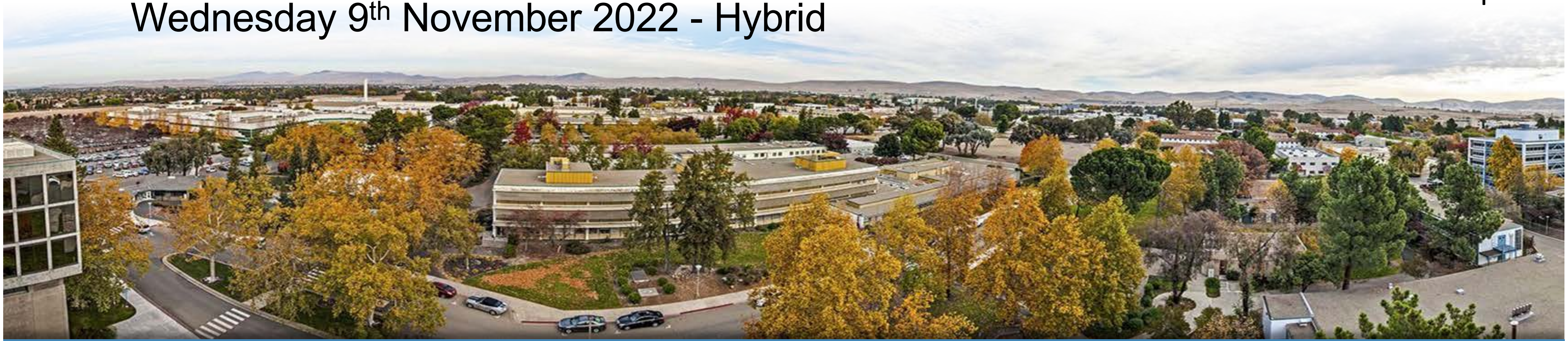
# CMIP6:

