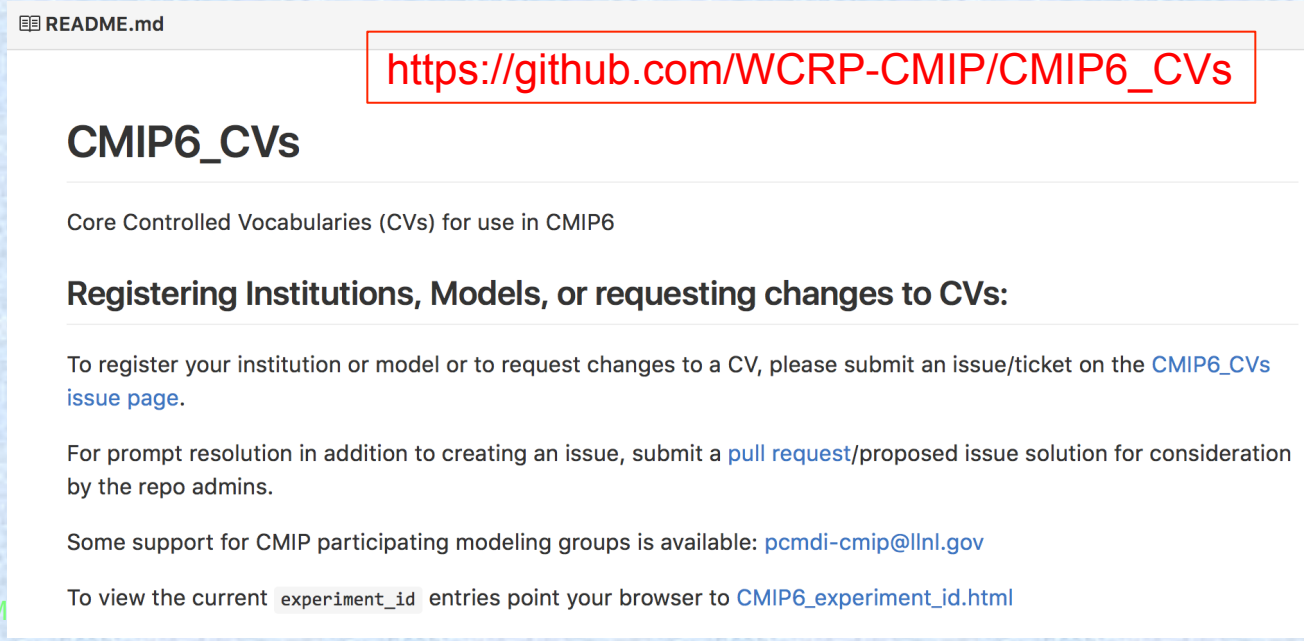
- Registration of institution and model(s)
- Model and experiment documentation (ES-DOC)
- Preparation for running experiments
- Preparation of CMIP6-conforming model output
- Publication of model output on ESGF
- Documention and correction of errors in output

Roadmap for contributing model results to CMIP6

- Registration of institution and model(s)
- Model and experiment documentation (ES-DOC)
- Preparation for running experiments
- Preparation of CMIP6-conforming model output
- Publication of model output on ESGF
- Documention and correction of errors in output

Registration of institution and model(s)

- Before contributing data on ESGF (i.e., "publishing") ,
  - Register model_id, which will be used in file names, search facets, model documentation, etc. (a "nickname" ≤ 16 characters)
  - Register institution.
  - The CMOR validator will prevent publication of models and institutions that are not found in the registry.


README.md

https://github.com/WCRP-CMIP/CMIP6_CVs

CMIP6_CVs

Core Controlled Vocabularies (CVs) for use in CMIP6

Registering Institutions, Models, or requesting changes to CVs:

To register your institution or model or to request changes to a CV, please submit an issue/ticket on the CMIP6_CVs issue page.

For prompt resolution in addition to creating an issue, submit a pull request/proposed issue solution for consideration by the repo admins.

