Lessons learned from 25 years of PMIP model-data distribution

Jean-Yves Peterschmitt, Pascale Braconnot, Masa Kageyama

doi.org/10.22498/pages.26.2.60

Lessons learned from 25 years of PMIP model-data distribution

Jean-Yves Peterschmitt, P. Braconnot and M. Kageyama

Open paleo data from both observations and models underlies the success of the Paleoclimate Modelling Intercomparison Project. We present how the project has evolved from a stand-alone database to an active member of a distributed international infrastructure following community standards.

Climate models are improved iteratively, as scientific knowledge, along with computing and storage technology progress. Sharing and comparing models and their output to paleo reconstructions is an essential part of this process. This can be done by sharing data directly between individuals, but is more efficient when formally organized as a MIP (Model Intercomparison Project), where all contributors and users adopt the same standards. The Paleoclimate Modelling Intercomparison Project (PMIP), started in 1990 (Joussaume and Taylor 1995), was one of the early MIPs, following the AMIP example (Gates et al. 1998).

PMIP has been successful in terms of participation, publications, and contributions to successive IPCC Working Group 1 reports, and is now in its fourth phase, with 20 modeling groups/models from 14 countries (Kageyama et al. 2018; Kageyama et al. 2016 [PMIP4 special issue]). The first studied periods were the mid-Holocene and the Last Glacial Maximum, with the pre-industrial period used as a control run. PMIP4 now includes five additional experiments: the last millennium, the Last Interglacial, the mid-Pliocene Warm Period, the last deglaciation and DeepMIP. Thanks to improvements in model complexity, resolution, and length of the simulations, the different phases of PMIP have targeted key scientific questions on climate sensitivity, the hydrological cycle, and abrupt event and inter-annual to multi-decadal variability.

For PMIP4, experimental protocols were co-designed by the modeling and data communities (Kageyama et al. 2018). They require that the same model version be used for PMIP4-CMIP6 experiments and future climate projections so that rigorous analyses of climate processes, including both physical and biogeochemical interactions, can be performed across the range of past and future climate. This is done in collaboration with other CMIP6 MIPs (Eyring et al. 2016).

PMIP simulations address the key CMIP6 overarching questions:

- How does the Earth system respond to forcing?
- What are the origins and consequences of systematic model biases?
- How can we assess future climate changes given climate variability, predictability and uncertainties in scenarios?

	PMIP 1	PMIP 2	PMIP 3	PMIP 4
DB online	1996	2005	2011	2018
Number of groups/ models	22	18	25	20
Number of countries	11	10	12	14
Main experiments	0 k 6 k 21 k	Same as PMIP 1	PMIP 2 + Last Millennium	PMIP3 + Last Interglacial + Mid Pliocene Warm Period + Last Deglaciation + DeepMIP
DB Size	1.7 GB	482 GB	**distributed** *several* TB	**distributed** LOTS of TB…
Data distribution	ftp server LSCE (+PCMDI)	DODS server LSCE	CMIP5 ESGF	CMIP6 ESGF
Data format & Convention	NetCDF AMIP/CF	NetCDF CMIP+PMIP2/CF	NetDCF CMIP5/CF	NetCDF CMIP6/CF
Example grid IPSL atmosphere	lmcelmd5 64x50 x L11	IPSL-CM4-V1-MR 96x72 x L19	IPSL-CM5A-LR 96x95 x L39	IPSL-CM6A-LR 144x143 x L79
Example grid NCAR atmosphere	ccsm3 128x64 x L18	CCSM 128x64 x L17	CCSM4 288x192 x L26	CESM2 288x192 x L32

Box 1: PMIP database factsheet

Current work places a particular emphasis on the assessment of the different sources of uncertainties resulting from, for example, model formulation, reconstructions of forcing, and internal model noise. Model-data comparisons are key in this process (Braconnot et al. 2012; Harrison et al. 2015).