# 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



LLNL-PRES-842123

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

Lawrence Livermore National Laboratory

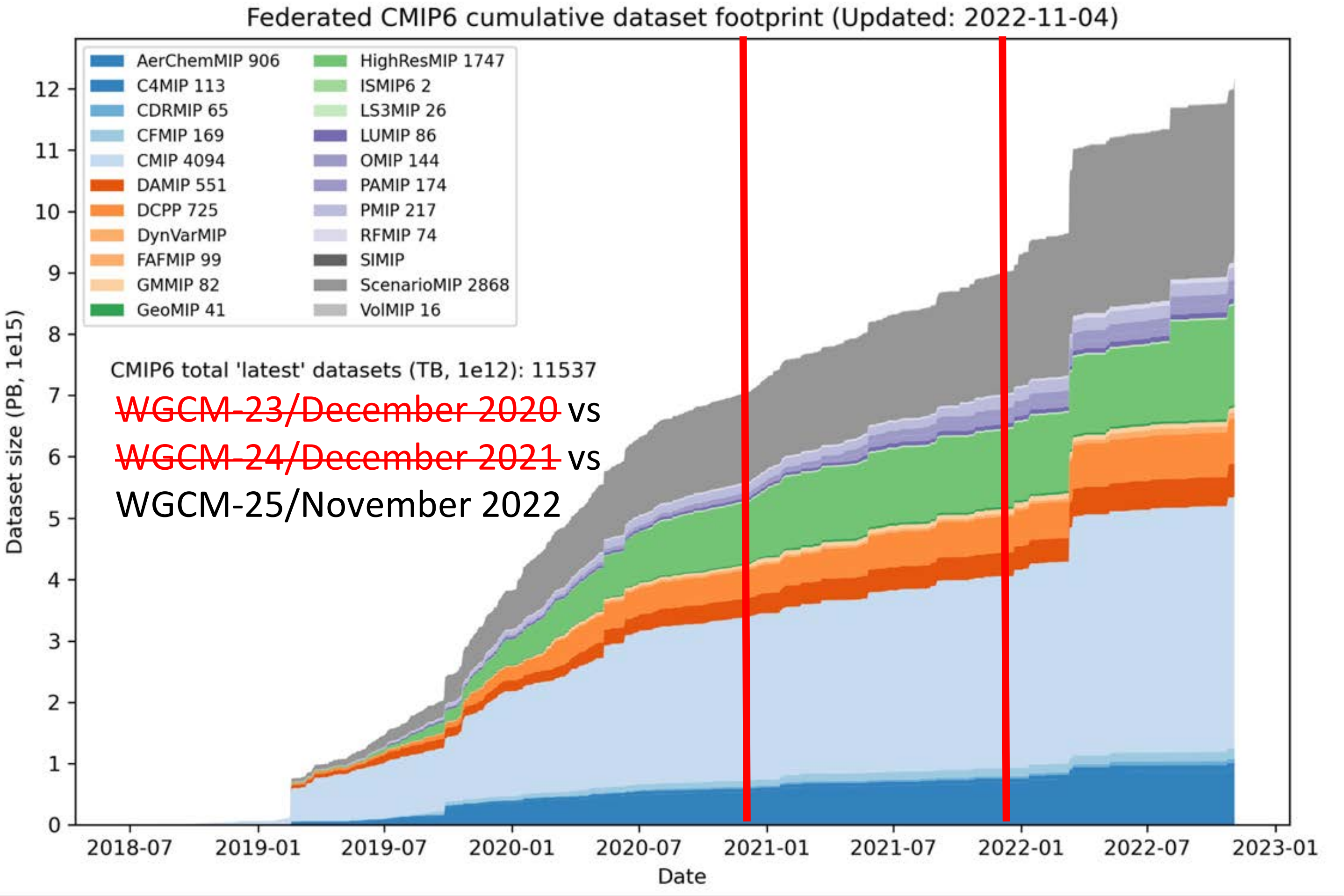
# CMIP: a team sport

- 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…

**Every modelling group, every forcing dataset provider, …**

# 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

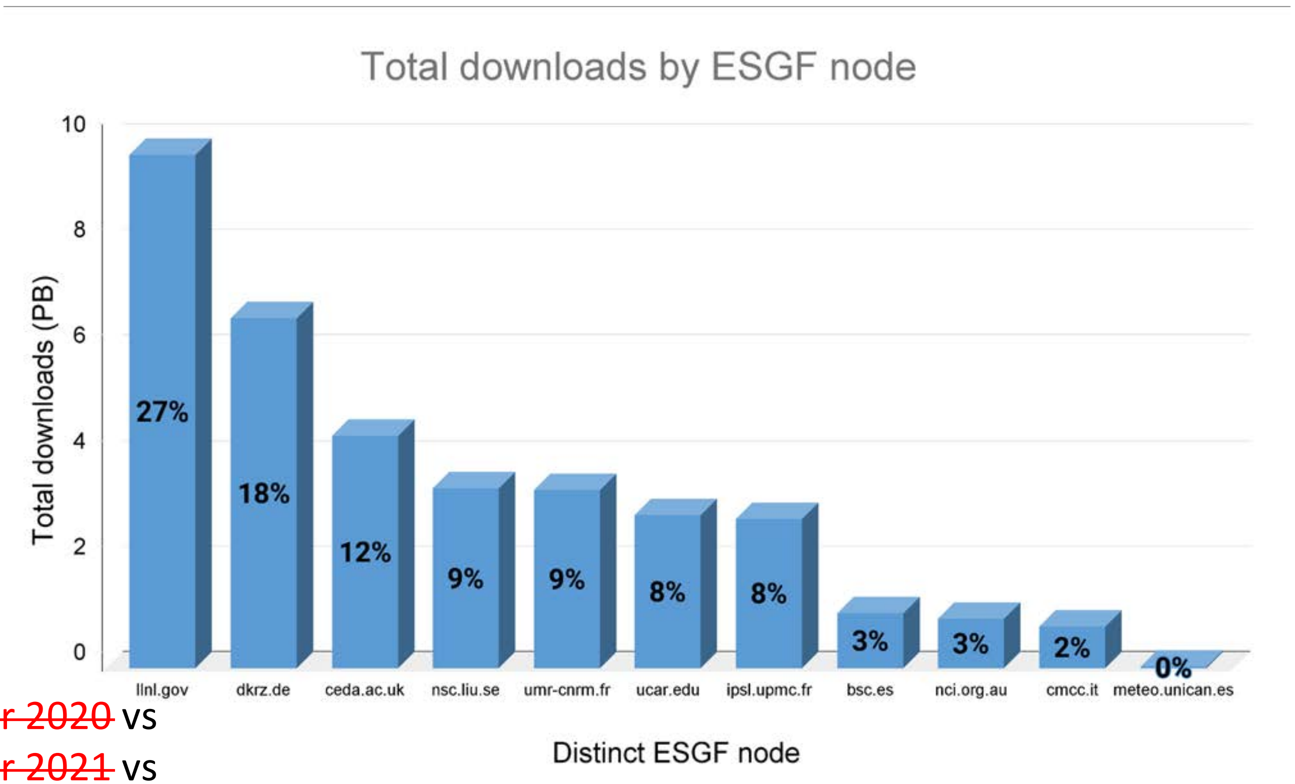

Federated CMIP6 cumulative dataset footprint (Updated: 2022-11-04)

CMIP6 total 'latest' datasets (TB, 1e12): 11537

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

# 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%

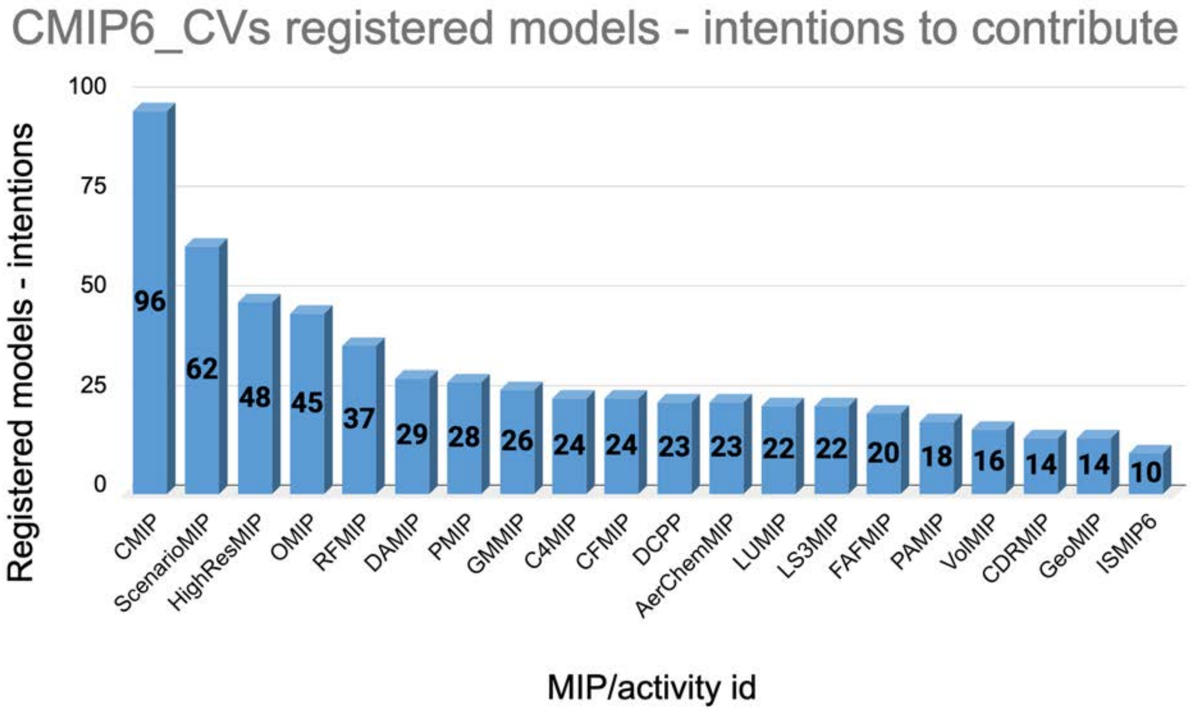


~~WGCM-23/December 2020~~ vs
~~WGCM-24/December 2021~~ vs
WGCM-25/November 2022

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

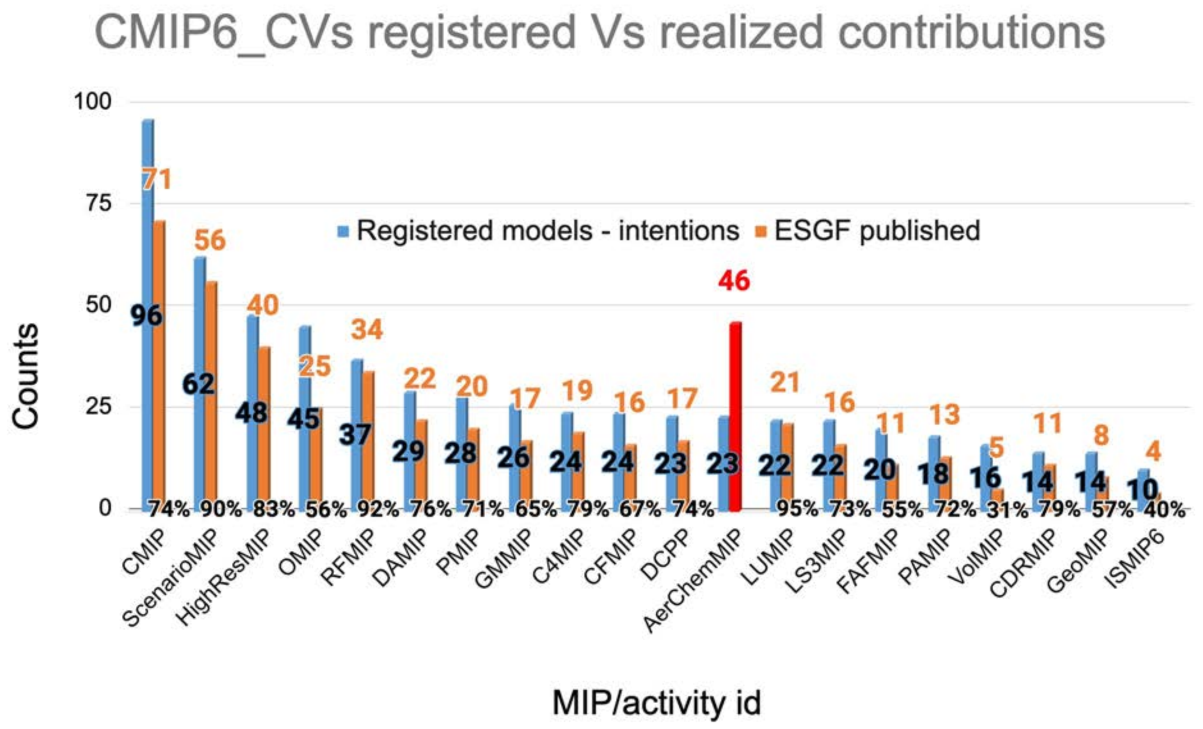


https://wcrp-cmip.github.io/CMIP6_CVs/docs/CMIP6_experiment_id.html
https://wcrp-cmip.github.io/CMIP6_CVs/docs/CMIP6_institution_id.html
https://wcrp-cmip.github.io/CMIP6_CVs/docs/CMIP6_source_id.html

