CMIP6:

Infrastructure, data and coordination

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

WGCM-25
Thursday 10th November 2022 - Hybrid

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

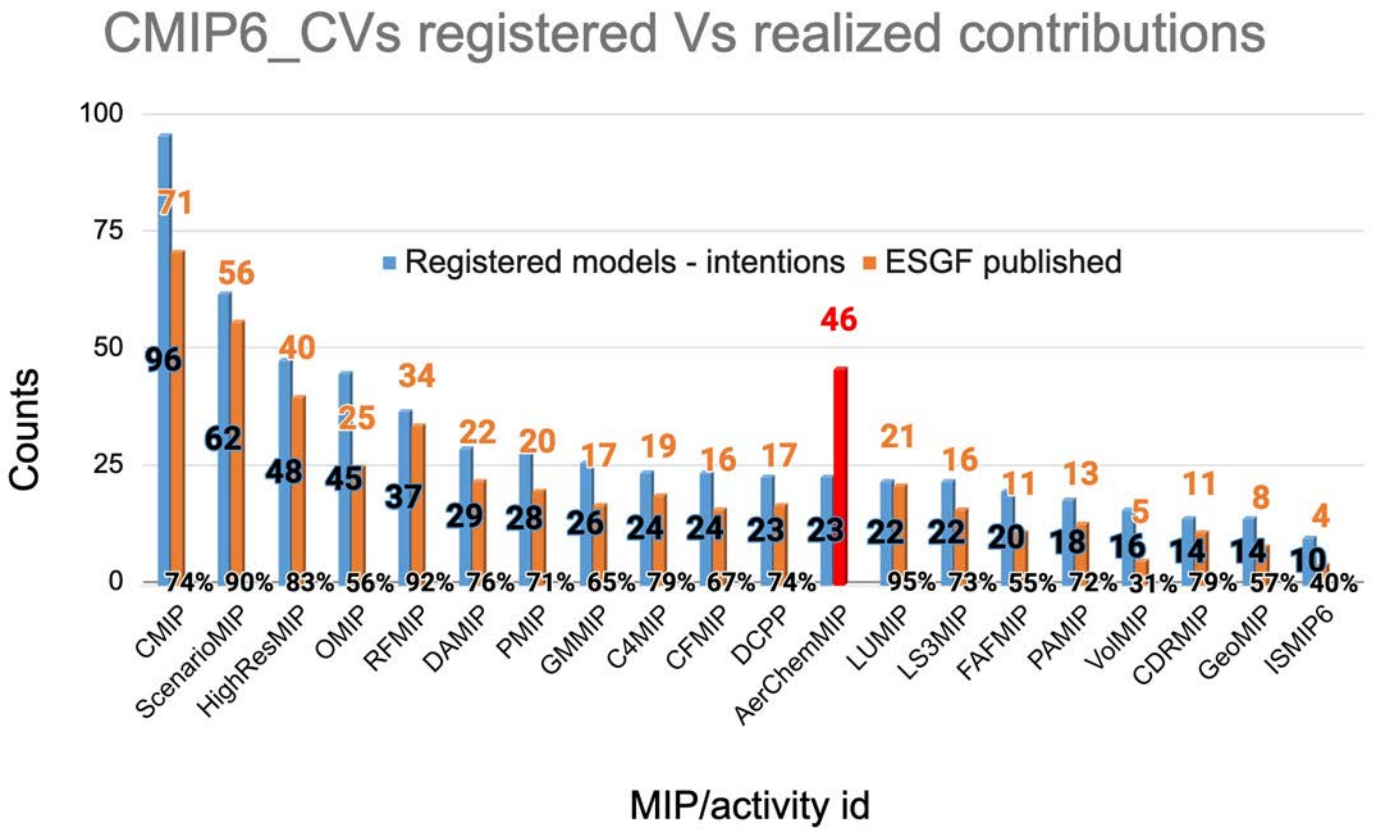
LLNL-PRES-842123
This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

Lawrence Livermore National Laboratory

Survey highlights

- Timeliness
  - Data request
  - Forcing errors and delays
  - IPCC deadlines strained ability for data use
  - Burden of 24 MIPs - need to stress DECK is only compulsory
- Desire for mechanism to support new MIPs
- Core MIPs expected for CMIP7 - critical review of arrange satellite MIPs
- Reuse infrastructure - reduce overheads to entry
- Potential ‘core’ MIPs aligned with IPCC with less centralized coordination of specialist MIPs that could be decoupled from the IPCC timeline - CMIP6Plus
- Website task team - aid documentation

