CMIP6's Reliance on ESGF Infrastructure: Present and Future

Karl E. Taylor

With contributions from
The WGCM Infrastructure Panel and PCMDI

Presented at the
8th Annual Earth System Grid Federation Conference

Washington D.C.
3 December 2018
CMIP and the climate modeling community rely on ESGF to serve critical ongoing needs

- Modeling centers have invested heavily in adapting their simulation work flow to be consistent with CMIP requirements.
- Users now expect easy access to multi-model simulation output.
- The WCRP advocates use of ESGF to serve data from international coordinated climate research efforts (e.g., obs4MIPs, CORDEX, input4MIPs)
- We must ensure sufficient resources continue to be invested in ESGF
Outline

• CMIP6 project status

• Review major components of infrastructure supporting CMIP6
  • Present status
  • Deficiencies

• Remarks on longer-term issues
CMIP6 design overview:

- **DECK**
  - Small set of benchmark runs
  - To evolve only slowly (e.g. OMIP, LMIP)

- **Historical CMIPX**
  - Forcing to be updated for each new phase

- **CMIP6-endorsed MIPs**
  - An evolving collection to address specific scientific issues
CMIP5/6 evolution: More institutions, more models, more experiments, more data

- 44 institutions/consortia have officially registered for CMIP6
- 100 models are registered
- 287 experiments defined
- order 20 PB of model output expected

CMIP6_CVs
https://github.com/WCRP-CMIP/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 following the instructions on the CMIP6_CVs issue page.

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

To view the current institution_id entries point your browser to CMIP6_institution_id.html

To view the current source_id entries point your browser to CMIP6_source_id.html

The CVs build on logic that is described in the CMIP6 Global Attributes, DRS, Filenames, Directory Structure, and CV’s document.
CMIP6 status: data availability and IPCC timeline

- Model output now being served by ESGF from 4 institutions (6 models) and 19 experiments
- Much output to be made available over the next year

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>2019</td>
<td>January 7: Second Lead Author Meeting</td>
</tr>
<tr>
<td></td>
<td>April 29: First order draft expert review</td>
</tr>
<tr>
<td></td>
<td>August 26: Third Lead Author Meeting</td>
</tr>
<tr>
<td>2020</td>
<td>March 2: Second order draft expert review</td>
</tr>
<tr>
<td></td>
<td>June 1: Fourth Lead Author Meeting</td>
</tr>
<tr>
<td></td>
<td>October 18: Submission of final draft</td>
</tr>
<tr>
<td>2021</td>
<td>April 16: IPCC acceptance/adopter/approval</td>
</tr>
</tbody>
</table>

- 31 December 2019: Journal articles submitted
- 30 September 2020: Journal articles accepted
Model output & data base specifications and global metadata requirements

• Status: In place!

• Should not be modified.

• Reasonably well documented:
  • Definition of CMIP6 netCDF global attributes
  • Specifications for file names, directory structures, and CMIP6 Data Reference Syntax (DRS)
  • Specifications for output file content, structure, and metadata.
  • Guidance on grid requirements
  • Guidance on time-averaging (with masking)
  • Specification for search facets

• See: https://pcmdi.llnl.gov/CMIP6/Guide/modelers.html#5-model-output-requirements
Infrastructure components and dependencies

ESGF services
- CoG search
- errata services
- PID services
- citation services
- ES-DOC

ESGF core software

Model output preparation
- netCDF model output files
- Climate Model Output Rewriter (CMOR3)

CMIP6 specifications

Reference Controlled Vocabularies (CVs)
- Data request database (DREQ)
- CF Metadata Conventions
- Model output & database specs.
Reference ”controlled vocabularies” (CVs)

- CVs allow users and individual infrastructure elements to communicate.
- Recorded in JSON files
- For ”institution” and ”source” vocabularies are registered by participating groups
- Dictionaries record allowed relationships (e.g., which sub-experiments are associated with a given experiment)
- Status: all needed CV’s defined, including activity, Institution, model, experiment, sub-experiment, realm, frequency, .....
Institutions and models must be registered in CVs

Status: Operational

CMIP6_CVs
https://github.com/WCRP-CMIP/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 following the instructions on the CMIP6_CVs issue page.