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

# CMIP6 controlled vocabulary

- ~~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

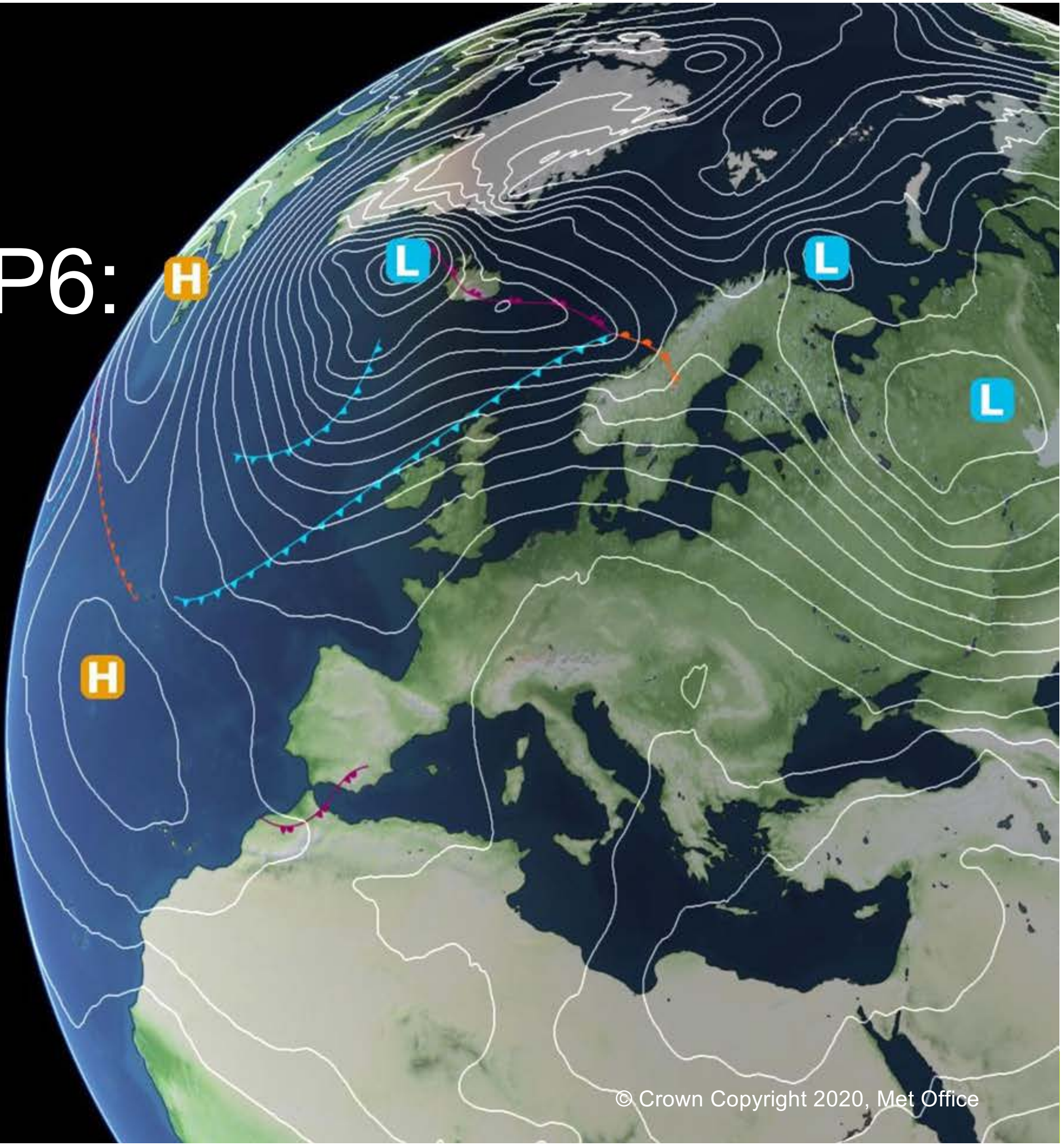


https://wcrp-cmip.github.io/CMIP6_CVs/docs/CMIP6_experiment_id.html
https://wcrp-cmip.github.io/CMIP6_CVs/docs/CMIP6_institution_id.html
https://wcrp-cmip.github.io/CMIP6_CVs/docs/CMIP6_source_id.html