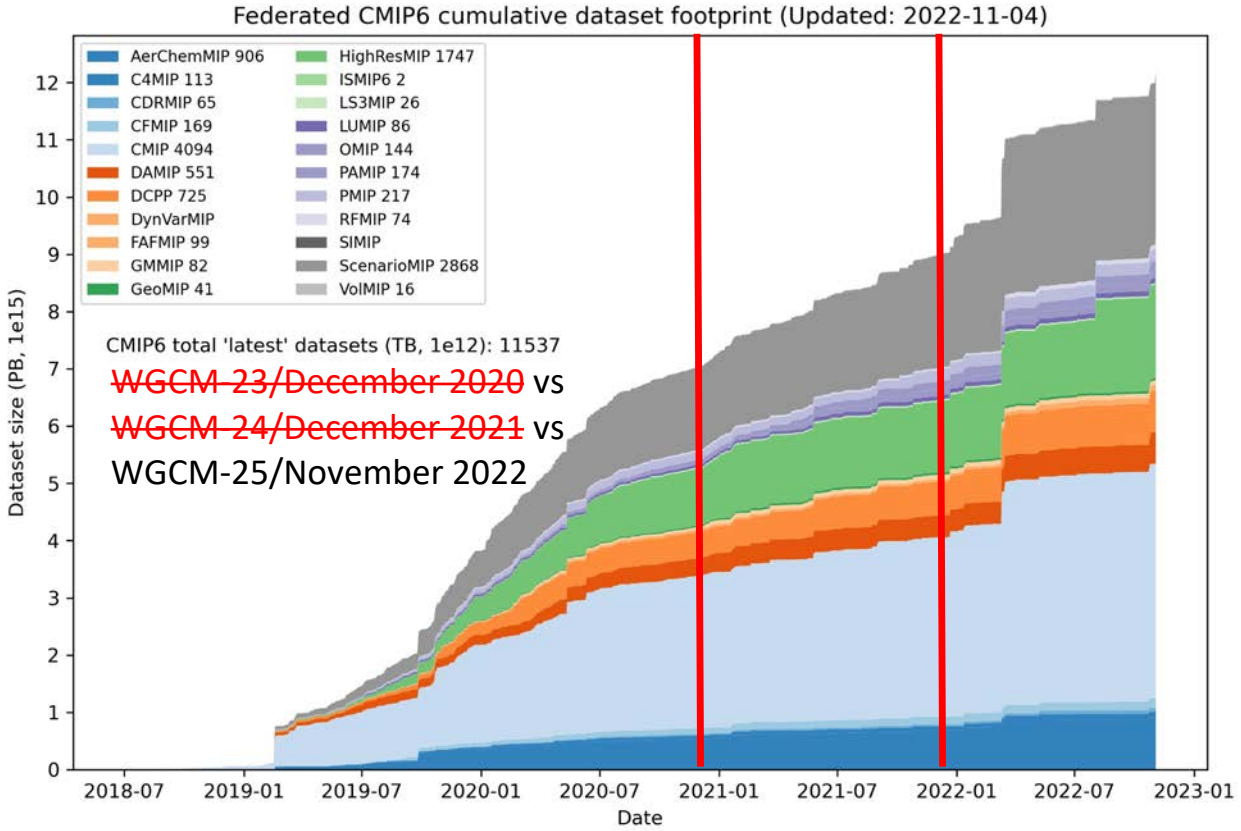
CMIP6 delivery



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

ESGF Published data

- 6.4 million datasets on ESGF across all CMIP6 activities/MIPs
- Delivery has been seamless – thanks to data challenges and ESGF stability testing
- Footprint – storage units in PBs



Licensing update

All models relaxed to a simple CC-BY 4.0 attribution license.

Details on WCRP-CMIP/CMIP6_CV github

WCRP-CMIP CMIP6_CVs version: 6.2.58.49

Show 10 entries

source id	institution id	release year	cohort	label	label extended	license	exceptions contact	history
4AOP-v1-5	IPSL	2019	Published	4AOP-v1-5	Line-By-Line Radiative Transfer Model v1.5, Laboratoire Meteorologie Dynamique, GEISA spectroscopic database	CC BY 4.0	@listes.ipsl.fr <-ipsl-cmip6	2020-06-11: initially published under CC B NC-SA 4.0; 2022-06-0 relaxed to CC BY 4.0
ACCESS-CM2	CSIRO-ARCCSS	2019	Published	ACCESS-CM2	Australian Community Climate and Earth System Simulator Climate Model Version 2	CC BY 4.0	@csiro.au <-access_csiro	2019-11-08: initially published under CC B SA 4.0; 2022-06-10: relaxed to CC BY 4.0
ACCESS-ESM1-5	CSIRO	2019	Published	ACCESS-ESM1.5	Australian Community Climate and Earth System Simulator Earth System Model Version 1.5	CC BY 4.0	@csiro.au <-access_csiro	2019-11-12: initially published under CC B SA 4.0; 2022-06-10: relaxed to CC BY 4.0
ACCESS-OM2	CSIRO-COSIMA	2020	Published	ACCESS-OM2	Australian Community Climate and Earth System Simulator Ocean Model Version 2	CC BY 4.0	@csiro.au <-access_csiro	2021-06-16: initially published under CC B SA 4.0; 2022-06-10: relaxed to CC BY 4.0
ACCESS-OM2-025	CSIRO-COSIMA	2020	Published	ACCESS-OM2-025	Australian Community Climate and Earth System Simulator Ocean Model Version 2 quarter degree	CC BY 4.0	@csiro.au <-access_csiro	2021-06-17: initially published under CC B SA 4.0; 2022-06-10: relaxed to CC BY 4.0
ARTS-2-3	UHH	2015	Published	ARTS 2.3	ARTS 2.3 (Current development version of the	CC BY 4.0	@uni-hamburg.de	2019-06-20: initially published under CC B

CMIP – Coupled Model Intercomparison Project's Post

CMIP – Coupled Model Intercomparison Project
43 followers
1w

We are delighted to announce that the World Climate Research Programme (#WCRP) #CMIP6 data license has been relaxed to CC BY 4.0 (Creative Commons Attribution; https://lnkd.in/e365Whs).

The #CMIP6 license relaxation enables a broader use the #CMIP archive, facilitating expanding #climate data usage across #research, #industry, and #policy communities, see https://lnkd.in/eYmHmS_8 for additional details.

We would like to thank the Working Group on Coupled Modelling (#WGCM) Infrastructure Panel co-chairs, Paul J. Durack and Matthew Mizielinski, and the #CMIP6 contributing modelling groups for being so helpful during this process.

CMIP – Coupled Model Intercomparison Project
@wcrpcmip

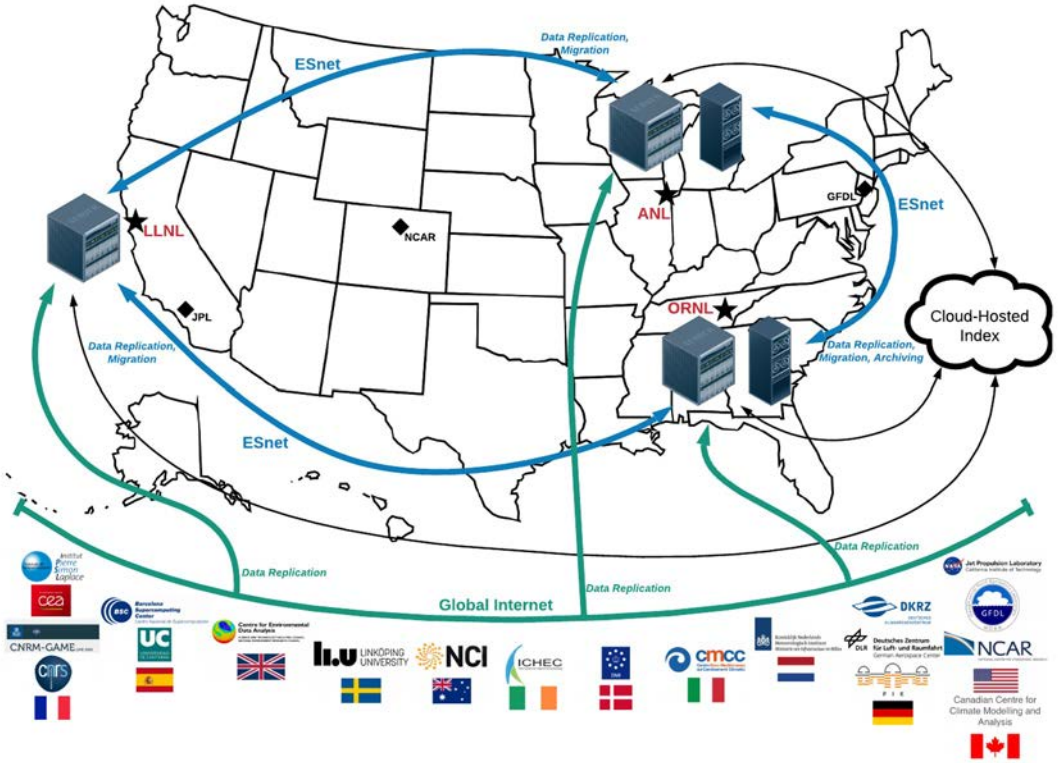
We're delighted to announce the #CMIP6 data license has been relaxed to CC BY 4.0 (Creative Commons Attribution; creativecommons.org/licenses/by/4....) 1/2