Some support for CMIP participating modeling groups is available: pcmdi-cmip@llnl.gov

To view the current `experiment_id` entries point your browser to CMIP6_experiment_id.html

Roadmap for contributing model results to CMIP6

- Registration of institution and model(s)
- ES-DOC
  - Model and experiment documentation
  - Errata service
- Preparation for running experiments
- Preparation of CMIP6-conforming model output
- Publication of model output on ESGF
- Documention and correction of errors in output

Status of ES-DOC for CMIP6

- Project to document CMIP6 well underway
- Builds on CMIP5 experience (both good and bad !)
  - Metadata in files published on ESGF will be ingested automatically to form a "stub" record of each simulation
  - Modeling groups will supplement information
    - Can start with CMIP5 descriptions
    - multiple tools available (python library or notebooks, questionnaire,…)
    - Beta testing underway (UKMO, GFDL, IPSL); invite additional groups
- March 2017 community release scheduled (for ocean, atmosphere, sea-ice components)
- Working on
  - Forcings description (with Tim Johns et al. e.g. IPCC Table 12.1)
  - Summary descriptions: tables for papers (draft for ocean available)

New ES-DOC errata service

- Records issues (problems) with published datasets
- Provides service for responding to queries about datasets identified by their "persistent identifiers" (PIDs)
  - Datasets are labeled with "persistent identifiers" (PIDs)
  - User can determine whether a queried version of dataset/file is safe to use or is
    - affected by an unresolved issue.
    - Has been superseded by a newer version
- In development:
  - Exposure of errata service to other services (such as the ESGF CoG front-end and Synda) to ensure real time, automated feedback on data status.
  - Incorporation of the issue declaration process in the conventional publishing workflow.
- March 2017 community release scheduled

Roadmap for contributing model results to CMIP6

- Registration of institution and model(s)
- Model and experiment documentation (ES-DOC)
- Preparation for running experiments
  - Obtain "forcing" datasets (input4MIPs)
  - Become familiar with CMIP data request software and requirements
  - Decide whether any data should be regridded
- Preparation of CMIP6-conforming model output
- Publication of model output on ESGF
- Documention and correction of errors in output

Obtain "forcing"

- input4MIPs encourages standard format and structure for "forcing" or "boundary condition" datasets (following specs for CMIP6 model output)

- Summary of forcing dataset status and where to get them:
  - https://pcmdi.llnl.gov/projects/input4mips/

Example: AMIP Boundary Forcing (from Summary)

Contacts: Paul J. Durack durack1@llnl.gov, pcmdi-cmip@lists.llnl.gov

Available at: https://pcmdi.llnl.gov/search/input4mips/

Status: ready for use version: 1.1.1 (2016-10-20)

Further information/documentation: http://www-pcmdi.llnl.gov/projects/amip/AMIP2EXPDSN/BCS

Characteristics of datasets in collection:

- Used in following expts.: AMIP
- Spatial domain: global
- Spatial resolution: 1x1 degree
- Temporal domain: 1870-01 through 2016-06
- Temporal resolution: monthly

Dataset list:

- Sea ice monthly-mean obs (siconc); Data volume: 45MB; 1 file; 1 variable
- SST monthly-mean obs (tos); Data volume: 178MB; 1 file; 1 variable
- Etc.

Usage notes:

Be sure to understand the difference between tosbcs and tos before using the data for AMIP simulations. (see http://www-pcmdi.llnl.gov/projects/amip/AMIP2EXPDSN/BCS

CMIP data request software and requirements

- Through an API, you can determine what variables to save by specifying
  - An experiment
  - A year of the simulation
  - The experiment suite planned for your model
- Metadata associated with each variable are retrievable:
  - e.g., standard_name, units, cell_methods
  - CMOR tables are generated based on the metadata recorded by the data request
- Status
  - MIPs have requested lists of variables
  - CMIP panel is reviewing

CMIP data request tools and documentation

- Primary source found at the WIP CoG site:

  https://www.earthsystemcog.org/projects/wip/CMIP6DataRequest

CMIP6 Data Request

The CMIP6 experimental design and organization has been agreed at the WGCM 18th Session in October 2014, see details on the CMIP Panel website at http://www.wcrp-climate.org/index.php/wgcm-cmip/about-cmip. Part of this covers the creation and timeline of the *CMIP6 Data Request*.

The data request is available through a repository, and the latest version is available here (updated October 21st, 2016):

http://proj.badc.rl.ac.uk/svn/exarch/CMIP6dreq/tags/latest

An overview of the pressure levels proposed for atmospheric diagnostics is available for discussion (here).