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

Met Office
Hadley Centre

# A first step beyond CMIP6: CMIP6Plus



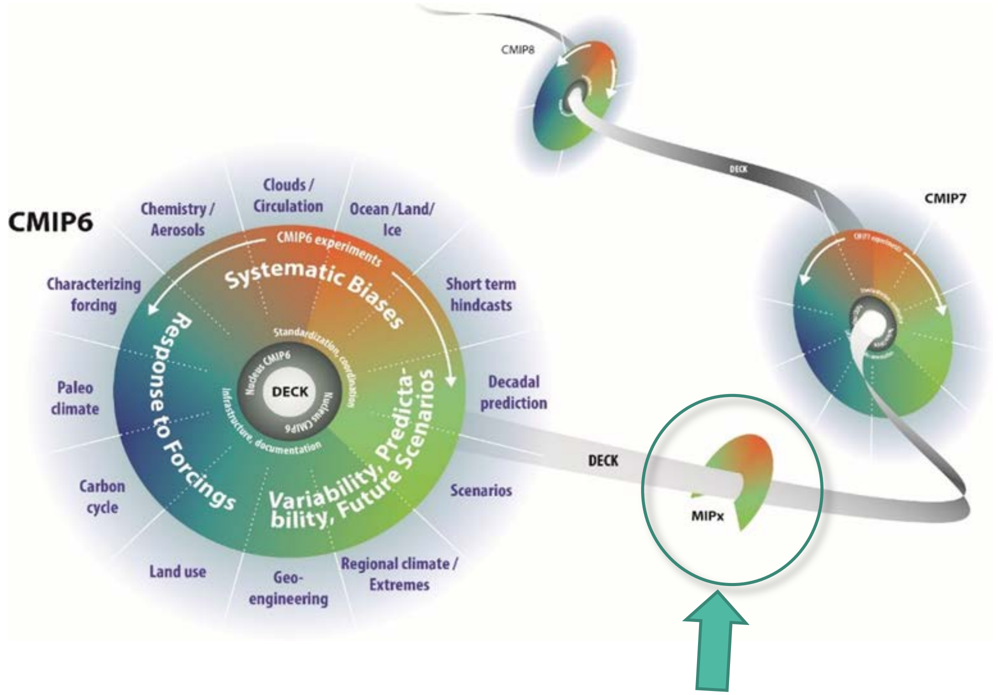
www.metoffice.gov.uk

Met Office
Hadley Centre

CMIP6+
Coupled Model Intercomparison Project next phase

# 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?



CMIP6+
Coupled Model Intercomparison Project next phase

# 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

Met Office
Hadley Centre

# 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

Met Office
Hadley Centre

# 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**

CMIP
Coupled Model
Intercomparison
Project

# Looking towards CMIP7

# CMIP7 Task Teams

- 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.*

CMIP

WCRP World Climate Research Programme

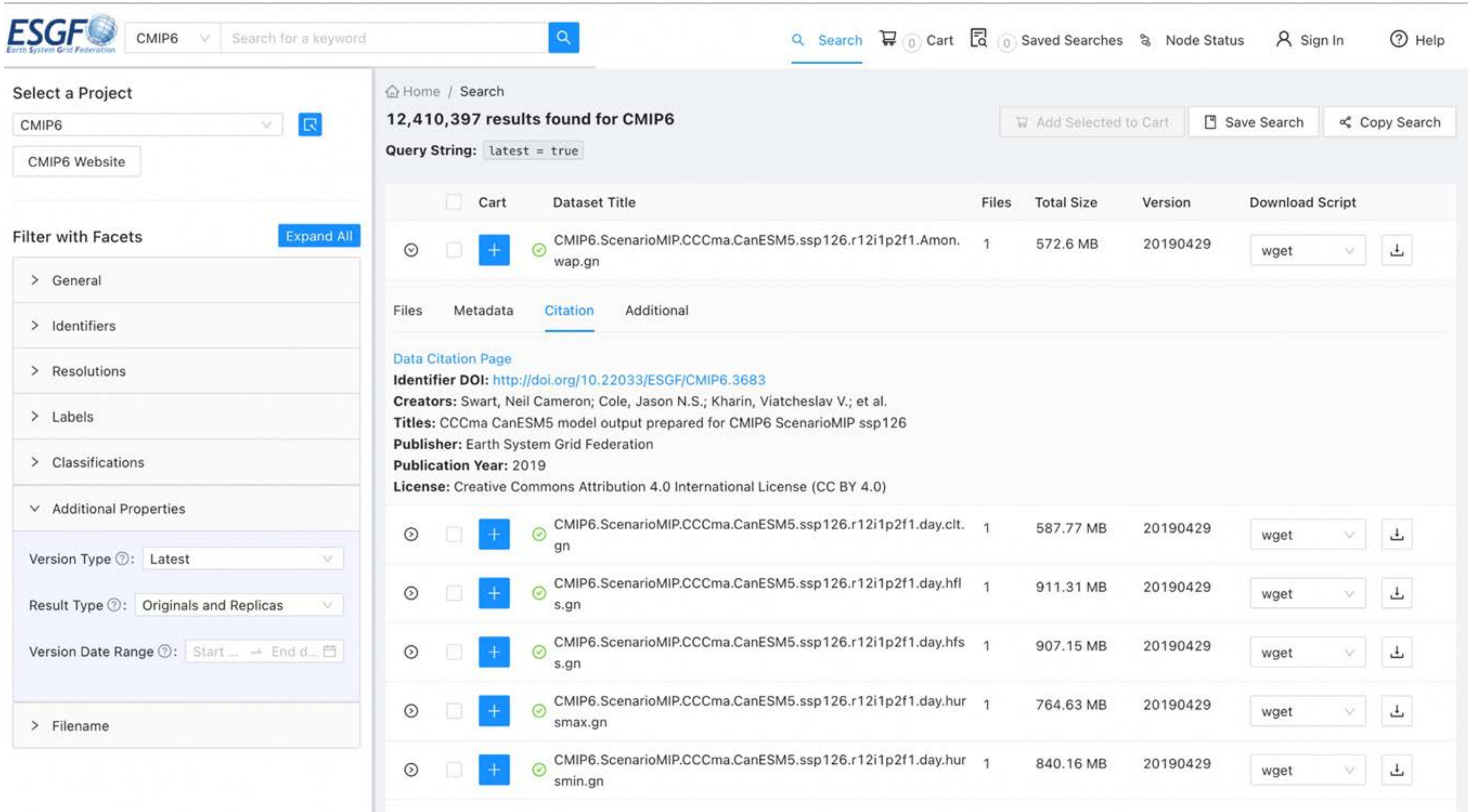
Paul J. Durack
durack1@llnl.gov

Matthew Mizielinski
matthew.mizielinski@metoffice.gov.uk

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

**Disclaimer**
This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

Lawrence Livermore
National Laboratory



Version Type: Latest

Result Type: Originals and Replicas

Version Date Range: Start ... → End d...

**12,410,397 results found for CMIP6**

**Query String:** `latest = true`

| Cart | Dataset Title | Files | Total Size | Version | Download Script |
|---|---|---|---|---|---|
| + | CMIP6.ScenarioMIP.CCCma.CanESM5.ssp126.r12i1p2f1.Amon.wap.gn | 1 | 572.6 MB | 20190429 | wget |

Data Citation Page

**Identifier DOI:** http://doi.org/10.22033/ESGF/CMIP6.3683

**Creators:** Swart, Neil Cameron; Cole, Jason N.S.; Kharin, Viatcheslav V.; et al.

**Titles:** CCCma CanESM5 model output prepared for CMIP6 ScenarioMIP ssp126

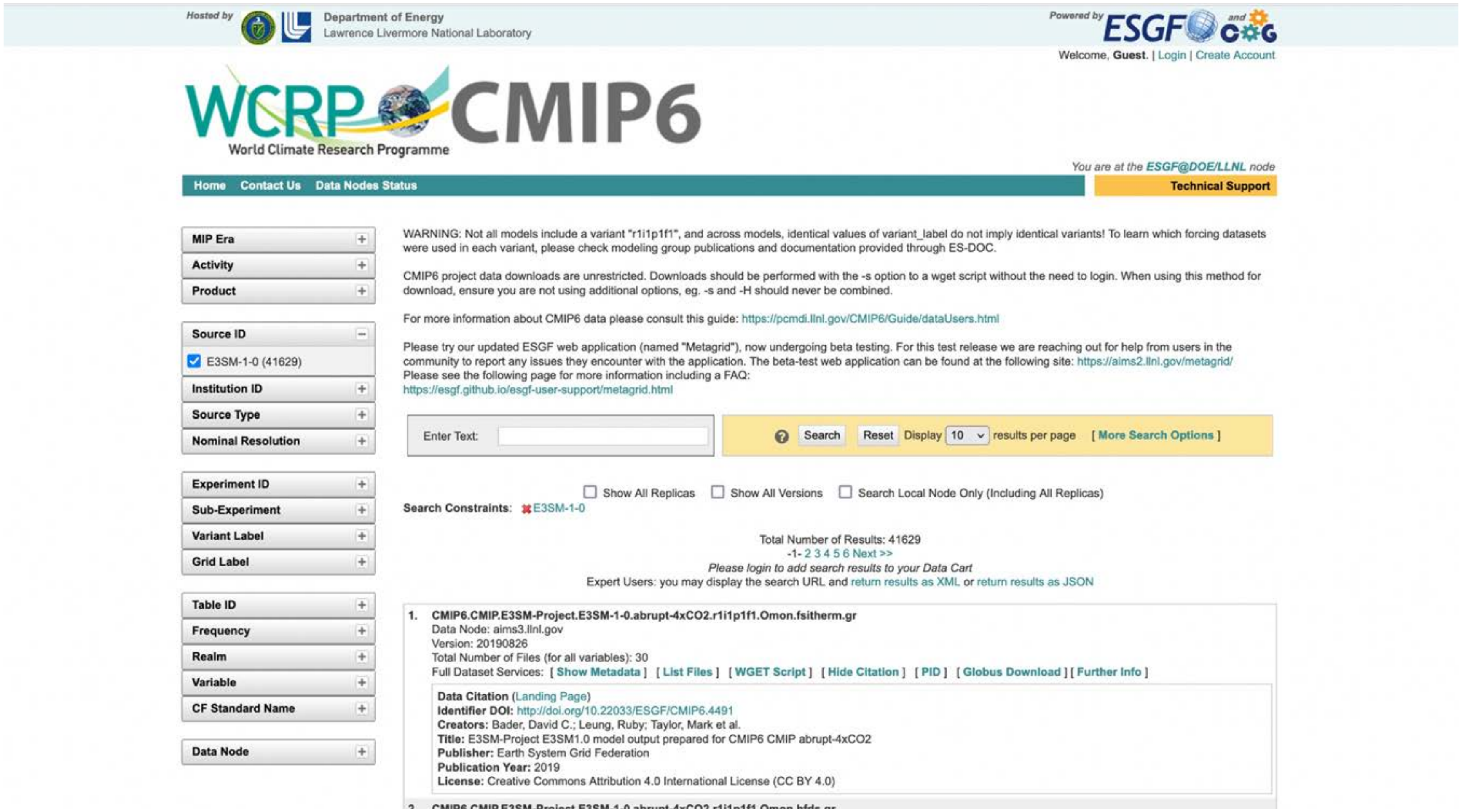
**Publisher:** Earth System Grid Federation