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

To view the current institution_id entries point your browser to CMIP6_institution_id.html

To view the current source_id entries point your browser to CMIP6_source_id.html

The CVs build on logic that is described in the CMIP6 Global Attributes, DRS, Filenames, Directory Structure, and CV’s document.
Infrastructure components and dependencies

ESGF services
- CoG search
- errata services
- PID services
- citation services
- ES-DOC

ESGF core software

Model output preparation
- netCDF model output files
- Climate Model Output Rewriter (CMOR3)

CMIP6 specifications
- Data request database (DREQ)
- CF Metadata Conventions
- Model output & database specs.

Reference Controlled Vocabularies (CVs)
CMIP data request tools and documentation (DREQ; Martin Juckes)

Specifies:

- Which variables should be saved for a given experiment and time period, and at what frequency.
- What variable-specific metadata should be included in output files.

See WIP CoG site

Status: In place and in use!

- "Cosmetic" refinements still being made
- New releases undergo review

---

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
Infrastructure components and dependencies

ESGF services
- CoG search
- errata services
- PID services
- citation services

ESGF core software

Model output preparation

CMIP6 specifications

ESGF archive, catalogue, and services

- Climate Model Output Rewriter (CMOR3)

- netCDF model output files

- Data request database (DREQ)

- CF Metadata Conventions

- Model output & database specs.

Reference Controlled Vocabularies (CVs)
Climate Model Output Rewriter (CMOR3)

- Facilitates (and checks) conformance of files to CMIP6 requirements

- **Status:** in place and in use!
  - Code available at [https://github.com/PCMDI/cmor](https://github.com/PCMDI/cmor)
  - Development phase is complete
  - Bugs corrected when discovered
Pre-Publication Attribute Reviewer for ESGF (PrePARE)

- Checks conformance of files to CMIP6 requirements
  - Available for use by groups not relying on CMOR
  - Included as part of the publication job stream

- Status: in place and in use (??)
  - Code is part of the CMOR build (??)
  - Development phase is complete
  - Bugs corrected when discovered
  - Certain checks of variable attributes are currently too stringent and will be relaxed.
ESGF core software stack

• Supports a federated data archive hosting the CMIP6 data

• Status: In place and operational!
  • Output from 6 models and 19 experiments now available
  • Served through 5 CMIP6 CoG data portals

• The work of the project scientists and the CDNOT in performing the “data challenge” tests ensured a smooth launch of ESGF for CMIP6.

THANK YOU!

• Replication procedure is working
  • PCMDI has replicated ~65% of available; DKRZ ~10%.

• To do (high priority): enable Globus grid ftp at all sites for all datasets.
Infrastructure components and dependencies

**ESGF services**
- CoG search
- errata services
- PID services
- citation services

**ESGF core software**

**Model output preparation**
- Climate Model Output Rewriter (CMOR3)
  - netCDF model output files
  - PrePARE

**CMIP6 specifications**
- Data request database (DREQ)
- CF Metadata Conventions
- Model output & database specs.

**Reference Controlled Vocabularies (CVs)**
ESGF services: CoG

• Provides users with an interface to browse and download data

• Status: All CMIP6 data now available via any one of 5 CoG sites

• All sites should have identical look and feel
  • Rules and format are given at https://goo.gl/P1e18T
  • IPSL, PCMDI/LLNL, and CEDA are in full compliance, and GFDL and DKRZ have been asked to make minor modifications.
Data citation services

1. CMIP6.CFMIP.IPSL.IPSL-CM6A-LR.abrupt-0p5xCO2.r1i1p1f1.CFmon.albiscrp.gr
   Data Node: vesg.ipsl.upmc.fr
   Version: 20180605
   Total Number of Files (for all variables): 1
   Full Dataset Services: [Show Metadata] [List Files] [THREDDS Catalog] [WGET Script] [LAS] [Show Citation] [PID] [Globus Download]
   Add to Data Cart