Key documents describing the request (in the "docs" directory of the repository) are:

- Examples
- Python Library (dreqPy)
- The Request XML document and Schema
- Spreadsheet view of the variable definitions
- A searchable list of variables in the request, linking to
- A browsable HTML view of the request
- Overview tables for tier 1, priority 1 and all tiers and priorities
- Discussion of issues: old forum, new github pages
- Registration for email list: CMIP6-DATAREQUEST@JISCMAIL.AC.UK
- Installation and usage of the python package

See Version 01.beta.38 Release Notes for more details

When problems are found, raise an issue!
"CMIP6_DataRequest_VariableDefinitions"

Modeling groups may or may not choose to regrid data

- Certain analyses require strictly conservative regridding
- Many scientists are unwilling to analyze data on anything but cartesian latitude x longitude grids
- Analysis of output from high resolution models may be impractical (and availability may be reduced) if data are not regridded to a coarser grid.
- Consider regridding:
  - Data produced on complicated grids
  - High resolution data
  - BUT, be careful to conserve certain quantities
- We will likely request "weights" be provided to regrid data to a few standard grids

Roadmap for contributing model results to CMIP6

- Registration of institution and model(s)
- Model and experiment documentation (ES-DOC)
- Preparation for running experiments
- Preparation of CMIP6-conforming model output
  - Become familiar with required global attributes and controlled vocabularies (CV's)
  - Make licensing choice: Creative Commons Attribution "[NonCommercial] Share Alike" 4.0 International License
  - Prepare requested output by processing through CMOR or checking for conformance with the CMOR validator
- Publication of model output on ESGF
- Documention and correction of errors in output

Standard metadata are recorded in all CMIP6 files

- Identify, for example:
  - Variable
  - Experiment
  - Model
  - Institution
  - Sponsoring MIP
  - Grid information
- Controlled vocabularies (CV's) ensure that metadata can be interpreted by infrastructure software.
  [Reference CVs hosted at: https://github.com/WCRP-CMIP/CMIP6_CVs]
- See "CMIP6_global_attributes_filenames_CVS" document
  - linked from https://www.earthsystemcog.org/projects/wip/position_papers

46 Global attributes are defined in a table (with notes)

The attributes provide critical information needed to interpret the model output and are key attributes are relied on by the infrastructure.

CMIP6 global attribute see note 1	**description**	**examples**	**corresponding attribute in CMIP5**	**form** see note 2	**when required?**	**further information and rationale**
activity_id	activity identifier(s)	"CMIP", "PMIP", "LS3MIP LUMIP" see note 3	project_id	CV	always	renamed more generically, since not all activities are projects; also multiple activities may now be listed separated by single spaces.
branch_method	branching procedure	"standard", "none provided", "no parent" see note 4	-	free form	whenever parent exists	in CMIP6 some branching methods will involve short spin-up periods or other non-standard procedures which need to be described. See note 4. If no parent, omit or set to "no parent"
branch_time_in_child	branch time with respect to child's time axis	365.0D0, 0.0D0 see note 5	-	double precision float	whenever parent exists	aids in interpreting branch times; units are the same as the units used for the child's time axis. If no parent, omit (preferred) or set to start time of the run.
branch_time_in_parent	branch time with respect to parent time axis	3650.0D0 see note 5	branch_time	double precision float	whenever parent exists	changed name to explicitly distinguish it from branch_time_in_child; units are specified in the attribute: parent_time_units. If no parent, omit or set to 0.0D0.

CMOR facilitates (and checks) conformance of files to CMIP6 requirements

- Code available from
  - https://github.com/PCMDI/cmor
- Documentation available at
  - http://cmor.llnl.gov/
- CMOR will be run to check conformance of datasets written by other software.
- Status
  - CMOR3 is being tested by the Hadley Centre
  - CMOR3 is being generalized to facilitate its use by obs4MIPs and input4MIPs

Roadmap for contributing model results to CMIP6

- Registration of institution and model(s)
- Model and experiment documentation (ES-DOC)
- Preparation for running experiments
- Preparation of CMIP6-conforming model output
- Publication of model output on ESGF
  - Become familiar with ESGF
  - Join CMIP Data Node Operations Team (CDNOT)
  - Commit to either Tier 1 or Tier2 node support
- Documention and correction of errors in output