**Publication Year:** 2019

**License:** Creative Commons Attribution 4.0 International License (CC BY 4.0)

| Cart | Dataset Title | Files | Total Size | Version | Download Script |
|---|---|---|---|---|---|
| + | CMIP6.ScenarioMIP.CCCma.CanESM5.ssp126.r12i1p2f1.day.clt.gn | 1 | 587.77 MB | 20190429 | wget |
| + | CMIP6.ScenarioMIP.CCCma.CanESM5.ssp126.r12i1p2f1.day.hfls.gn | 1 | 911.31 MB | 20190429 | wget |
| + | CMIP6.ScenarioMIP.CCCma.CanESM5.ssp126.r12i1p2f1.day.hfss.gn | 1 | 907.15 MB | 20190429 | wget |
| + | CMIP6.ScenarioMIP.CCCma.CanESM5.ssp126.r12i1p2f1.day.hursmax.gn | 1 | 764.63 MB | 20190429 | wget |
| + | CMIP6.ScenarioMIP.CCCma.CanESM5.ssp126.r12i1p2f1.day.hursmin.gn | 1 | 840.16 MB | 20190429 | wget |

# CMIP



☑ E3SM-1-0 (41629)

WARNING: Not all models include a variant "r1i1p1f1", and across models, identical values of variant_label do not imply identical variants! To learn which forcing datasets were used in each variant, please check modeling group publications and documentation provided through ES-DOC.

CMIP6 project data downloads are unrestricted. Downloads should be performed with the -s option to a wget script without the need to login. When using this method for download, ensure you are not using additional options, eg. -s and -H should never be combined.

For more information about CMIP6 data please consult this guide: https://pcmdi.llnl.gov/CMIP6/Guide/dataUsers.html

Please try our updated ESGF web application (named "Metagrid"), now undergoing beta testing. For this test release we are reaching out for help from users in the community to report any issues they encounter with the application. The beta-test web application can be found at the following site: https://aims2.llnl.gov/metagrid/
Please see the following page for more information including a FAQ:
https://esgf.github.io/esgf-user-support/metagrid.html

Enter Text:

Display 10 results per page

□ Show All Replicas □ Show All Versions □ Search Local Node Only (Including All Replicas)

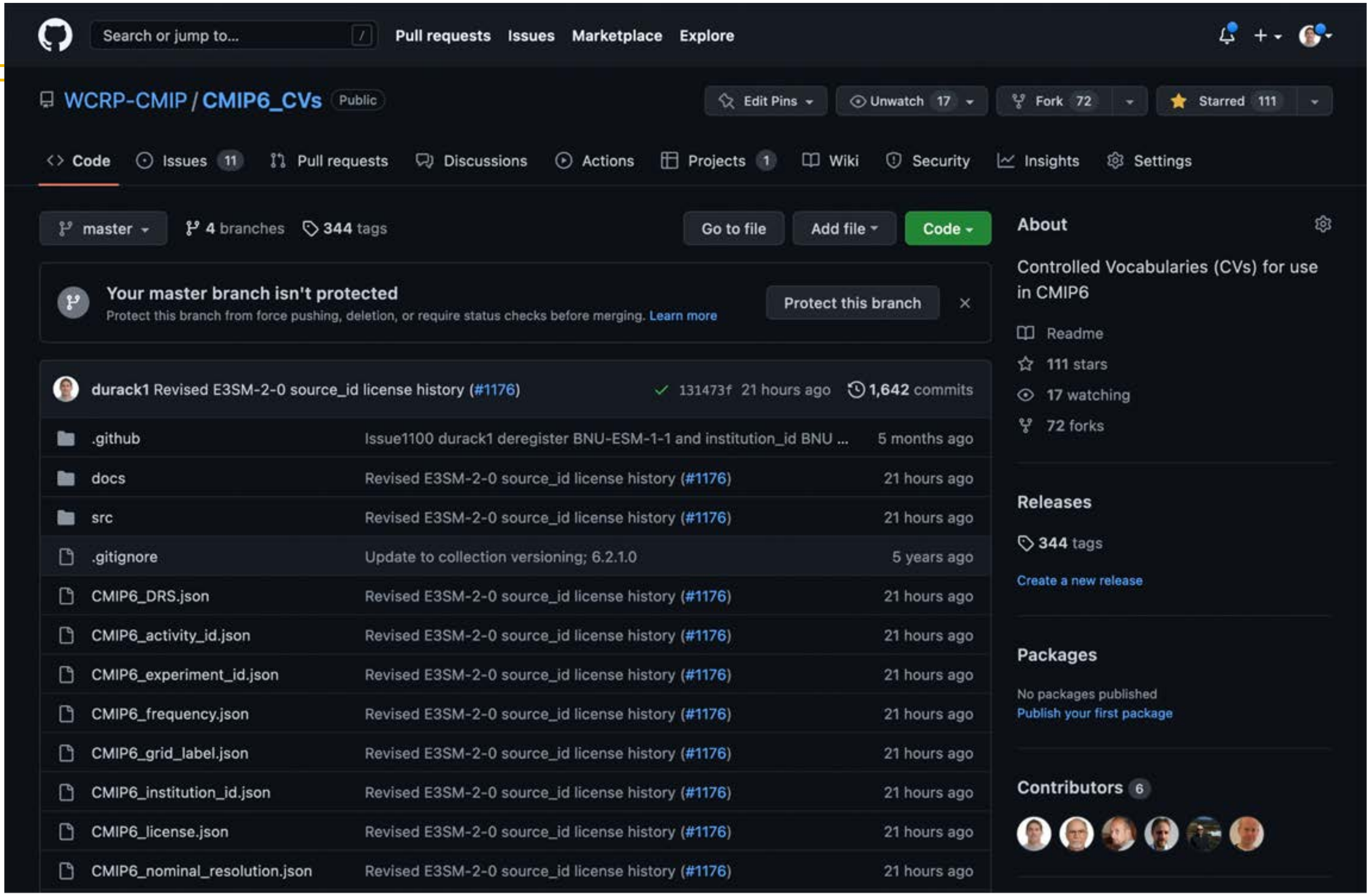
Search Constraints: ✖ E3SM-1-0

Total Number of Results: 41629
-1- 2 3 4 5 6 Next >>
*Please login to add search results to your Data Cart*
Expert Users: you may display the search URL and return results as XML or return results as JSON

