CMIP6: WIP Status Update

A summary of CMIP6 to-date from the WGCM Infrastructure Panel (WIP)

WGCM-25
Wednesday 9th November 2022 - Hybrid

Paul J. Durack, Matthew Mizielinski and the WIP membership
CMIP: a team sport

- PCMDI
  - U.S. DOE has provided 33-years of *MIP support
- ESGF
  - Originated by U.S. DOE
  - Major recent contributions from numerous others
- IS-ENES and ENES-RI follow-on
  - European contribution to ESGF & CMIP infrastructure
- Numerous other projects and institutions, including DKRZ, IPSL, CEDA, ES-DOC, NASA, NOAA, ...
- 30+ ESGF nodes, 17 countries
- 131 models, 48 institutions representing 26 countries, and many, many more...
ESGF Published data

- Over 3.7 5.6 6.4 million datasets on ESGF across all CMIP6 activities/MIPs
- Delivery has been seamless – thanks to data challenges and ESGF stability testing
- Datasets - unique variable collections per experiment RIPF
- Footprint – storage units in PBs
ESGF publication and replication

- 17.7 21.9 24.5 PB CMIP6 data available including 10 12 13.6 PB unique and 8 10 10.8 PB replicated
- 16.2 27.9 36 PB CMIP6 downloads (to November 2022)
- LLNL 27 25 27% downloads to date
- DKRZ 20 18 18%
- CEDA 12 14 12%
- LIU 9%
- CNRM 15 11 9%
- UCAR 9 8%

WGCM-23/December 2020 vs
WGCM-24/December 2021 vs
WGCM-25/November 2022

http://esgf-ui.cmcc.it/esgf-dashboard-ui/federated-view.html - 30 nodes in total, 11 reporting statistics
CMIP6 controlled vocabulary

- 137 140 132 models registered with CMIP6 CVs
- Each model involved in 6* activities on average
- Experiments grown from ~280 to 322 including six “CovidMIP” experiments added to DAMIP
- Added CDRMIP and PAMIP in March 2018
- Added “COVIDMIP” in November 2020

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

https://wcrp-cmip.github.io/CMIP6_CVs/docs/CMIP6_experiment_id.html
https://wcrp-cmip.github.io/CMIP6_CVs/docs/CMIP6_institution_id.html
https://wcrp-cmip.github.io/CMIP6_CVs/docs/CMIP6_source_id.html
137 140 132 models registered with CMIP6 CVs
Each model involved in 5* activities on average
Experiments grown from ~280 to 322 including six “CovidMIP” experiments added to DAMIP
Added CDRMIP and PAMIP in March 2018
Added “COVIDMIP” in November 2020

https://github.com/WCRP-CMIP/CMIP6_CVs
https://wcrp-cmip.github.io/CMIP6_CVs/docs/CMIP6_experiment_id.html
https://wcrp-cmip.github.io/CMIP6_CVs/docs/CMIP6_institution_id.html
https://wcrp-cmip.github.io/CMIP6_CVs/docs/CMIP6_source_id.html
A first step beyond CMIP6: CMIP6Plus
CMIP6Plus

• Structure provided by CMIP6 allows users to build tools and workflows to analyse large amounts of data from many models

• Activities (MIPs) are keen to have their data usable and accessible alongside CMIP6 data

• How can we extend the structures we have to support new MIPs and the extension of existing ones?
CMIP6Plus

- Ability of the CMIP6 infrastructure to extend to support new science demonstrated by CovidMIP
- Explore benefits and challenges of a more operationalised infrastructure
- Opportunity to think about changes and trial them in advance of CMIP7
Proposed changes: Variable definitions & CVs

- CMOR standards rely on
  - Variable definitions (MIP tables)
  - Controlled vocabularies (CVs)
- Currently MIP tables and CVs are combined in a project specific repository
- Separate the variable definitions such that they can be re-used by different projects

### mip-cmor-tables

**# Variable definitions**
- `table1.json`
- `table2.json`
- `...`
- `tableN.json`

coordinates, grids and formula terms

**# Generic controlled vocabularies**
- frequencies
- calendars
- source type
- institution ids
- esgf node ids
- modeling realms

regions (CF or CORDEX)

variant_label and associated indices

**# List of known compatible projects**
(linking to project repositories)

Equivalent XML documents as required

Contains all variable quantities, generic identifiers defined across all WCRP projects

### Projects

**# one repo per project**

CMIP6Plus_CVs
- input4MIPs_CVs
- obs4MIPs_CVs
  - (CORDEX_CVs)
  - (ESMO_CVs)
  - (GEWEX_CVs)
  - (SPARC_CVs)
  - (OtherProject_CVs)
  - ...