CMIP Coupled Model Intercomparison Project

ESGF - international federation

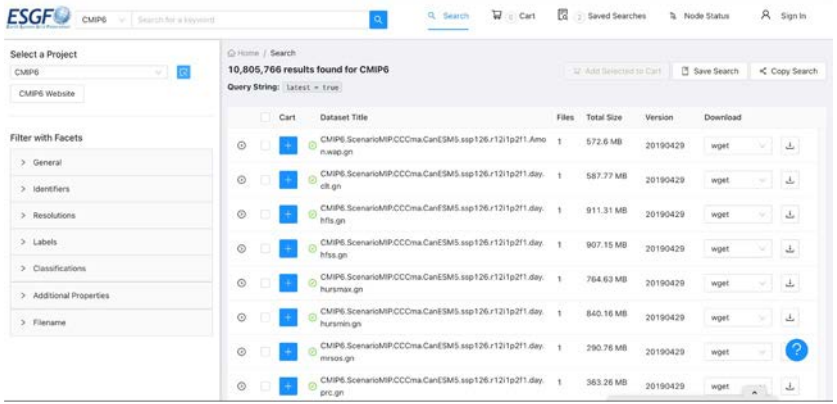
ESGF2 Project Launched, CMIP data copied

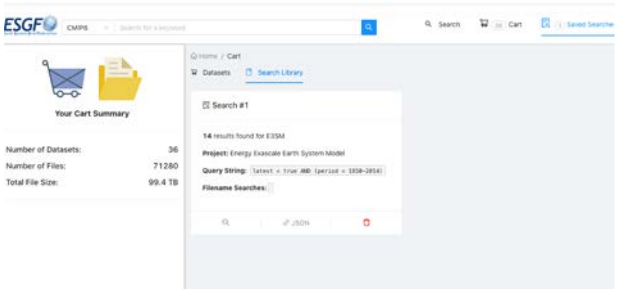
- Project under ORNL leadership restarted July 2022 once budget resolved
- 7.5 PB of CMIP data copied from LLNL to ORNL/ANL Feb-May 2022
- Work underway to relaunch compute platform hosted at LLNL
  - JupyterHub available attached to CMIP replica storage
- Successfully deployed test node at ORNL in OpenShift environment
  - Production status pending arrival of storage hardware and replica publication



Metagrid “Beta” UI Released to Community

- Metagrid web interface shared March 2022 ; gives us key feedback for improvements
- Google Analytics reports 250-300 new users each week (since mid-June)
- Includes online tutorial walkthrough of key functionality
- New function for users to save and share searches
- Goal to replace CoG (current web interface) mid 2023 pending Globus integration





ESGF2: Next-gen Index with Globus Search

Input form

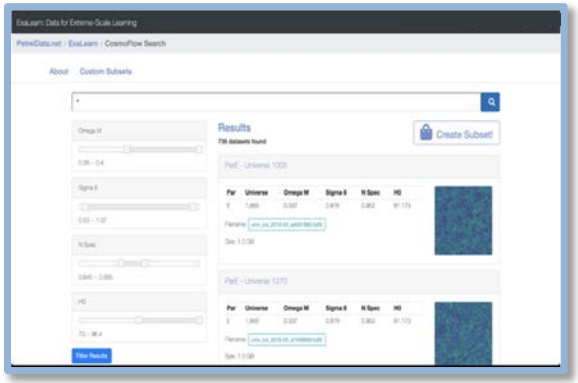
Automated Extraction

Bulk ingest

Globus Search

Ingest metadata, set visibility policies

Sample webapp



Enabling discoverability via Faceted Search

- Cloud-hosted index
- CMIP5/6 search records ingested for search demo

ESGF in Europe

- ESGF *Future Architecture* set out in 2019 after meeting of technical leads from ESGF partner organisations
  - https://doi.org/10.5281/zenodo.3928223

- Progress and achievements implementing Future Architecture:
  - *Container*-based deployment – live at CEDA, GFDL (on Amazon Web Services) and in test at partners DKRZ, ORNL and NCI
  - Implemented new centralised search service with community standard API (STAC)
  - Implemented new centralised identity service and simplified access control for CORDEX data using tokens (no complicated certificates!)

- Restarted Working Team to look at Compute Services
  - Developed services which allow sub-setting of data

ENES Research Infrastructure (ENES-RI)

Paid by membership fees

ENES-RI coordination

1 FTE - Sci. Officer
(2.2 FTE - Sci. officer + TF leads + technical leads + admin officer)

ENES Climate Data Infrastructure

12.5 FTE

- Governance activities
- Data Distribution, long term archival and user support
- Metadata
- Compute services