1. **CMIP6.CMIP.E3SM-Project.E3SM-1-0.abrupt-4xCO2.r1i1p1f1.Omon.fsitherm.gr**
Data Node: aims3.llnl.gov
Version: 20190826
Total Number of Files (for all variables): 30
Full Dataset Services: [ Show Metadata ] [ List Files ] [ WGET Script ] [ Hide Citation ] [ PID ] [ Globus Download ] [ Further Info ]

**Data Citation** (Landing Page)
**Identifier DOI:** http://doi.org/10.22033/ESGF/CMIP6.4491
**Creators:** Bader, David C.; Leung, Ruby; Taylor, Mark et al.
**Title:** E3SM-Project E3SM1.0 model output prepared for CMIP6 CMIP abrupt-4xCO2
**Publisher:** Earth System Grid Federation
**Publication Year:** 2019
**License:** Creative Commons Attribution 4.0 International License (CC BY 4.0)

2. CMIP6.CMIP.E3SM-Project.E3SM-1-0.abrupt-4xCO2.r1i1p1f1.Omon.hfds.gr



WCRP-CMIP / CMIP6_CVs

**Your master branch isn't protected**

Protect this branch from force pushing, deletion, or require status checks before merging. Learn more

durack1 Revised E3SM-2-0 source_id license history (#1176) ✓ 131473f 21 hours ago 1,642 commits

| | | |
|---|---|---|
| .github | Issue1100 durack1 deregister BNU-ESM-1-1 and institution_id BNU ... | 5 months ago |
| docs | Revised E3SM-2-0 source_id license history (#1176) | 21 hours ago |
| src | Revised E3SM-2-0 source_id license history (#1176) | 21 hours ago |
| .gitignore | Update to collection versioning; 6.2.1.0 | 5 years ago |
| CMIP6_DRS.json | Revised E3SM-2-0 source_id license history (#1176) | 21 hours ago |
| CMIP6_activity_id.json | Revised E3SM-2-0 source_id license history (#1176) | 21 hours ago |
| CMIP6_experiment_id.json | Revised E3SM-2-0 source_id license history (#1176) | 21 hours ago |
| CMIP6_frequency.json | Revised E3SM-2-0 source_id license history (#1176) | 21 hours ago |
| CMIP6_grid_label.json | Revised E3SM-2-0 source_id license history (#1176) | 21 hours ago |
| CMIP6_institution_id.json | Revised E3SM-2-0 source_id license history (#1176) | 21 hours ago |
| CMIP6_license.json | Revised E3SM-2-0 source_id license history (#1176) | 21 hours ago |
| CMIP6_nominal_resolution.json | Revised E3SM-2-0 source_id license history (#1176) | 21 hours ago |

**About**

Controlled Vocabularies (CVs) for use in CMIP6

Readme

111 stars

17 watching

72 forks

**Releases**

344 tags

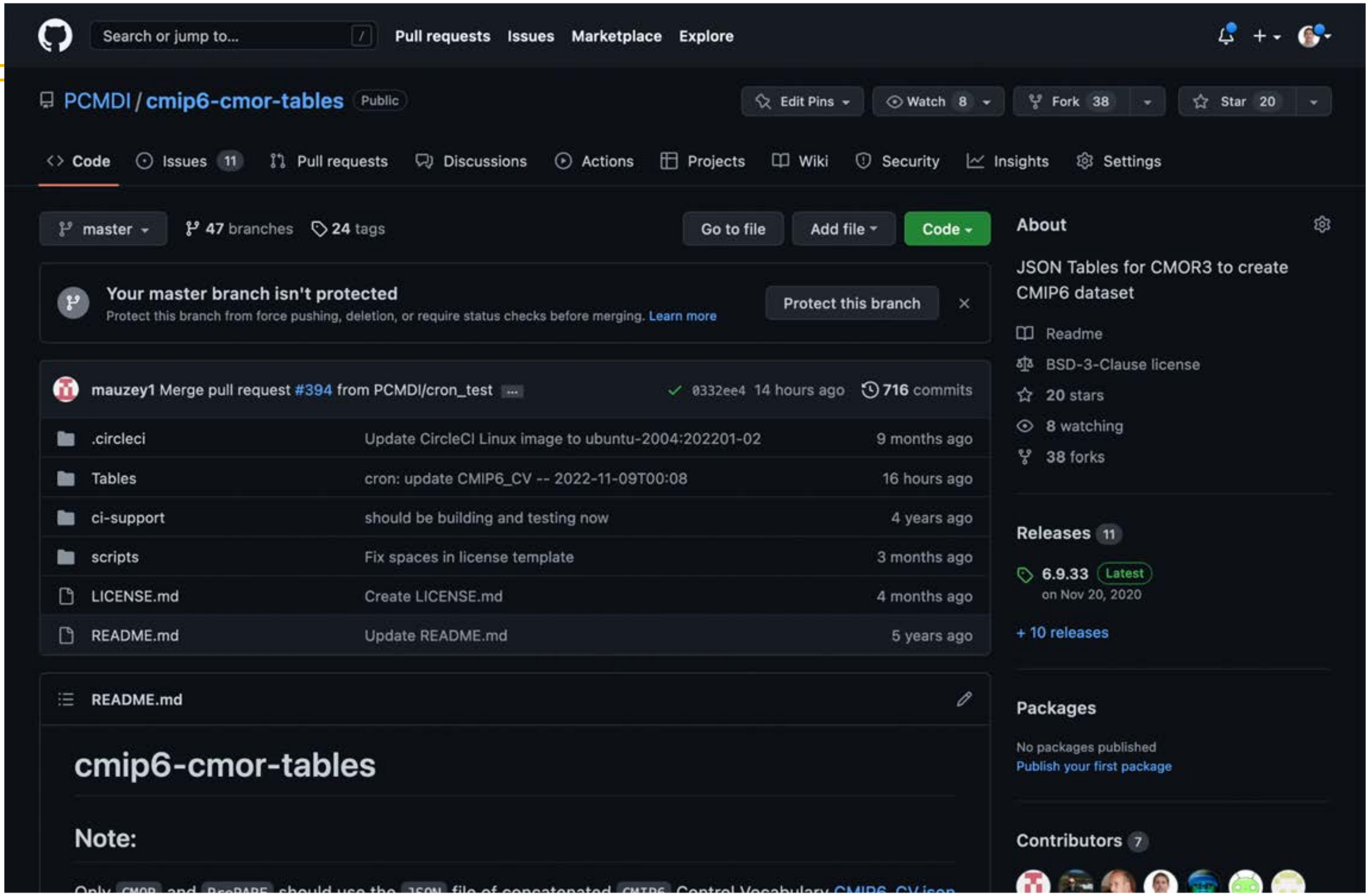
Create a new release

**Packages**

No packages published

Publish your first package

**Contributors** 6



PCMDI / cmip6-cmor-tables Public

**Your master branch isn't protected**

Protect this branch from force pushing, deletion, or require status checks before merging. Learn more

mauzey1 Merge pull request #394 from PCMDI/cron_test ... ✓ 0332ee4 14 hours ago 716 commits

| | | |
|---|---|---|
| .circleci | Update CircleCI Linux image to ubuntu-2004:202201-02 | 9 months ago |
| Tables | cron: update CMIP6_CV -- 2022-11-09T00:08 | 16 hours ago |
| ci-support | should be building and testing now | 4 years ago |
| scripts | Fix spaces in license template | 3 months ago |
| LICENSE.md | Create LICENSE.md | 4 months ago |
| README.md | Update README.md | 5 years ago |

