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-829833
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