ENES Models, Tools, and HPC

5.8 FTE

Training and workshops

FTE tbc

Institutional in-kind / external projects

Model Evaluation

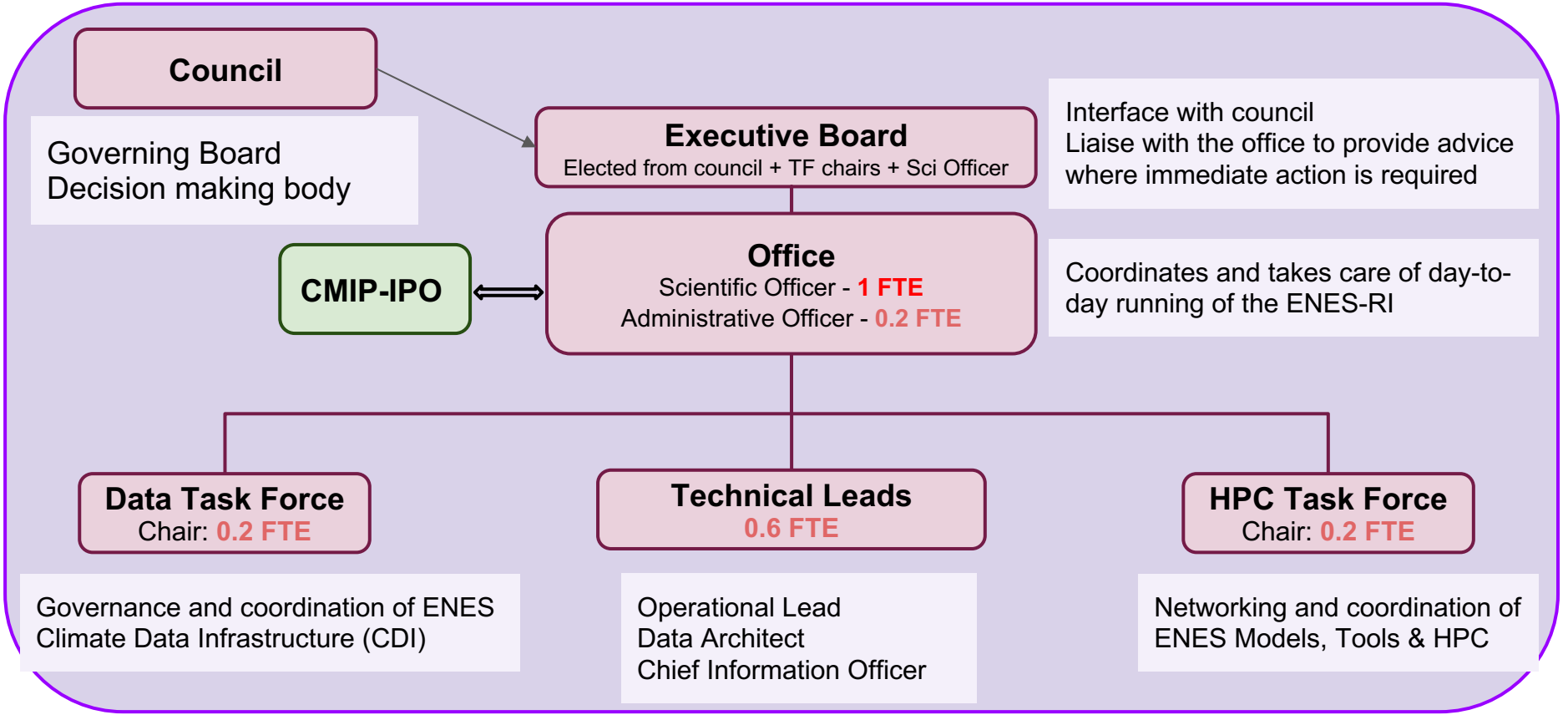
6 FTE

User engagement

FTE tbc

ENES-RI coordination **agreed**

Funded by partners



CMIP6 infrastructure activities

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

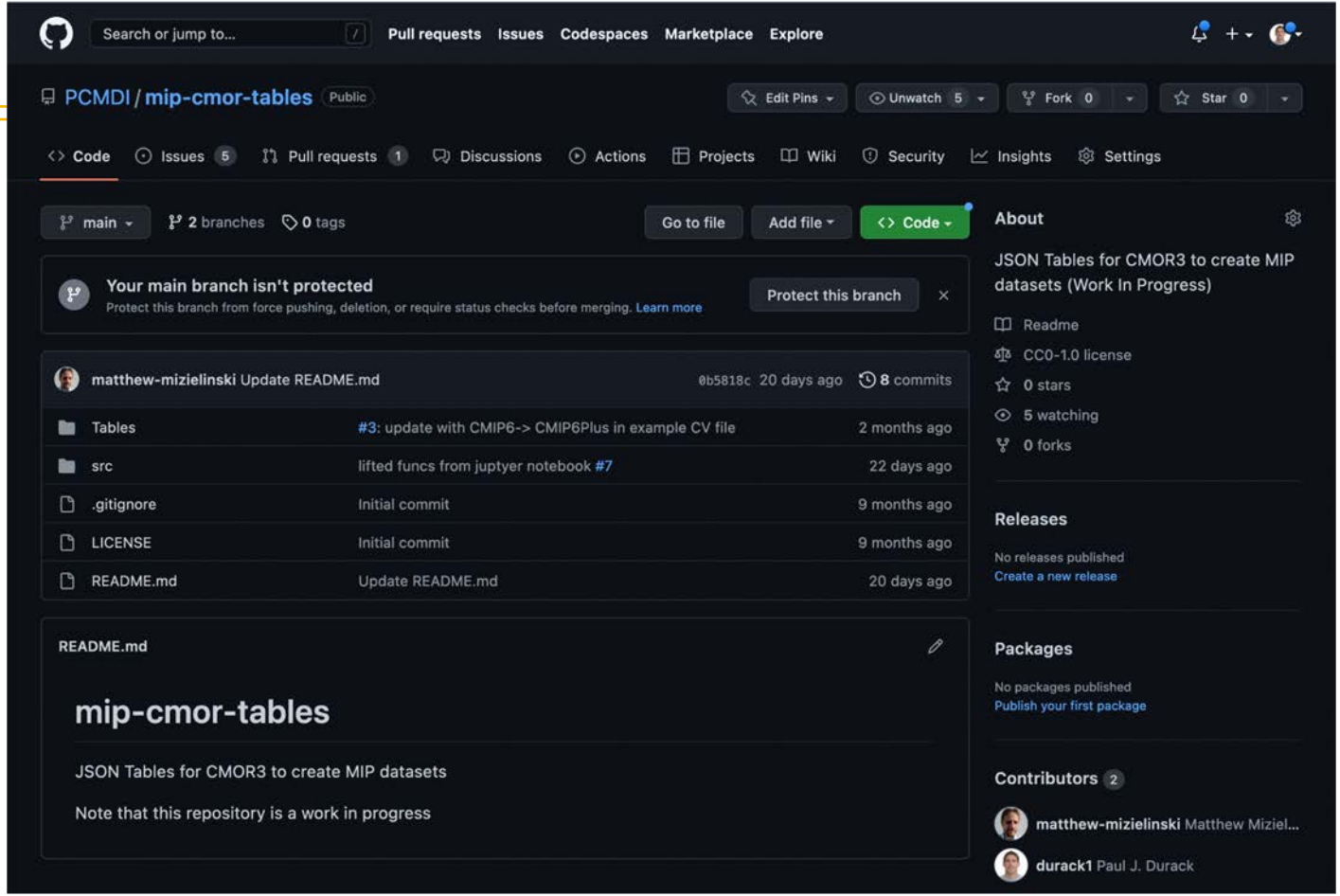
Projects

 # one repo per project
CMIP6Plus_CVs
input4MIPs_CVs
obs4MIPs_CVs
(CORDEX_CVs)
(ESMO_CVs)
(GEWEX_CVs)
(SPARC_CVs)
(OtherProject_CVs)
…