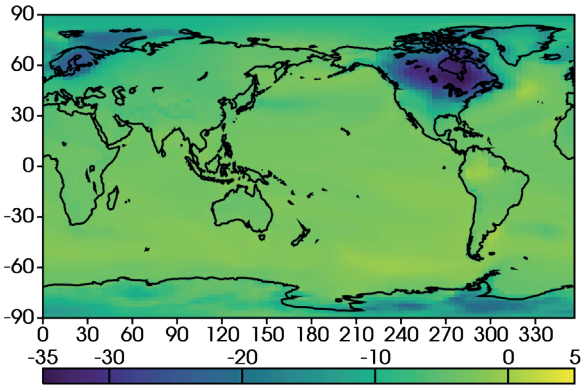
The PMIP model database has progressed from almost 2 GB for PMIP1 (~14,500 files) to a frightening (and unknown!) number of terabytes for PMIP4 (Box 1). Standards and good data-distribution tools are the key to dealing with the massive amount of data generated, along with good communication tools (mailing lists and websites), and invaluable help from the Earth System Grid Federation (ESGF; Balaji et al. 2018) community that maintains the CMIP database.

Using standards

The database of model output is too large to be accessed by ordinary database queries. Nevertheless, users need to easily access the subset of the data they need for their analyses, regardless of which research group generated it. In PMIP, this is achieved through the use of community standards. Standards are sometimes viewed as a hindrance to data production, but they are necessary to avoid chaos when working with multi-model data – the essence of a MIP. Data that is consistent across all the models and experiments eases reuse by users, and is required to automatically process numerous files, easily ingest new files, and to reprocess files when a bug is found. Such standardization also generally makes any analyses more reproducible.

Standardization is a key aspect of the long history of PMIP in international collaborations. PMIP currently follows the CMIP6 standards for file format (NetCDF format) and metadata (Climate and Forecast conventions, CMIP6 Data Reference Syntax, Controlled Vocabulary and Data Request). The NetCDF binary format has many advantages: self-describing, easily and efficiently writable/readable by programs, capacity to hold several gigabytes of data, and suitable for long-term archiving. Thanks to these choices it is still possible to access the content of PMIP1 files created more than 20 years ago. It is not easy for the modeling groups to meet the CMIP6 requirements, but the Climate Model Output Rewriter (CMOR3) library and project-specific configuration


```
In [1]:  import cdms2, cdutil, vcs
In [2]:  f = cdms2.open('/home/scratch01/jypeter/tas_IPSL-CM5A-LR_21-0.nc')
In [3]:  v = f('tas_diff')
In [4]:  v_annual_mean = cdutil.YEAR(v)
In [12]: x = vcs.init(geometry=(600,400), bg=True)
In [16]: x.plot(v_annual_mean)
Out [16]:
```




```
In [17]: x.clear()
         x.plot(v_annual_mean(latitude=(20,80), longitude=(-20,40)), ratio='autot')
Out [17]:
```

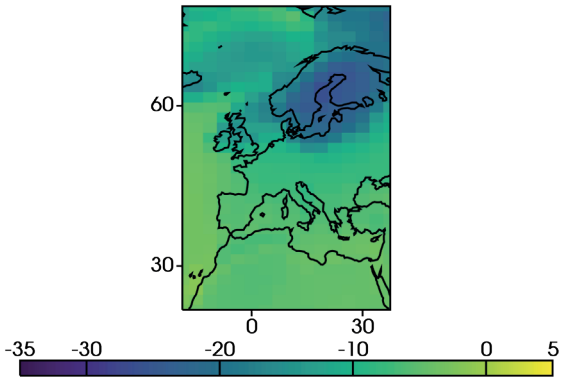


Figure 1: An example of a script using the CDAT climate-oriented Python distribution in a notebook to read PMIP3 data from a NetCDF file (surface temperature, anomaly for the Last Glacial Maximum minus the pre-industrial control, for the IPSL-CM5A-LR model), compute the annual mean and plot it on a global grid and a smaller region.

tables facilitate the creation of CMIP-compliant files.

