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
- Documentation 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

• Documentation 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.

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

• Documentation 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
• Documentation 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 tosbc 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

When problems are found, raise an issue! “CMIP6_DataRequest_VariableDefinitions”

See Version 01.beta.38 Release Notes for more details
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
- Documentation 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
The attributes provide critical information needed to interpret the model output and are key attributes are relied on by the infrastructure.

<table>
<thead>
<tr>
<th>CMIP6 global attribute</th>
<th>description</th>
<th>examples</th>
<th>corresponding attribute in CMIP5</th>
<th>form</th>
<th>when required?</th>
<th>further information and rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>activity_id</td>
<td>activity identifier(s)</td>
<td>“CMIP”, “PMIP”, “LS3MIP LUMIP” see note 3</td>
<td>project_id</td>
<td>CV</td>
<td>always</td>
<td>renamed more generically, since not all activities are projects; also multiple activities may now be listed separated by single spaces.</td>
</tr>
<tr>
<td>branch_method</td>
<td>branching procedure</td>
<td>“standard”, “none provided”, “no parent” see note 4</td>
<td>-</td>
<td>free form</td>
<td>whenever parent exists</td>
<td>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”</td>
</tr>
<tr>
<td>branch_time_in_child</td>
<td>branch time with respect to child’s time axis</td>
<td>365.0D0, 0.0D0 see note 5</td>
<td>-</td>
<td>double precision float</td>
<td>whenever parent exists</td>
<td>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.</td>
</tr>
<tr>
<td>branch_time_in_parent</td>
<td>branch time with respect to parent time axis</td>
<td>3650.0D0 see note 5</td>
<td>branch_time</td>
<td>double precision float</td>
<td>whenever parent exists</td>
<td>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.</td>
</tr>
</tbody>
</table>
CMOR facilitates (and checks) conformance of files to CMIP6 requirements

- Code available from
  - [https://github.com/PCMDI/cmor](https://github.com/PCMDI/cmor)
- Documentation available at
- 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
- Documentation 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

<table>
<thead>
<tr>
<th>Task Team</th>
<th>Focus</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CoG User Interface</strong></td>
<td>Improved ESGF search and data cart management and interface</td>
</tr>
<tr>
<td><strong>Compute</strong></td>
<td>Developing the capability to enable data analytics within ESGF</td>
</tr>
<tr>
<td><strong>Dashboard</strong></td>
<td>Statistics related to ESGF usage</td>
</tr>
<tr>
<td><strong>Data Transfer</strong></td>
<td>ESGF data transfer and enhancement of the web-based download</td>
</tr>
<tr>
<td><strong>Documentation</strong></td>
<td>Document the ESGF software stack</td>
</tr>
<tr>
<td><strong>Identity Entitlement Access</strong></td>
<td>ESGF X.509 certificate-based authentication and improved interface</td>
</tr>
<tr>
<td><strong>Installation</strong></td>
<td>Installation of the components of the ESGF software stack</td>
</tr>
<tr>
<td><strong>International Climate Network</strong></td>
<td>Increase data transfer rates between the ESGF climate data centers</td>
</tr>
<tr>
<td><strong>Metadata and Search</strong></td>
<td>ESGF search engine based on Solr5; discoverable search metadata</td>
</tr>
<tr>
<td><strong>Node Manager</strong></td>
<td>Management of ESGF nodes and node communications</td>
</tr>
<tr>
<td><strong>Provenance Capture</strong></td>
<td>ESGF provenance capture for reproducibility and repeatability</td>
</tr>
<tr>
<td><strong>Publication</strong></td>
<td>Capability to publish data sets for CMIP and other projects to ESGF</td>
</tr>
<tr>
<td><strong>Quality Control</strong></td>
<td>Integration of external information into the ESGF portal</td>
</tr>
<tr>
<td><strong>Replication</strong></td>
<td>Replication tool for moving data from one ESGF center to another</td>
</tr>
<tr>
<td><strong>Software Security</strong></td>
<td>Security scans to identify vulnerabilities in the ESGF software</td>
</tr>
<tr>
<td><strong>Tracking / Feedback Notification</strong></td>
<td>User and node notification of changed data in the ESGF ecosystem</td>
</tr>
<tr>
<td><strong>User Support</strong></td>
<td>User frequently asked questions regarding ESGF and housed data</td>
</tr>
<tr>
<td><strong>Versioning</strong></td>
<td>Managing multiple versions of ESGF published data sets</td>
</tr>
</tbody>
</table>
# 2016 ESGF user survey

## Users Rated Importance of ESGF Capabilities (1-5)

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

## 327 Responses from a variety of users

### Table 1. Which of the following best describes you?

<table>
<thead>
<tr>
<th>Category</th>
<th>Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Provider</td>
<td>8.26% 27</td>
</tr>
<tr>
<td>Data Consumer</td>
<td>63.00% 206</td>
</tr>
<tr>
<td>Both Provider and Consumer</td>
<td>28.75% 94</td>
</tr>
<tr>
<td>Total</td>
<td>327</td>
</tr>
</tbody>
</table>

### Table 2. Which of the following best describes you?

<table>
<thead>
<tr>
<th>Category</th>
<th>Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Undergraduate Student</td>
<td>2.11% 6</td>
</tr>
<tr>
<td>Graduate Student</td>
<td>13.33% 38</td>
</tr>
<tr>
<td>Post-Doc</td>
<td>23.86% 68</td>
</tr>
<tr>
<td>Academic Scientist/Professional</td>
<td>32.28% 92</td>
</tr>
<tr>
<td>Government Scientist/Professional</td>
<td>23.51% 67</td>
</tr>
<tr>
<td>Private Scientist/Professional</td>
<td>2.46% 7</td>
</tr>
<tr>
<td>Other (please specify)</td>
<td>2.46% 7</td>
</tr>
<tr>
<td>Total</td>
<td>285</td>
</tr>
</tbody>
</table>

### Table 3. Which best describes your affiliation?

<table>
<thead>
<tr>
<th>Category</th>
<th>Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Government Agency</td>
<td>38.11% 109</td>
</tr>
<tr>
<td>University</td>
<td>56.64% 162</td>
</tr>
<tr>
<td>Private Sector</td>
<td>5.24% 15</td>
</tr>
<tr>
<td>Total</td>
<td>286</td>
</tr>
</tbody>
</table>
ESGF planning documents for CMIP6

- ESGF Governance Policy (http://esgf.llnl.gov/governance.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 ESDOC 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 (PIIDs) will be assigned at fine granularity
  ➤ Web service planned for recording lists of PIIDs along with citation info. for CMIP6 publications.
  ➤ ES-DOC errata services will be PID-based
  ➤ Potential use of PIIDs 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