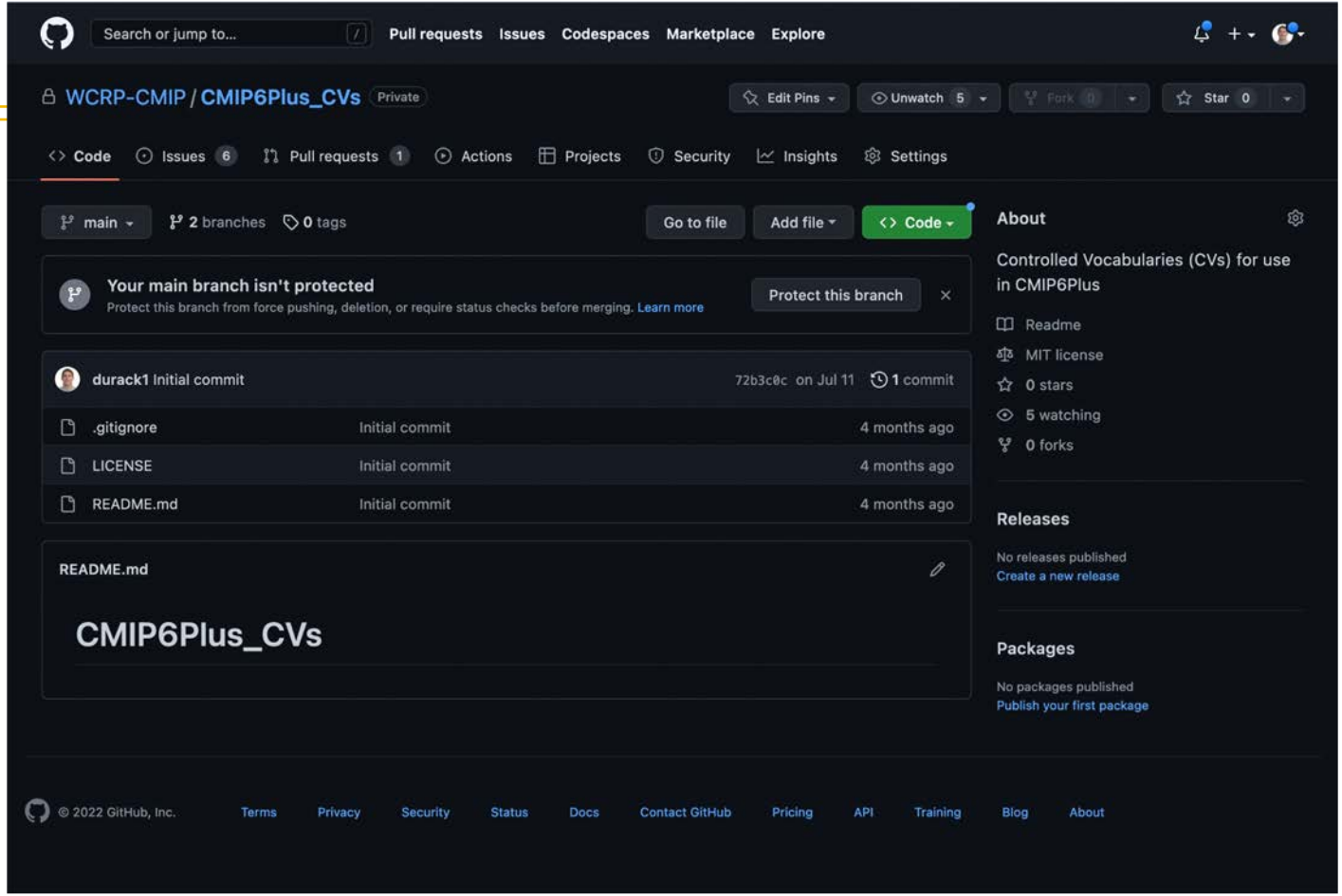
 # Required entries
DRS (directory structure and file naming conventions), <activity_id>, <experiment_id>, license, <mip_era>, required_global_attributes, source_id, tracking_id_prefix, <further_info_url>, <sub_experiment_id>,prime_esgf_node_id, permitted variables
(<..> denotes optional fields that may 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


Search or jump to...
Pull requests
Issues
Codespaces
Marketplace
Explore
PCMDI / mip-cmor-tables
Public
Edit Pins
Unwatch 5
Fork 0
Star 0
Code
Issues 5
Pull requests 1
Discussions
Actions
Projects
Wiki
Security
Insights
Settings
main
2 branches
0 tags
Go to file
Add file
Code
Your main branch isn't protected
Protect this branch from force pushing, deletion, or require status checks before merging. Learn more
Protect this branch
matthew-mizielinski Update README.md
0b5818c 20 days ago
8 commits
Tables
#3: update with CMIP6-> CMIP6Plus in example CV file
2 months ago
src
lifted funcs from juptyer notebook #7
22 days ago
.gitignore
Initial commit
9 months ago
LICENSE
Initial commit
9 months ago
README.md
Update README.md
20 days ago
README.md
mip-cmor-tables
JSON Tables for CMOR3 to create MIP datasets
Note that this repository is a work in progress
About
JSON Tables for CMOR3 to create MIP datasets (Work In Progress)
Readme
CC0-1.0 license
0 stars
5 watching
0 forks
Releases
No releases published
Create a new release
Packages
No packages published
Publish your first package
Contributors 2
matthew-mizielinski Matthew Miziel...
durack1 Paul J. Durack

CMIP


Search or jump to...
Pull requests
Issues
Codespaces
Marketplace
Explore
WCRP-CMIP / CMIP6Plus_CVs
Private
Edit Pins
Unwatch 5
Fork 0
Star 0
Code
Issues 6
Pull requests 1
Actions
Projects
Security
Insights
Settings
main
2 branches
0 tags
Go to file
Add file
Code
Your main branch isn't protected
Protect this branch from force pushing, deletion, or require status checks before merging. Learn more
Protect this branch
durack1 Initial commit
72b3c0c on Jul 11
1 commit
.gitignore
Initial commit
4 months ago
LICENSE
Initial commit
4 months ago
README.md
Initial commit
4 months ago
README.md
CMIP6Plus_CVs
About
Controlled Vocabularies (CVs) for use in CMIP6Plus
Readme
MIT license
0 stars
5 watching
0 forks
Releases
No releases published
Create a new release
Packages
No packages published
Publish your first package
© 2022 GitHub, Inc.
Terms
Privacy
Security
Status
Docs
Contact GitHub
Pricing
API
Training
Blog
About