Accessing model data

Once the data files are available in the standard format, the next goal is to ensure they move as smoothly as possible from the data provider to the data user. This is accomplished through a number of developments:

- Model-output data providers need an automatic service to answer user requests.
- Users want to determine easily if the required data is available, and then to easily access the files. Given the size of the database (Box 1) there are ongoing developments to provide computation and analysis services directly on the servers holding the data.
- Users need a good documentation of the models and how the PMIP experiments were run. For PMIP4-CMIP6, this information will be centralized on the Earth System Documentation (es-doc) site.

For CMIP5-6 (PMIP3-4), the data files are sent by the modeling groups to the closest ESGF Data Node and, after review ranging from a basic validation to an exhaustive quality control, they can be searched and downloaded from any other node of the federation. This distributed repository is scalable and is the only practical way to handle the 10-50 petabytes of data expected for CMIP6 (including PMIP4 data). ESGF also offers a fast web-search interface and bulk data-download tools. This infrastructure is powerful, but it requires substantial manpower for customized software development and local node administration, as well as sufficient storage and computing resources.

In addition to standardization, the PMIP data policy has evolved over time. For PMIP1, the full database was initially available only for the groups which had submitted data during an embargo period, prior to public release. For PMIP2, the database was also available for people proposing an analysis project. PMIP3-4 followed the CMIP5-6 data policy, which allows anyone to use the data from modeling groups, with some restrictions for commercial applications. In turn, the results of the study that uses the model output must be shared with the same open policy, without forgetting to credit the producers.

Using PMIP data

There are many ways to use PMIP model data, depending on the analyses to be done. The data complexity (number of available variables and file size) has increased substantially since the beginning of PMIP, but the programing complexity has decreased. It is now much easier to use a high-level scripting language (Fig. 1) than it was to use Fortran programs. Users can also process PMIP data with the Graphical User Interfaces provided by some programs (e.g. GIS programs such as QGIS), but they may be quickly limited by data size and available operations. There is also an ongoing effort by the PMIP community to provide some higher-level web interface; this will receive more attention in the coming years.

Conclusion

PMIP has benefited from CMIP5-6 and the ESGF infrastructure, which has eased the comparison between past and future climate simulations. One of the next challenges is to make using the data easier for non-modelers, especially experts in paleoclimate reconstructions. This will require the deployment of specific web servers similar to the ones used for impact studies, but customized for paleoclimate needs. Another challenge will be to deal with the long, transient climate simulations (thousands of years of model data) generated by the PMIP4 experiments (deglaciation, the Eemian and the Holocene) when performing model-model and model-data comparisons.

RESOURCES

PMIP: pmip.lsce.ipsl.fr
AMIP and CMIPn: pcmdi.llnl.gov/mips
PMIP3 publications: citeulike.org/user/jypeter/order/year
NetCDF: unidata.ucar.edu/software/netcdf
CF conventions: cfconventions.org
CMIP6 DRS: goo.gl/v1drZl
CMIP6 DR: earthsystemcog.org/projects/wip/CMIP6DataRequest
CMIP6 CV: github.com/WCRP-CMIP/CMIP6_CVs
CMOR3 library: cmor.llnl.gov
es-doc: search.es-doc.org
CDAT: cdat.llnl.gov
QGIS: qgis.org
DeepMIP: deepmip.org

AFFILIATIONS

Laboratoire des Sciences du Climat et de l'Environnement (LSCE/IPSL, CEA-CNRS-UVSQ), Université Paris-Saclay, Gif-sur-Yvette, France

CONTACT

Jean-Yves Peterschmitt: Jean-Yves.Peterschmitt@lsce.ipsl.fr

REFERENCES

Balaji V et al. (2018) Geosci Mod Dev 11: 3659-3680
Braconnot P et al. (2012) Nat Clim Change 2: 417-424
Eyring V et al. (2016) Geosci Mod Dev 9: 1937-1958
Gates WL et al. (1998) BAMS 73: 1962-1970
Harrisson SP et al. (2015) Nat Clim Change 5: 735-743
Joussaume J, Taylor K (1995) Proceedings of the first international AMIP scientific conference, Monterey, USA: 425-430
Kageyama M et al. (2016) CP/GMD inter-journal SI
Kageyama M et al. (2018) Geosci Mod Dev 11: 1033-1057