**# Required entries**

DRS (directory structure and file naming conventions), `<activity_id>`, `<experiment_id>`, license, `<mip_era>`, required _global_attributes_, `source_id`, `tracking_id_prefix`, `<further_info_url>`, `<sub_experiment_id>`, `prime_esgf_node_id`, `permitted_variables`

(<..> denotes optional fields that may not be needed for all projects)

**# Synthesis files**
(for CMOR and ESGF use)

- e.g. `<project>_CVs.json`
- `<project>_ESGF.json` (ESGF.ini)

Contains all project-specific information relevant for data being written and ESGF to publish/host
Shared and project specific information

Common items
• Variable definitions
• Frequencies, calendars
• Coordinate descriptions
• Institutions
• Various labels (grid, variant)
• Provenance (link to CMIP3/5/6 versions and Data Requests)

Project specific items
• DRS (directory structure and file naming conventions)
• licenses
• Global attributes
• Experiments
• Models
• Connected services
Proposed entry requirements

1. Minimum number of institutions/modelling groups involved
2. MIP definition paper similar to CMIP6 GMD
3. Define new forcing data and publish to input4MIPs
4. Define list of variables required
5. **Estimate and agree on data volumes, and arrange for storage/publication (on ESGF or elsewhere)** – Funding required
Looking towards CMIP7
In process of establishing a number of CMIP Task Teams to drive forward definition of CMIP7 in an open and collaborative manner.

An open call to the community for applications was launched in August 2022.

Over 120 applications were received.

Evaluation and shortlisting is now complete.

Successful applicants to be invited during November.

First meetings (online) before Christmas.
CMIP7 Task Teams

- Forcings (*Paul Durack and Vaishali Naik*)
- Data Request (*Martin Juckes and Chloe Mackallah*)
- Model benchmarking (*Birgit Hassler and Forrest Hoffman*)
- Data citation (*Martina Stockhause and Sasha Ames*)
- Model documentation (*David Hassell and Guillaume Levavasseur*)
- Strategic ensemble design (*Ben Sanderson and Isla Simpson*)

*A further task team on Data Access will be opened for applications before Christmas and led by Robert Pincus and co-lead tbc.*
Work completed by the PCMDI project is funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Regional and Global Model Analysis Program.
<table>
<thead>
<tr>
<th>Dataset Title</th>
<th>Files</th>
<th>Total Size</th>
<th>Version</th>
<th>Download Script</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMIP6.ScenarioMIP:CCCma.CanESM5.ssp126.r121p2f1.day.ctl.gn</td>
<td>1</td>
<td>587.77 MB</td>
<td>20190429</td>
<td>wget</td>
</tr>
<tr>
<td>CMIP6.ScenarioMIP:CCCma.CanESM5.ssp126.r121p2f1.day.hfsls.gn</td>
<td>1</td>
<td>911.31 MB</td>
<td>20190429</td>
<td>wget</td>
</tr>
<tr>
<td>CMIP6.ScenarioMIP:CCCma.CanESM5.ssp126.r121p2f1.day.hrfls.gn</td>
<td>1</td>
<td>907.15 MB</td>
<td>20190429</td>
<td>wget</td>
</tr>
<tr>
<td>CMIP6.ScenarioMIP:CCCma.CanESM5.ssp126.r121p2f1.day.hur.smin.gn</td>
<td>1</td>
<td>840.16 MB</td>
<td>20190429</td>
<td>wget</td>
</tr>
</tbody>
</table>
Your master branch isn’t protected
Protect this branch from forced-pushing, deletion, or require status checks before merging. Learn more

