Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2018-52
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 22 March 2018

Geoscientific Model Development Discussions

Requirements for a global data infrastructure in support of CMIP6

Venkatramani Balaji[1,2], Karl E. Taylor[3], Martin Juckes[4], Michael Lautenschlager[5], Chris Blanton[6,2], Luca Cinquini[7], Sébastien Denvil[8], Paul J. Durack[3], Mark Elkington[9], Francesca Guglielmo[8], Eric Guilyardi[8,10], David Hassell[10], Slava Kharin[11], Stefan Kindermann[5], Bryan N. Lawrence[10,4], Sergey Nikonov[1,2], Aparna Radhakrishnan[6,2], Martina Stockhause[5], Tobias Weigel[5], and Dean Williams[3]

[1]Princeton University, Cooperative Institute of Climate Science, Princeton NJ, USA
[2]NOAA/Geophysical Fluid Dynamics Laboratory, Princeton NJ, USA
[3]PCMDI, Lawrence Livermore National Laboratory, Livermore, CA, USA
[4]Science and Technology Facilities Council, Abingdon, UK
[5]Deutsches KlimaRechenZentrum GmbH, Hamburg, Germany
[6]Engility Inc., NJ, USA
[7]Jet Propulsion Laboratory (JPL), 4800 Oak Grove Drive, Pasadena, CA 91109, USA
[8]Institut Pierre-Simon Laplace, CNRS/UPMC, Paris, France
[9]Met Office, FitzRoy Road, Exeter, EX1 3PB, UK
[10]National Center for Atmospheric Science and University of Reading, UK
[11]Canadian Centre for Climate Modelling and Analysis, Atmospheric Environment Service, University of Victoria, BC, Canada

Correspondence to: V. Balaji (balaji@princeton.edu)

Abstract. The World Climate Research Programme (WCRP)'s Working Group on Climate Modeling (WGCM) Infrastructure Panel (WIP) was formed in 2014 in response to the explosive growth in size and complexity of Coupled Model Intercomparison Projects (CMIPs) between CMIP3 (2005-06) and CMIP5 (2011-12). This article presents the WIP recommendations for the global data infrastructure needed to support CMIP design, future growth and evolution. Developed in close coordination with those who build and run the existing infrastructure (the Earth System Grid Federation), the recommendations are based on several principles beginning with the need to separate requirements, implementation, and operations. Other important principles include the consideration of data as a commodity in an ecosystem of users, the importance of provenance, the need for automation, and the obligation to measure costs and benefits.

This paper concentrates on requirements, recognising the diversity of communities involved (modelers, analysts, software developers, and downstream users). Such requirements include the need for scientific reproducibility and accountability alongside the need to record and track data usage for the purpose of assigning credit. One key element is to generate a dataset-centric rather than system-centric focus, with an aim to making the infrastructure less prone to systemic failure.

With these overarching principles and requirements, the WIP has produced a set of position papers, which are summarized here. They provide specifications for managing and delivering model output, including strategies for replication and versioning, licensing, data quality assurance, citation, long-term archival, and dataset tracking. They also describe a new and more formal approach for specifying what data, and associated metadata, should be saved, which enables future data volumes to be estimated.

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2018-52
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 22 March 2018

Geoscientific Model Development Discussions

The paper concludes with a future-facing consideration of the global data infrastructure evolution that follows from the blurring of boundaries between climate and weather, and the changing nature of published scientific results in the digital age.

1 Introduction

CMIP6 (Eyring et al., 2016a), the latest Coupled Model Intercomparison Project (CMIP), can trace its genealogy back to the Charney Report (Charney et al., 1979). This seminal report on the links between CO_2 and climate was an authoritative summary of the state of the science at the time, and produced findings that have stood the test of time (Bony et al., 2013). It is often noted that the range and uncertainty bounds on equilibrium climate sensitivity generated in this report have not fundamentally changed, despite the enormous increase in resources devoted to analysing the problem in decades since.

Beyond its prescient findings on climate sensitivity, the Charney Report also gave rise to a methodology for the treatment of uncertainties and gaps in understanding, which has been equally influential, and is in fact the basis of CMIP itself. The Report can be seen as one of the first uses of the *multi-model ensemble*. At the time, there were two models capable of representing the equilibrium response of the climate system to a change in CO_2 forcing, one from Syukuro Manabe's group at NOAA's Geophysical Fluid Dynamics Laboratory, and the other from James Hansen's group at NASA's Goddard Institute for Space Studies. Then as now, these groups marshaled vast state-of-the-art computing and data resources to run very challenging simulations of the Earth system. The Report's results were based on an ensemble of 3 runs from Manabe, labeled M1-M3, and two from Hansen, labeled H1-H2.

By the time of the IPCC First Assessment Report (FAR) in 1990, the process had been formalized. At this stage, there were 5 models participating in the exercise, and some of what has now been formalized as the "Diagnosis, Evaluation, and Characterization of Klima" (DECK) experiments[1] had been standardized (a pre-industrial control, 1% per year CO_2 increase to doubling, etc). The "scenarios" had emerged as well, for a total of 5 different experimental protocols. Fast-forwarding to today, CMIP6 expects more than 75 models from around 35 modeling centers (in 14 countries, a stark contrast to the US monopoly in Charney et al., 1979) to participate in the DECK and historical experiments (Table 2 of Eyring et al., 2016a), and some subset of these to participate in one or more the 21 MIPs endorsed by the CMIP Panel (Table 3 of Eyring et al., 2016a). The MIPs call for over 200 experiments, a considerable expansion over CMIP5.