README.md

# cmip6-cmor-tables

## Note:

Only `CMOR` and `PrePARE` should use the `JSON` file of concatenated `CMIP6` Control Vocabulary CMIP6_CV.json

**About**

JSON Tables for CMOR3 to create CMIP6 dataset

- Readme
- BSD-3-Clause license
- 20 stars
- 8 watching
- 38 forks

**Releases** 11

6.9.33 Latest
on Nov 20, 2020

+ 10 releases

**Packages**

No packages published
Publish your first package

**Contributors** 7

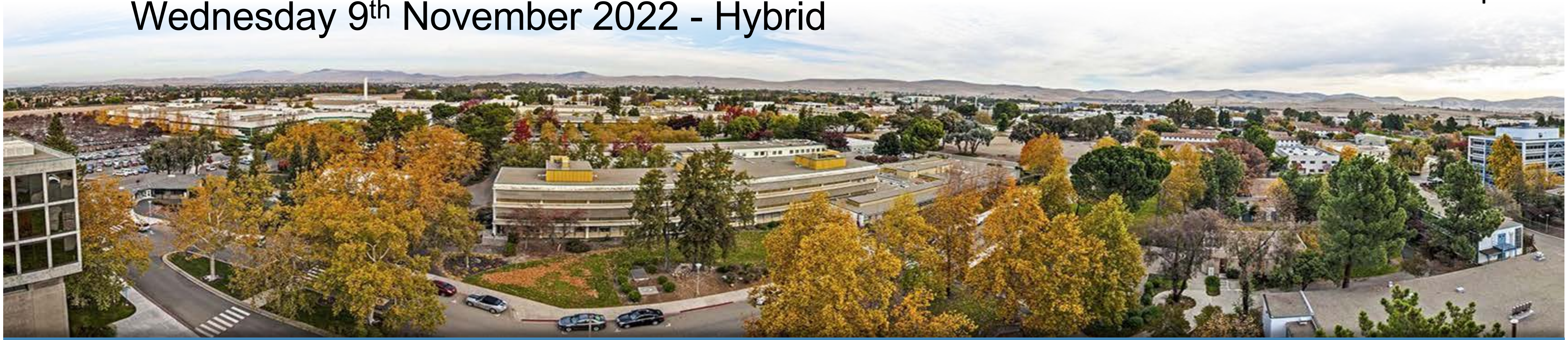
# CMIP immediate future?

## 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



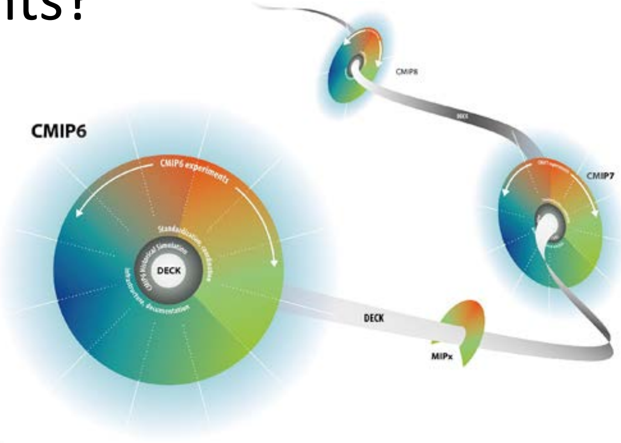
LLNL-PRES-842123
This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

Lawrence Livermore National Laboratory

# CMIP6+

CMIP6+
Coupled Model Intercomparison Project next phase

## How can we leverage existing infrastructure investments?


CMIP8
DECK
CMIP6
CMIP6 experiments
CMIP7
Standardization, coordination
CMIP6 Historical Simulation
DECK
Infrastructure, documentation
DECK
MIPx

- 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

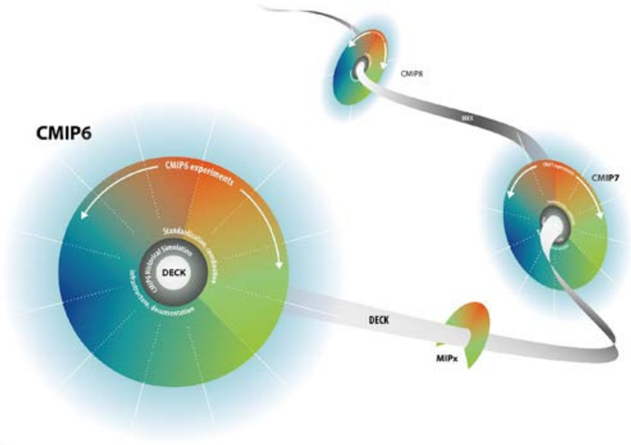
# CMIP6+

Coupled Model Intercomparison Project next phase

## How can we leverage existing communities?

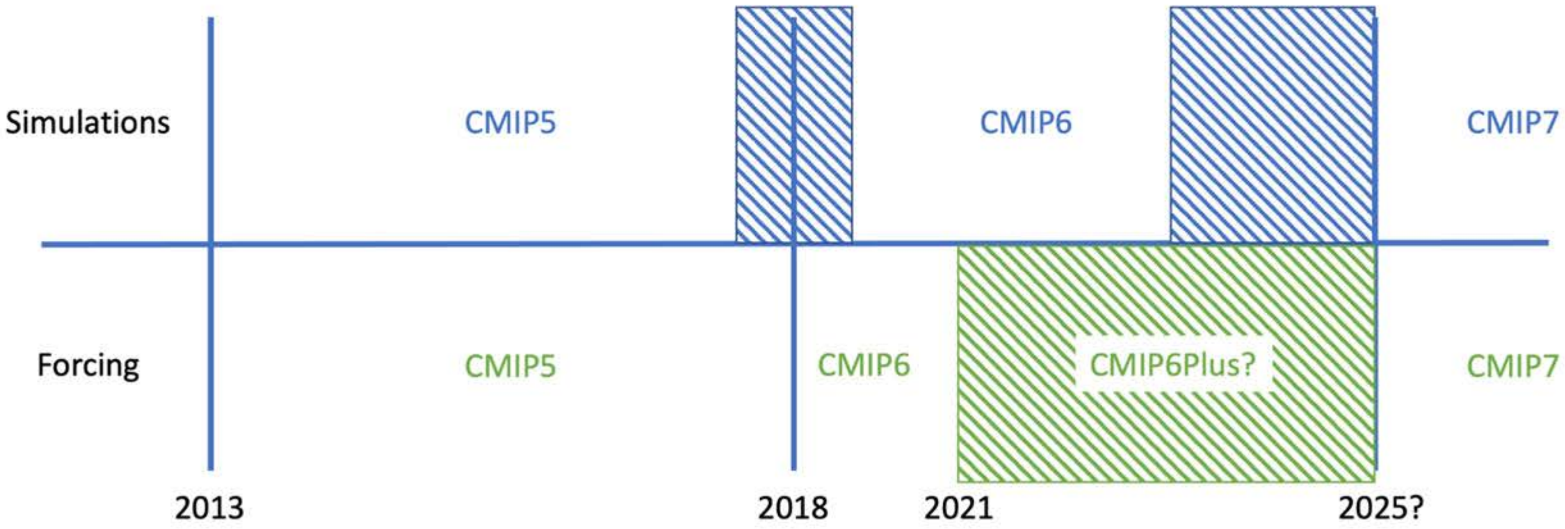
- Continue beyond CMIP6, few changes as CMIP7 is discussed and planned
  - Reduce time pressures - loosen CMIP*x* - IPCC AR*x* linkage
  - Continuous “CMIP science” - not monolithic phases every ~7 years