Show Citation
**General Information**

<table>
<thead>
<tr>
<th>Name</th>
<th>CMIP6.CFMIP.IPSL.IPSL-CM6A-LR.abrupt-0p5xCO2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>Coupled Model Intercomparison Project Phase 6 (CMIP6) data sets. These data includes all datasets published for 'CMIP6.CFMIP.IPSL.IPSL-CM6A-LR.abrupt-0p5xCO2' according to the Data Reference Syntax defined as 'mip_era.activity_id.institution_id.source_id.experiment_id.member_id.table_id.variable_id.grid_label.version'. The Earth System Model IPSL-CM6A-LR, released in 2017, includes the components: atmos: LMDZ (NPv6, N96; 144 x 143 longitude/latitude; 79 levels; top level 40000 m), land: ORCHIDEE (v2.0, Water/Carbon/Energy mode), ocean: NEMO-OPA (eORCA1.3, tripolar primarily 1deg; 362 x 332 longitude/latitude; 75 levels; top grid cell 0-2 m), ocnBchem: NEMO-PISCES, seaIce: NEMO-LIM3. The model was run by the Institut Pierre Simon Laplace, Paris 75252, France (IPSL) in native nominal resolutions: atmos: 250 km, land: 250 km, ocean: 100 km, ocnBchem: 100 km, seaIce: 100 km. Project: These data have been generated as part of the internationally-coordinated Coupled Model Intercomparison Project Phase 6 (CMIP6) see also CMD Special Issue: <a href="http://www.eso.org/magazine/pr/2018/06/index.html">http://www.eso.org/magazine/pr/2018/06/index.html</a>. The...</td>
</tr>
</tbody>
</table>

### ES-DOC Documentation

<table>
<thead>
<tr>
<th>Category</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIP Era</td>
<td>CMIP6</td>
</tr>
<tr>
<td>Institution</td>
<td>IPSL</td>
</tr>
<tr>
<td>Model</td>
<td>IPSL-CM6A-LR</td>
</tr>
<tr>
<td>Experiment</td>
<td>abrupt-0p5xCO2</td>
</tr>
<tr>
<td>Ensemble Description</td>
<td>N/A</td>
</tr>
<tr>
<td>Machine Performance</td>
<td>N/A</td>
</tr>
</tbody>
</table>

### Dataset Documentation

<table>
<thead>
<tr>
<th>Category</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset ESGF Search</td>
<td>N/A</td>
</tr>
<tr>
<td>Dataset Errata</td>
<td>N/A</td>
</tr>
<tr>
<td>Dataset Citation(s)</td>
<td><a href="https://cera-www.dkrz.de/WDCC/meta/CMIP6/CMIP6.CFMIP.IPSL.IPSL-CM6A-LR.abrupt-0p5xCO2">https://cera-www.dkrz.de/WDCC/meta/CMIP6/CMIP6.CFMIP.IPSL.IPSL-CM6A-LR.abrupt-0p5xCO2</a></td>
</tr>
</tbody>
</table>

### Other Documentation

<table>
<thead>
<tr>
<th>Category</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>WCRP CMIP6 Homepage</td>
<td><a href="https://www.wcrp-climate.org/wgcm-cmip/wgcm-cmip6">https://www.wcrp-climate.org/wgcm-cmip/wgcm-cmip6</a></td>
</tr>
<tr>
<td>ES-DOC CMIP6 Homepage</td>
<td><a href="https://es-doc.org/cmip6">https://es-doc.org/cmip6</a></td>
</tr>
</tbody>
</table>
Summary:

- The major infrastructure elements enabling scientists to access data are in place and working satisfactorily.

- Users continue to complain that
  - Globus gridftp is unavailable at some nodes
  - Server-side computation capability is missing
    - subsetting
    - simple reduction (climatology, zonal mean, etc.)

- The ES-DOC software has advanced considerably, but modeling groups have not yet provided content.

- Better documentation (and tutorials) across the infrastructure should help us make up for our lack of user support
CMIP6 website provides practical guidance for all parties interested in CMIP

Example: Guidance document for data users

Points to information on experimental design

Describes and links to controlled vocabularies

Documents essential model output specifications

Describe how to access output

Remarks on longer-term issues

• Despite growing international investment in climate modeling infrastructure, it remains fragile: “single points of failure” need to be addressed because they can lead to enormous disruptions
  • Some individuals are irreplaceable
  • Some software is not well documented

• ESGF has become essential to the climate research community:
  • CMIP, input4MIPs, obs4MIPs, etc.
  • Modeling groups have invested in it

• Given resource constraints, we should
  • Treat ESGF as part of an operational climate research enterprise; it must be reliable
  • Avoid pursuing revolutionary changes, even when promising, if that would threaten or weaken support for the working infrastructure. [Beware of the cloud!]