CMIP

CMIP6 citation

Data Citation Service

Data Citation Service:

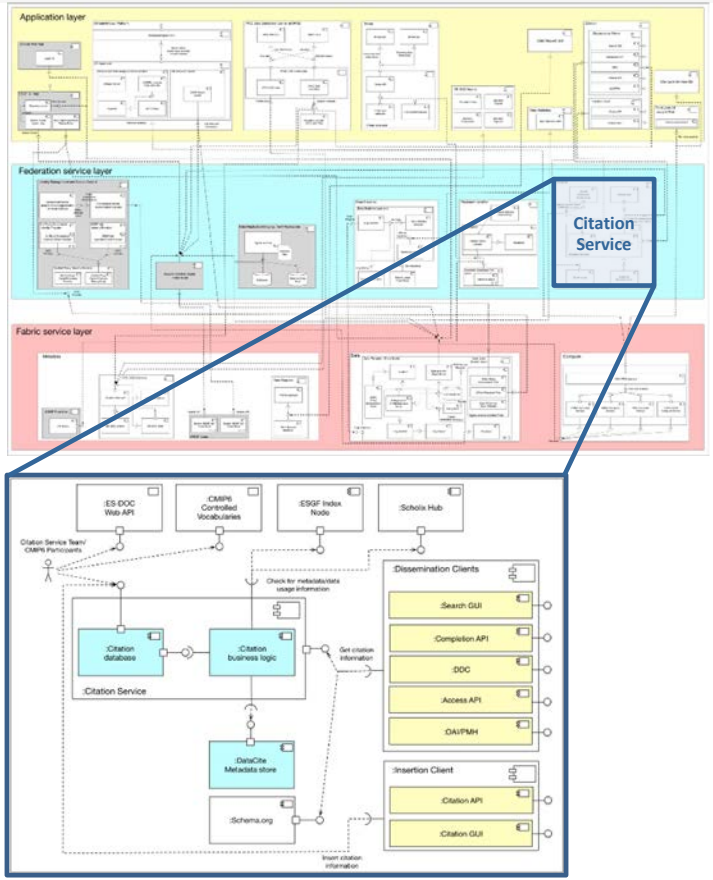
- Provide data DOIs on data collection granularities
- Provide information on data usage in papers
- Disseminate data citation information

Benefit:

- Data is citable in scientific publications.
- Receive credit for data creation
- Enhanced data discoverability

Data Citation Service as infrastructure component:

- Integrated in ESGF and ES-DOC
- Interfaces to DataCite and Scholix
- GUI [API] for maintaining citation information
- APIs for disseminating of citation information
- GUI for discovery of citation information
- Schema.org compliance enables integration in e.g. Google Dataset Search

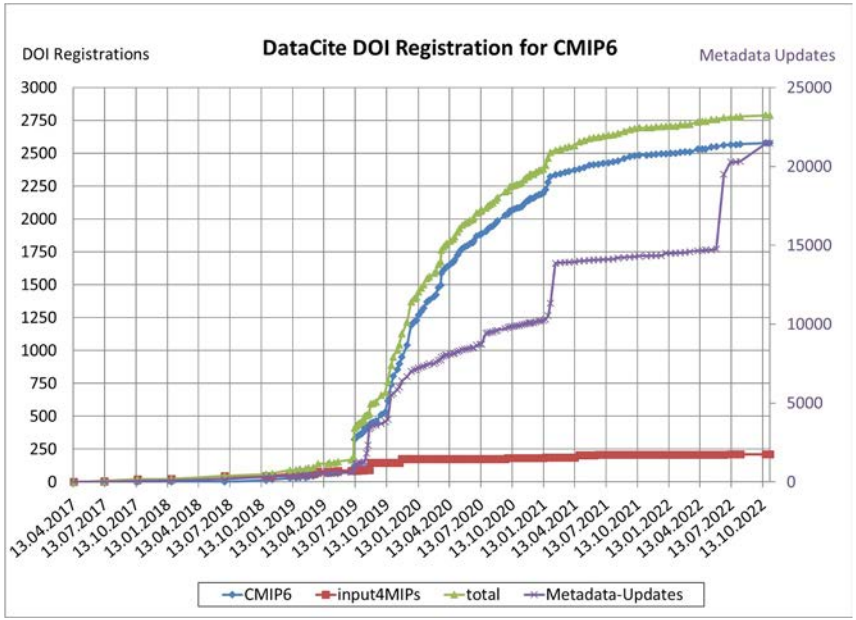


General Information: cmip6cite.wdc-climate.de
Available Data References: bit.ly/CMIP6_Citation_Search
Data Citation Statistics: bit.ly/CMIP6_DOI_Statistics

Data Citation Status

Data DOI Statistics:

- 2578 CMIP6 DOIs have been registered
- At WGI AR6 literature and data cut-off data, all datasets were citable (current coverage ~98%).

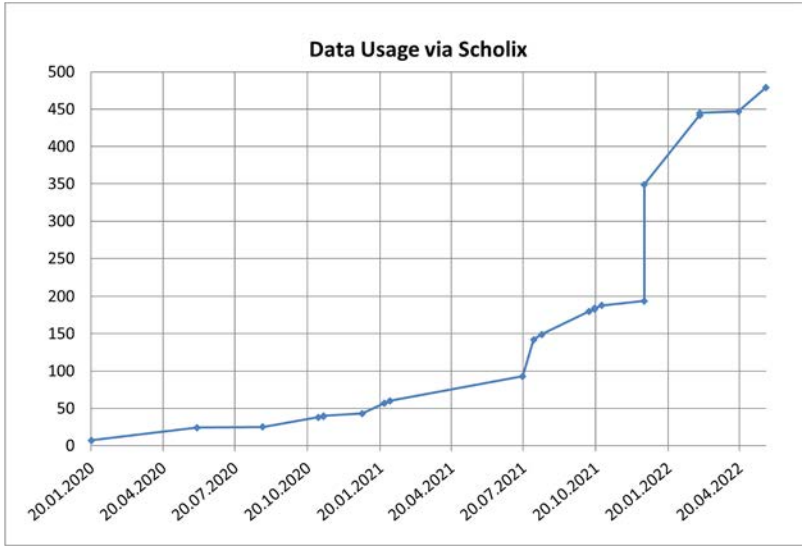


Data Citation Usage Status:

- IPCC WGI AR6 cited CMIP6 data in table AII.10
- 479 papers referencing CMIP6 data have been added

Incomplete data usage due to

- CMIP6 data citations not included in reference lists
- Publisher does not publish data references



Data Citation Future

Data Citation Task Team:

- Develop options for sustainable citation service
- Federate data citation
- Better integrate data citation into other infrastructure components and researchers' workflows
- Guidance for data citation Governance Board