<table>
<thead>
<tr>
<th>Author</th>
<th>Description</th>
<th>Last commit</th>
<th>Commits</th>
</tr>
</thead>
<tbody>
<tr>
<td>mazuy1</td>
<td>Merge pull request #304 from PCMD/ncar_test</td>
<td>14 hours ago</td>
<td>716</td>
</tr>
<tr>
<td>.circleci</td>
<td>Update CircleCI Linux image to ubuntu-2004/202201-02</td>
<td>9 months ago</td>
<td></td>
</tr>
<tr>
<td>Tables</td>
<td>cron: update CMIP6-CV -- 2022-11-09T00:08</td>
<td>16 hours ago</td>
<td></td>
</tr>
<tr>
<td>ci-support</td>
<td>should be building and testing now</td>
<td>4 years ago</td>
<td></td>
</tr>
<tr>
<td>scripts</td>
<td>Fix spaces in license template</td>
<td>3 months ago</td>
<td></td>
</tr>
<tr>
<td>LICENSE.md</td>
<td>Create LICENSE.md</td>
<td>4 months ago</td>
<td></td>
</tr>
<tr>
<td>README.md</td>
<td>Update README.md</td>
<td>5 years ago</td>
<td></td>
</tr>
</tbody>
</table>

Note:
Only mazuy1 and if this is should use the CI/CD file of the repository process. CMIP6-CV is chosen.
CMIP immediate future?

CMIP6+, input4MIPs and obs4MIPs

Some steps toward an “operational” CMIP from the WGCM Infrastructure Panel (WIP)

WGCM-25
Wednesday 9th November 2022 - Hybrid

Matthew Mizielinski, Paul J. Durack, and the WIP membership
How can we leverage existing infrastructure investments?

- Agile, responsive evolution
  - Continuous DECK is a start
  - Facilitate, respond and enable science opportunities
    - COVIDMIP, ZECMIP/C4MIP

- Allow CMIP to evolve and “operationalize”
  - Incremental change (e.g. maintain ESGF dependence)
  - Next gen forcings and obs
  - Change little, increment, allowing modeling groups to focus on science

- Best prepare CMIP for exascale and the AI/ML onslaught
CMIP6+

How can we leverage existing communities?

• Continue beyond CMIP6, few changes as CMIP7 is discussed and planned
  • Reduce time pressures - loosen CMIPx - IPCC ARx linkage
  • Continuous “CMIP science” - not monolithic phases every ~7 years

• Facilitate and recognize contributors
  • Ensure ALL contributions are recognized
  • How can we aid forcing data providers?
    • Funding? (“CMIP endorsed” data provider)
    • Infrastructure support?
CMIP6+ a mud map

A forcing evolution following the continuous CMIP DECK paradigm

Transition between MIP-era model simulations

Transition between MIP-era forcing datasets (broader, prototype datasets need iteration before “formal” model simulations begin)
input4MIPs

How can we leverage existing communities and infrastructure?

• Allow the CMIP DECK (and *MIPs?) to evolve
  • CMIP6-era forcings conclude in 2014, but data providers have updates
    • PCMDI AMIP data updated to June 2021, six-monthly updates scheduled
    • PNNL/UMD CEDS/Emissions data updated to near real-time (~May 2021)
    • NASA GloSSAC v2/SAOD updated to December 2018
  • Update CMIP6+ forcing data to near real-time
    • CMIP6-era models re-run with new forcings - piControl, AMIP, historical-ext
      • Be responsive to science opportunities - e.g. Pinatubo 2.0/COVIDMIP
      • Evaluate new forcing datasets before CMIP7 “prime time”
      • Potentially more than a single endorsed forcing can be evaluated
    • “CMIP7” model development aided with latest-generation forcing
Feedback from modelling groups

• CMORize forcing data
  • Many datasets don’t align with single variable CMIP data standard
  • Is data format provided fit for purpose or rewritten?
• Extend ESGF data search capabilities
• Better document/more transparent IAM-generated scenario data
  • Are IAM inconsistencies a problem?

Other ideas

• Missing forcings? (IPSL: N-cycle, water isotopes, )
• Forcing data problems? (Led to 3 CMIP6 releases: 6.0 Dec 2016, 6.1 May ‘17, 6.2.1 Oct ‘17)
obs4MIPs update

- Limited progress in the last few years has led to a rethink of how to make obs4MIPs more useful. A revitalization of the effort is underway.

- A new emphasis strives to streamline how products can be made compliant with CMIP/obs4MIPs.