The Earth System Grid Federation (ESGF) status

- Includes funded partners worldwide.
  - DOE, IS-ENES, NASA, NCI, NOAA
  - International Executive Committee
    - Williams (Chair, DOE)
    - Lautenschlager (co-Chair, DKRZ)
    - Denvil (IPSL)
    - Juckes (STFC)
    - Cinquini, Duffy, Duffy (NASA)
    - Balaji, Cinquini, DeLuca (NOAA)
    - Trenham (NCI)
- Development is organized around 18 task teams

ESGF task teams

Task Team	Focus
CoG User Interface	Improved ESGF search and data cart management and interface
Compute	Developing the capability to enable data analytics within ESGF
Dashboard	Statistics related to ESGF usage
Data Transfer	ESGF data transfer and enhancement of the web-based download
Documentation	Document the ESGF software stack
Identity Entitlement Access	ESGF X.509 certificate-based authentication and improved interface
Installation	Installation of the components of the ESGF software stack
International Climate Network	Increase data transfer rates between the ESGF climate data centers
Metadata and Search	ESGF search engine based on Solr5; discoverable search metadata
Node Manager	Management of ESGF nodes and node communications
Provenance Capture	ESGF provenance capture for reproducibility and repeatability
Publication	Capability to publish data sets for CMIP and other projects to ESGF
Quality Control	Integration of external information into the ESGF portal
Replication	Replication tool for moving data from one ESGF center to another
Software Security	Security scans to identify vulnerabilities in the ESFF software
Tracking / Feedback Notification	User and node notification of changed data in the ESGF ecosystem
User Support	User frequently asked questions regarding ESGF and housed data
Versioning	Managing multiple versions of ESGF published data sets

2016 ESGF user survey

327 Responses from a variety of users

Table 1. Which of the following best describes you?	Responses	
Data Provider	8.26%	27
Data Consumer	63.00%	206
Both Provider and Consumer	28.75%	94
Total	327	

Table 2. Which of the following best describes you?	Responses	
Undergraduate Student	2.11%	6
Graduate Student	13.33%	38
Post-Doc	23.86%	68
Academic Scientist/Professional	32.28%	92
Government Scientist/Professional	23.51%	67
Private Scientist/Professional	2.46%	7
Other (please specify)	2.46%	7
Total	285	

Table 3. Which best describes your affiliation?	Responses	
Government Agency	38.11%	109
University	56.64%	162
Private Sector	5.24%	15
Total	286	

Users Rated Importance of ESGF Capabilities (1-5)

Table 1. Top Needs Identified by the Survey	
Survey Question	Average Rating or Percentage in Highest Needed Category
Distributed global search	4.54
Globus download (currently available only for a few data sets)	4.49
Unified data discovery for all ESGF data sources to support your research	4.36
Reliability and resilience of resources	4.25
Ingest and access to large volumes of scientific data (i.e., from data archive to super computer)	4.21
Data access and usage	4.21
User Interface (the web sites, or "CoG")	4.20
Synda download client	4.18
LAS analysis and visualization engine	4.18
Improved designs and principles of user interfaces to enable easier access to computer and software capabilities (e.g. recommendation systems, more flexible and interactive interfaces)	4.13
User support	4.05
Awareness and information of availability of these resources	4.03
Access to sufficient observational and experimental resources	4.01
Direct data delivery into ESGF computing systems from distributed data resources	3.98
Data sharing	3.95
Web documentation	3.91
Web documentation	3.93
Access to enough computational and storage resources	3.88
Data publishing	3.88
Quality Control algorithms for data	3.85
Availability of ancillary data products such as data plots, statistical summaries, data quality information and other documents	3.83
Discovery mechanisms for uncatalogued resources such as software, data in file systems etc.	3.79
User Interface (the web sites, or "CoG")	3.78
Across institutions and communities: Libraries, repositories that allow for community-wide authentication and access	3.78
Easy way to publish and archive your data using one of the ESGF data centers	3.76