Current Situation:

- Citation Service relies on a single person
- No funding

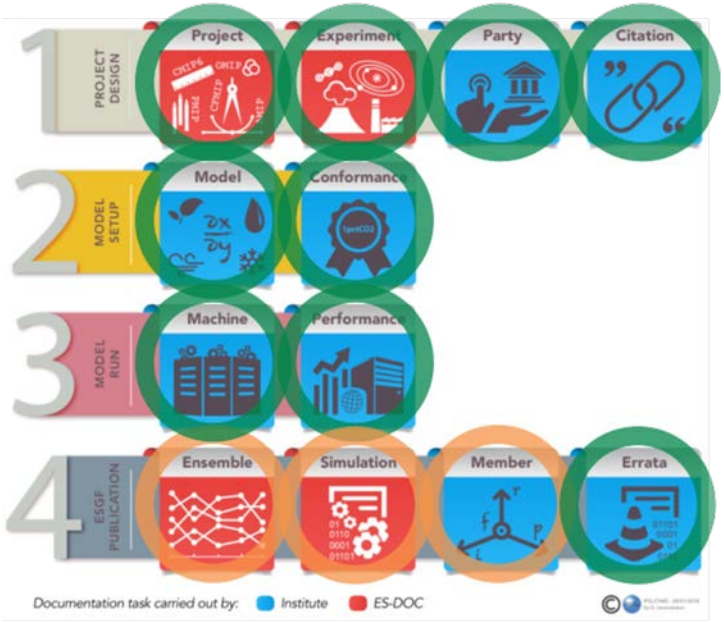
Future Situation:

- Federated approach with shared efforts and costs
- The Governance Board coordinates the federation and establishes a uniform and reliable citation service across WCRP projects.

CMIP6 - ES-Docs

ES-DOC (Earth System Documentation)

- The ES-DOC project supports the creation, dissemination and analysis of documentation describing the entire modelling workflow
- Status:
  - Infrastructure (from the ES-DOC team)
    - Delivered (2018 – 2022)
    - Delivery (2022: All **Simulation** and **Ensemble** descriptions will soon be linked automatically to the `further_info_URL`)
    - Content creation (by the CMIP6 groups): **Model**: 24 groups, **Machine & Performance**: 16 groups,**Conformance**: 4 groups
- Updates:
  - Since WGCM-24, spreadsheets for collecting **Machine**, **Performance** and **Conformance** documentation have been released.
  - New content in the last year: **Model**: 3 groups, **Machine & performance**: 5 groups, **Conformance**: 4 groups
- Plans:
  - The CMIP-IPO is coordinating a catch-up procedure with all CMIP6 modelling groups to fill in as much of the missing content as possible.
  - CMIP7 Task Team being set up (Use case analysis, infrastructure assessment, *etc.*)

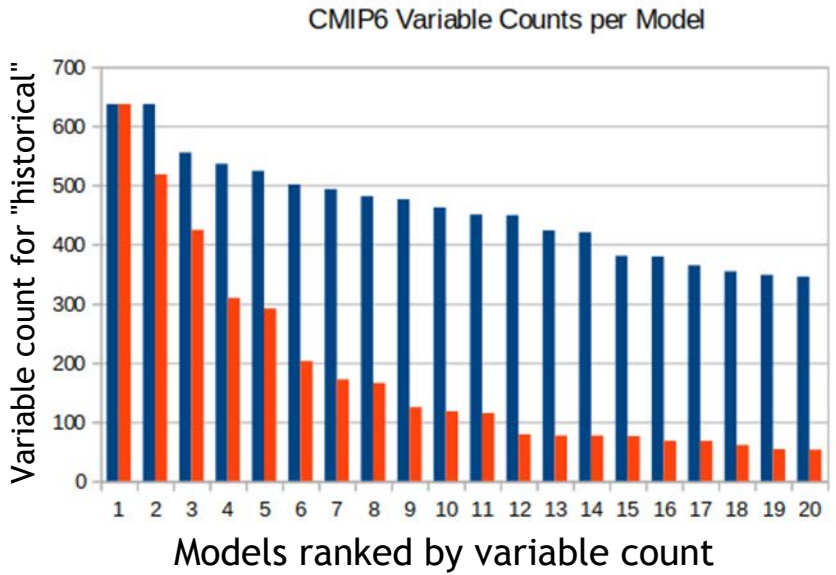


Documentation types and the status of their collection infrastructure

CMIP6 - Data Request

Variable output by Model for CMIP6 Historical

- Around 100 priority 1 variables had not been published by anybody (March, 2022).
- There are 673 variables archived from the IPSL-CM6A-LR model historical simulation.
- 20th ranked AWI-ESM-1-1-LR has 346 variables archived.
- The intersection of the variables output by the first **20 models is only 53.**
- The level of consistency is reduced if multiple experiments are considered.



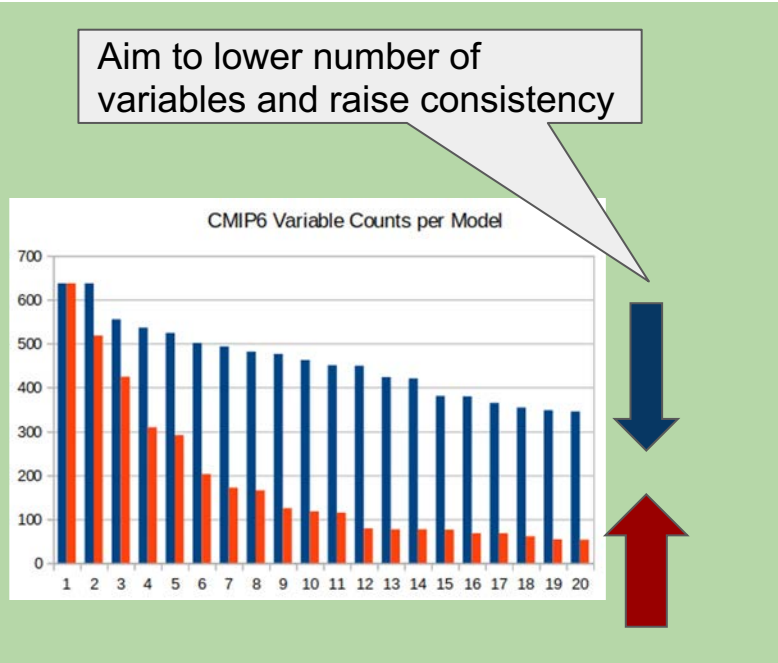
Blue: number of variables saved from model
Red: intersection of variables output for models 1 to n

Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

Ambition and Objectives

- **Ambition**: Address the "grade inflation" that led to the majority of 2000+ variables defined in CMIP6 to be rated as top priority

- **Objectives:**
  - Identify a set of core variables and associated metadata to enhance accessibility of climate projections to direct and indirect users.
  - Define a process for update and maintenance of the list.
  - Author team includes representation from climate science, infrastructure and modeling communities to ensure balance between expectations and awareness of constraints.