- Prioritize adherence to data standards with a more agnostic approach to data quality.
Three tiers of obs4MIPs

1. Version controlled obs4MIPs compliant datasets

1. Compliant datasets published on ESGF

1. Reviewed ESGF-published datasets (primarily assessing compliance with standards, with quality judgements mostly made elsewhere, e.g., GEWEX/GDAP Assessments)
obs4MIPs in 2022

- Project site to be overhauled and migrated from CoG to WCRP
- 3rd party contributions are being enabled (i.e., not required to be processed by original data curators)
- Codes used to process each dataset to be included in the version control - shared experience via code repo expected to expedite new contributions
- Many new/updated datasets to be made available via 3rd party contributions
- Reformulation of a project team underway including new contributors (P. Gleckler, LLNL-ret; S. Pinnock, ESA; N. Caltabiano, WCRP; S. Ames, LLNL; P. Durack, LLNL; R. Ferraro, JPL; G. Elsaesser, GISS). There are other contributors joining and we welcome the involvement of interested parties.
CMIP7 some ideas

Can we optimize to meet the science goals, rather than bloat the archive?

• Not just data request - rather *MIPs provide diagnostics/code to implement
  • Rather than requesting data, request the targeted diagnostic
  • Plus, less data; minus, locks out spontaneous science opportunities
• MIPs define diagnostics to implement within models
  • Advance the inclusion of key simulators (ala COSP)
  • Encourage MIP diagnostic team development - move workload to MIP chairs, not modellers
• How best to leverage community diagnostics:
  • ESMValTool and CMEC (Coordinated Model Evaluation Capabilities)
• Can we amalgamate efforts to reduce overheads (input4MIPs, obs4MIPs, ...)?
Harmonizing CMIP Data Holdings Across Phases and Activities

WGCM-24
Thursday 9th December 2021 - Virtual

Karl. E. Taylor, Paul J. Durack, Matthew Mizielinski, and the WIP membership
Background

• The collection of WCRP-endorsed MIP data now spans more than three decades
  • Multiple activities: AMIP, PMIP, CMIP, CORDEX, DCPP, obs4MIPs, input4MIPs
  • Data requirements have become increasingly stringent and refined
  • More comprehensive descriptions of models and experiments have been captured in metadata

• Our rich collection of model output should continue to be exploited in scientific studies
  • For example, serving the needs of machine learning exercises
Current state of MIP data collections

• Fortunately, except for the earliest datasets, all output files are netCDF and compliant with the CF standards.

• Use of older MIP datasets is hampered, however, by
  • Incomplete metadata (model names, configurations etc), primarily in early MIP phases
  • Incomplete documentation of forcing datasets
  • Renaming of some metadata attributes across eras
  • Differences in templates for constructing file names
  • Differences in controlled vocabularies (if they exist)

We could facilitate research by harmonizing the archive across generations!
PCMDI, with WIP guidance, is developing a harmonization strategy

• We have analyzed the metadata of all past recent phases of CMIP, CORDEX, obs4MIPs, and input4MIPs, which includes:
  • Data reference syntax (DRS) used to uniquely identify datasets
  • Global attributes, including DRS elements, but also additional information about a model and its simulation output
  • File and directory structures
27 data descriptors have been defined across 6 WCRP activities

<table>
<thead>
<tr>
<th>data descriptor generic name</th>
<th>WCRP Activity</th>
<th>Global Attribute name</th>
<th>In File Name?</th>
<th>In Directory Structure?</th>
<th>ESGF CoG Search Facet name</th>
<th>required by documentation service</th>
<th>required by citation service</th>
<th>CV defined by activity</th>
<th>CV entries registered</th>
<th>CV in JSON file?</th>
<th>multiple values allowed?</th>
</tr>
</thead>
<tbody>
<tr>
<td>sourceDD</td>
<td>CMIP3</td>
<td>source</td>
<td>yes</td>
<td>yes</td>
<td>Model</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>CMIP5</td>
<td>model_id</td>
<td>yes</td>
<td>yes</td>
<td>Source ID</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>CMIP6</td>
<td>source_id</td>
<td>yes</td>
<td>yes</td>
<td>Source ID</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>input4MIPs</td>
<td>source_id</td>
<td>yes</td>
<td>yes</td>
<td>Source ID</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>CORDEX</td>
<td>model_id</td>
<td>yes</td>
<td>yes</td>
<td>RCM Model</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>obs4MIPs</td>
<td>source_id</td>
<td>yes</td>
<td>yes</td>
<td>Source ID</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>realmDD*</td>
<td>CMIP3</td>
<td>modelingRealm</td>
<td>yes</td>
<td>Realm</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>CMIP5</td>
<td>modelingRealm</td>
<td>yes</td>
<td>yes</td>
<td>Realm</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>CMIP6</td>
<td>realm</td>
<td>yes</td>
<td>yes</td>
<td>Realm</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>input4MIPs</td>
<td>realm</td>
<td>yes</td>
<td>yes</td>
<td>Realm</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>CORDEX</td>
<td>realm</td>
<td>yes</td>
<td>yes</td>
<td>Realm</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>obs4MIPs</td>
<td>realm</td>
<td>yes</td>
<td>yes</td>
<td>Realm</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
</tbody>
</table>
What has led to inconsistencies in MIP metadata?