ESGF planning documents for CMIP6

- ESGF Governance Policy (http://esgf.llnl.gov/governance.html)
- Logo Requirement and Usage Guidelines (http://esgf.llnl.gov/logo_requirements.html)
- ESGF Strategic Roadmap (http://esgf.llnl.gov/media/pdf/2015-ESGF-Strategic-Plan.pdf)
- Software Security Plan
  (http://esgf.llnl.gov/media/pdf/ESGF-Software-Security-Plan-V1.0.pdf)
- ESGF Federation Policies and Guidelines
  (http://esgf.llnl.gov/media/pdf/ESGF-Policies-and-Guidelines-V1.0.pdf)
- Root Certificate Authorities Policy (DRAFT:
  https://docs.google.com/document/d/16dxkvZy4J83j1nVL8vc_AwqGUSL52m-n8ulXU9eKhpQ/edit)
- ESGF Tier 1 and Tier 2 Node Requirements (under development)
- Data Storage and Replication Plan (under development)
- User Training Plan (under development)
- ESGF CMIP6 Readiness Document (under development).

ESGF-supporting websites

- ESGF public website (esgf.llnl.gov)
- ESGF reports (esgf.llnl.gov/reports.html)
- Software repository website (github.com/esgf)
- International network website (icnwg.llnl.gov)
- CoG tutorial (www.earthsystemcog.org/projects/cog/tutorials_web)

Be sure to engage with the CDNOT

- A technical consortium charged with applying and operationalizing ESGF for CMIP6
- Sébastien Denvil (IPSL) chairs
- Members representing each site hosting CMIP6 data (i.e., most modeling centers and major data centers)
- Membership overlaps with bodies responsible for requirements (WIP) and software development (ESGF, ESDOC, ...)
- Serves to:
  - Communicate WIP discussion to all those of interest
  - Provide input to the WIP of data node/modeling center concerns

Establish (or partner with) an ESGF data node

- Tier 1: Serves multiple models and provides full suite of ESGF services
  - ≥ 10 petabytes of spinning disk storage space
  - ≥ 10 gigabits per second connection to a wide-area network provider
  - Run a 10 gbits/s perfSONAR host
  - Deploy at least four 10 gbits/s Data Transfer Nodes (DTNs)
  - Publish data using GridFTP and Globus URLs in addition to wget URLs,
  - Use Synda for data replication between Tier 1 sites.
- Tier 2
  - For centers that typically have fewer physical or staff resources available for ESGF but need to distribute CMIP6 data
  - Document describing minimum requirements under development

Roadmap for contributing model results to CMIP6

- Registration of institution and model(s)
- Model and experiment documentation (ES-DOC)
- Preparation for running experiments
- Preparation of CMIP6-conforming model output
- Publication of model output on ESGF
- Discovery, documentation, and correction of errors in output

Modeling groups are primarily responsible for "quality assurance", but

- CMOR checker can help with metadata
- ESGF publisher will reject files that don't meet minimum requirements
- Users may discover errors and can report them to the ES-DOC errata service
- ESGF supports versioning, and modeling groups can choose to retain, withdraw, or replace files with errors.

Citation and data tracking

- DOI's will be assigned at a fairly high level (model/ experiment?)
  - A reasonably short list of DOI's can be included in publications.
  - Main requirement: ensure proper citation of data acknowledging contributions by modeling groups
- Persistent IDs (PIDs) will be assigned at fine granularity
  - Web service planned for recording lists of PIDs along with citation info. for CMIP6 publications.
  - ES-DOC errata services will be PID-based
  - Potential use of PIDs in replication workflow.

CMIP6 infrastructure Conclusions

- International governance and oversight in place
- Better coordinated and better tested than for CMIP5.
- Not too far behind schedule
- Considerable documentation available describing
  - Plans
  - Requirements
  - CMIP6 specifications
  - Software

We invite questions/input.

CMIP Data Node Operations Team (CDNOT)

- A technical consortium charged with operationalizing the CMIP6 ESGF Federation
- Sébastien Denvil (IPSL) chairs
- Members representing each site hosting CMIP6 data (i.e., most modeling centers and major data centers)
- Membership overlaps with bodies responsible for requirements (WIP) and software development (ESGF, ESDOC, ...)
- Serves to:
  - Communicate WIP discussion to all those of interest
  - Provide input to the WIP of data node/modeling center concerns