- 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

input4MIPs

input datasets for Model Intercomparison Projects

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 realtime (~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

# input4MIPs

input4MIPs

input datasets for Model Intercomparison Projects

**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 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, …)?

Paul J. Durack
durack1@llnl.gov

Peter J. Gleckler
gleckler1@llnl.gov

Matthew Mizielinski
matthew.mizielinski@metoffice.gov.uk

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

Lawrence Livermore National Laboratory

**Disclaimer**
This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

# Harmonizing CMIP Data Holdings Across Phases and Activities

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

WGCM-24
Thursday 9th December 2021 - Virtual

LLNL-PRES-842123
This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

Lawrence Livermore National Laboratory

# 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)

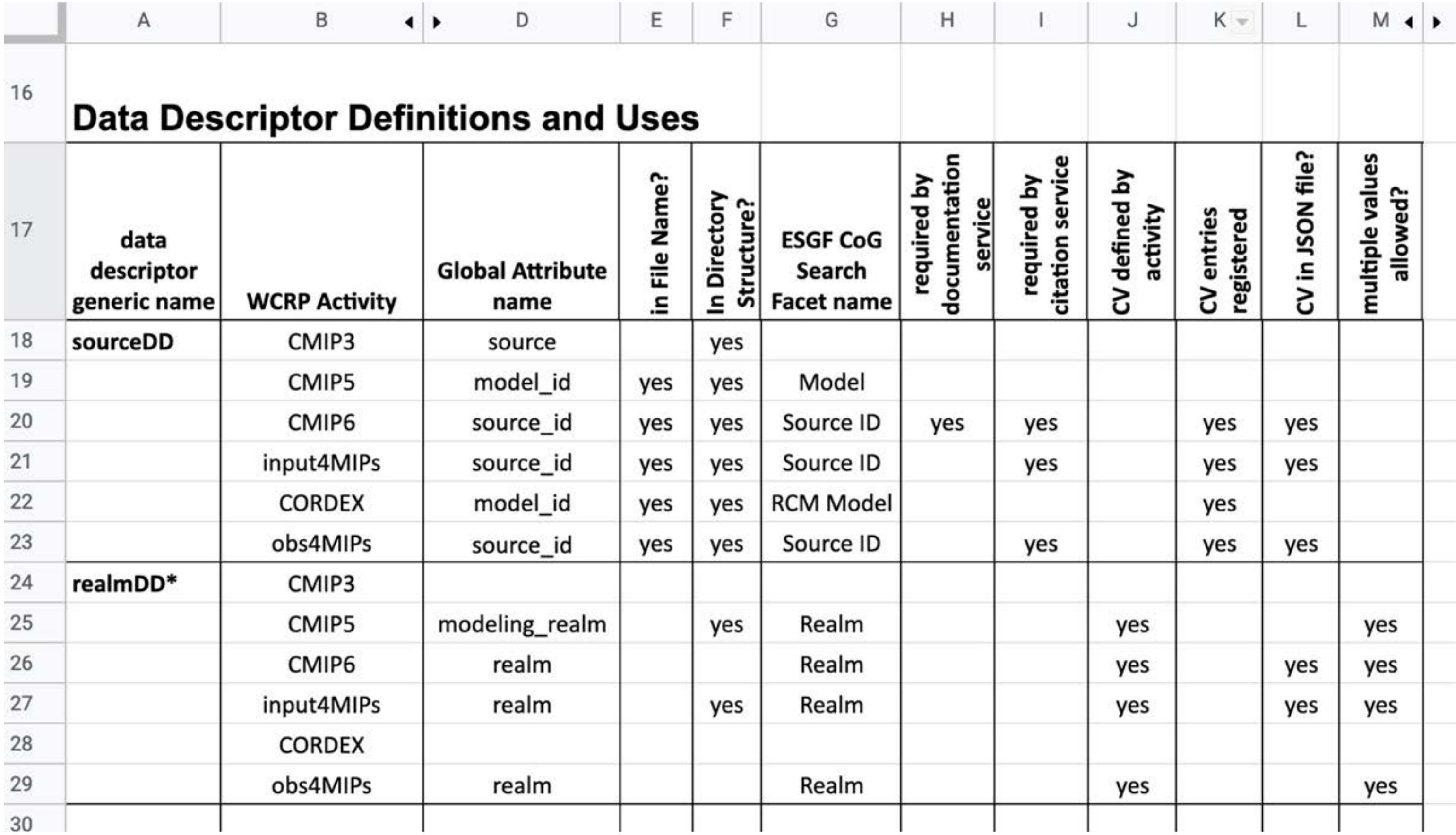
We could facilitate research by harmonizing
the archive across generations!

# 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

2 examples of data descriptors



| | A | B | D | E | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 16 | **Data Descriptor Definitions and Uses** | | | | | | | | | | | |
| 17 | **data descriptor generic name** | **WCRP Activity** | **Global Attribute name** | **in File Name?** | **In Directory Structure?** | **ESGF CoG Search Facet name** | **required by documentation service** | **required by citation service** | **CV defined by activity** | **CV entries registered** | **CV in JSON file?** | **multiple values allowed?** |
| 18 | **sourceDD** | CMIP3 | source | | yes | | | | | | | |
| 19 | | CMIP5 | model_id | yes | yes | Model | | | | | | |
| 20 | | CMIP6 | source_id | yes | yes | Source ID | yes | yes | | yes | yes | |
| 21 | | input4MIPs | source_id | yes | yes | Source ID | | yes | | yes | yes | |
| 22 | | CORDEX | model_id | yes | yes | RCM Model | | | | yes | | |
| 23 | | obs4MIPs | source_id | yes | yes | Source ID | | yes | | yes | yes | |
| 24 | **realmDD*** | CMIP3 | | | | | | | | | | |
| 25 | | CMIP5 | modeling_realm | | yes | Realm | | | yes | | | yes |
| 26 | | CMIP6 | realm | | | Realm | | | yes | | yes | yes |
| 27 | | input4MIPs | realm | | yes | Realm | | | yes | | yes | yes |
| 28 | | CORDEX | | | | | | | | | | |
| 29 | | obs4MIPs | realm | | | Realm | | | yes | | | yes |
| 30 | | | | | | | | | | | | |

# 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 scenarioMIP 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)

Karl E. Taylor
taylor13@llnl.gov

Paul J. Durack
durack1@llnl.gov

Matthew Mizielinski
matthew.mizielinski@metoffice.gov.uk

PCMDI project work is funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Regional and Global Model Analysis Program

Lawrence Livermore National Laboratory

**Disclaimer**
This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.