• Specifications for data produced by WCRP-endorsed projects have become increasingly complex due to increasing diversity of
  • Activities (CMIP, CORDEX, obs4MIPs, input4MIPs, ...)
  • Experiments
  • Model types (AOGCMs, ice sheet, offline radiation ...)
  • Data fields (gridded vs. site, mean vs. synoptic ...)

• The increased diversity has led to an evolution of metadata used
  • To uniquely identify datasets
  • In search facets (e.g., by ESGF search engine)

• Some descriptors are not always relevant across projects (e.g., experiment_id)
What about the future metadata needs?

• We will likely need more flexibility in the types of data collected and in the data structures required ("CMORization" may not be appropriate in all cases)

• The WIP seeks to
  • Stabilize data requirements, while
  • establishing a flexible framework to accommodate future requirements

• Advantages in modifying current metadata requirements will need to be gauged against their impact on modeling groups and users
  • Will modeling groups need to modify their workstreams
  • Will data users seeking to analyze data from multiple activities/phases be confused by nuanced changes in search terms and metadata.
What needs fixing? CMIP6 shortcomings:

- **Anticipated issues:**
  - Proliferation of CMOR tables (43 in CMIP6); somewhat obscure table names
  - Some fields recorded on more than one grid (e.g., native + 1x1 deg)
  - Some fields recorded with and without masking (e.g., surface fluxes for atmosphere, ocean, land, sea ice, etc.)
  - Multiple institutions contributing with a common model

- **Unanticipated issues:**
  - Experiments performed using CMIP5 forcing fields
  - New experiments added by activities after CMIP panel approval (e.g., COVIDMIP, 11/20 - partially resolved by adding experiments to DAMIP)
  - New forcing datasets created (e.g., extending AMIP boundary conditions)
Harmonizing the past and accommodating the future metadata needs: some specifics

• Facilitate recognition of aliases

• Record controlled vocabularies (CVs) for previous CMIP phases and all activities in commonly structured json files

• Expand registered CVs for "source_id" to include documentation essential for analysis of results:
  • Define the meaning of each integer appearing in an “ripf” variant identifier
  • Define the meaning of each integer appearing in a ”grid_id”

• Replace use of the “CMOR table name” in uniquely identifying datasets with more descriptive independent elements (e.g., frequency, realm, sampling)

• Enforce a uniform definition of attributes (for identification and search services) but allowing flexibility in the subset required by each activity
Improving adaptability of the infrastructure

• Accommodate flexibility in the requirements for data and metadata.
  • Strict and extensive requirements for historical and scenario MIP type experiments
  • Looser and fewer requirements for experiments serving a specialized community

• Implement the concept of ”data collections” that wrap together data from related activities into searchable databases
  • e.g., each MIP might have its own data collection, and some subset of the experiments might also be included as part of a CMIP7 collection
  • Activities could generate data of specialized interest, which might not be fully "cmorized"
The WIP welcomes modeling group input

• Please report shortcomings of the current infrastructure
  
  Complete CMIP6 survey (when available)
  
  Contact WIP co-chairs: durack1@llnl.gov and matthew.mizielinski@metoffice.gov.uk

• Please provide feedback about future plans
  
  A report from the WIP detailing plans will be circulated within the next few months

• We will need help checking the source_id and institution_id CV's generated for past CMIP phases (for harmonization purposes)