Paper Structure

1. **Introduction**
2. **Process and Methodology**
3. **The form and role of the core variables**
4. **Outlook and Scientific Context (or: context and horizon scan)**
5. **Results: The Core Variables**
6. **Validation tests and tools**
7. **Updates and Extensions to the list [pending initial process development and CMIP7 outline]**

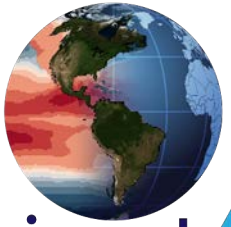
Appendix 1: Stakeholder Consultation Process

Appendix 2: The Core Variables

Appendix 3: Spatial and Temporal Structures

CMIP - next-gen forcing

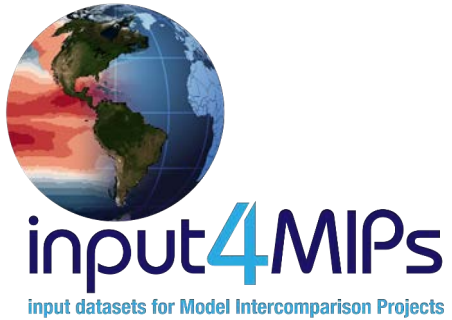
input4MIPs



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 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



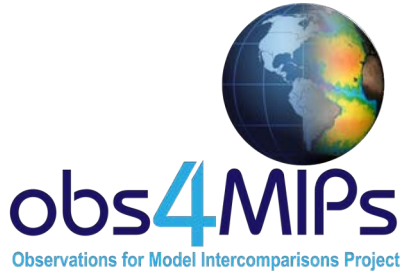
Feedback from modelling groups

- 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

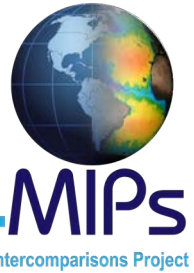
- Missing forcings? (IPSL: N-cycle, water isotopes,)
- Forcing data problems? (Led to 3 CMIP6 releases: 6.0 Dec 2016, 6.1 May '17, 6.2.1 Oct '17)

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)

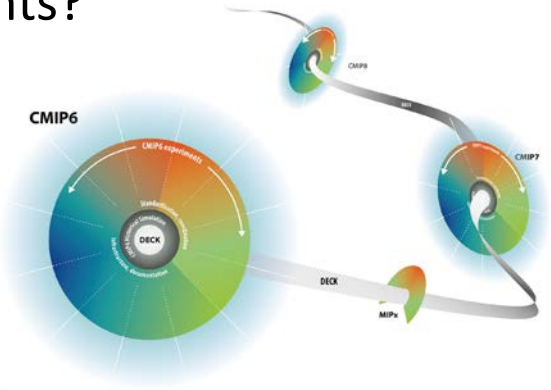
obs4MIPs in 2022

- 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.**

CMIP6+ - facilitate science

CMIP6+

How can we leverage existing infrastructure investments?



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

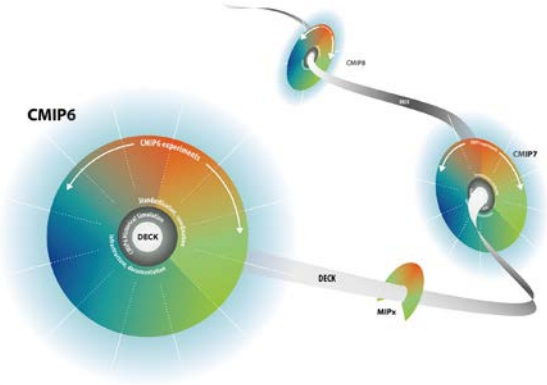
- Allow CMIP to evolve and “operationalize”
  - Incremental change (e.g. maintain ESGF dependence)
  - Next gen forcings and obs
  - Change little, increment, allowing modeling groups to focus on science

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

CMIP6+

How can we leverage existing communities?

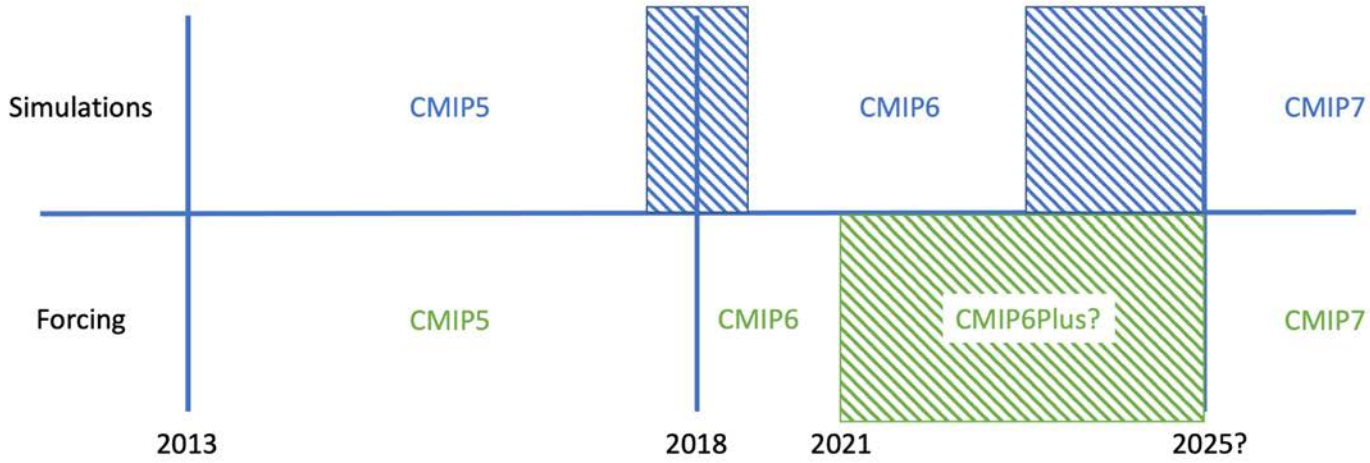
- Continue beyond CMIP6, few changes as CMIP7 is discussed and planned
    - Reduce time pressures - loosen CMIP*x* - IPCC AR*x* linkage
    - Continuous “CMIP science” - not monolithic phases every ~7 years



- 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)

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

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