Alongside the experiments themselves is the data request which defines, for each CMIP experiment, what output each model should provide for analysis. The complexity of this data request has also grown tremendously over the CMIP era. A typical dataset from the FAR archive (from the GFDL R15 model) lists climatologies and time series of two variables, and the dataset size is about 200 MB. The CMIP6 Data Request Juckes et al. (2015) lists literally thousands of variables from the hundreds of experiments mentioned above. This growth in complexity is testament to the modern understanding of many physical, chemical and biological processes which were simply absent from the Charney Report era models.

The simulation output is now a primary scientific resource for researchers the world over, rivaling the volume of observed weather and climate data from the global array of sensors and satellites (Overpeck et al., 2011). Climate science, and observed

[1] "Klima" is German for "climate".

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2018-52
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 22 March 2018

Geoscientific Model Development Discussions

and simulated climate data in particular, have now become primary elements in the "vast machine" (Edwards, 2010) serving the global climate and weather enterprise.

Managing and sharing this huge amount of data is an enterprise in its own right – and the solution established for CMIP5 was the global "Earth System Grid Federation" (ESGF, (Williams et al., 2015)). ESGF was identified by the WCRP Joint Scientific Committee in 2013 as the recommended infrastructure for data archiving and dissemination for the Programme. The larger gateways currently participating in the ESGF are shown in in Figure 1, which also lists (some of) the many projects these nodes support. With multiple agencies and institutions, and many uncoordinated and possibly conflicting requirements, the ESGF itself is a complex and delicate component to manage.

The sheer size and complexity of this infrastructure emerged as a matter of great concern at the end of CMIP5, when the growth in data volume relative to CMIP3 (from 40 TB to 2 PB, a 50-fold increase in 6 years) suggested the community was on an unsustainable path. These concerns led to the 2014 recommendation of the WGCM to form an *infrastructure panel* (based upon , a proposal at the 2013 annual meeting). The WGCM Infrastructure Panel (WIP) was tasked with examining the global computational and data infrastructure underpinning CMIP, and improving communication between the teams overseeing the scientific and experimental design of these globally coordinated experiments, and the teams providing resources and designing that infrastructure. The communication was intended to be two-way: providing input both to the provisioning of infrastructure appropriate to the experimental design, and informing the scientific design of the technical (and financial) limits of that infrastructure.

This paper is a summary of the requirements identified by the WIP in the first three years of activity since its formation in 2014, alongside the recommendations which have arisen. In Section 2, the principles and scientific rationale underlying the requirements for global data infrastructure are articulated. In Section 3 the CMIP6 Data Request is covered: standards and conventions, requirements for modeling centers to process a complex data request, and projections of data volume. In Section 4, recent evolution in how data are archived is reviewed alongside a licensing strategy consistent with current practice and scientific principle. In Section 5 issues surrounding data as a citable resource are discussed, including the technical infrastructure for the creation of citable data, and the documentation and other standards required to make data a first-class scientific entity. In Section 6 the implications of data replicas and in Section 7 issues surrounding data versioning, retraction, and errata are addressed. Section 8 provides an outlook for the future of global data infrastructure, looking beyond CMIP6 towards a unified view of the "vast machine" for weather and climate computation and data.

2 Principles underlying the infrastructure requirements

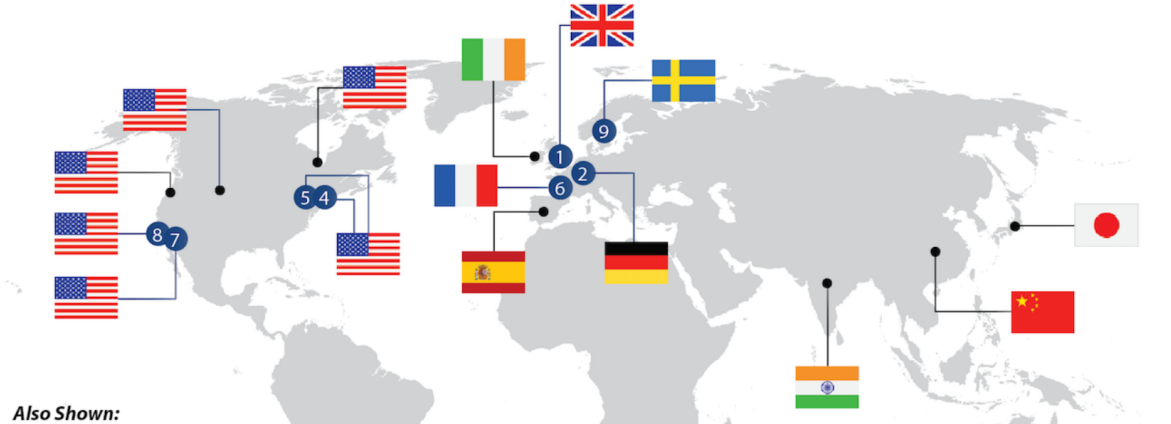
In the pioneering days of CMIP, the community of participants was small and well-knit, and all the issues involved in generating datasets for common analysis from different modeling groups could be settled by mutual agreement (Ron Stouffer, personal communication). Analysis was performed by the same community that performed the simulations. The Program for Climate Model Diagnostics and Intercomparison (PCMDI), established in 1989, had championed the idea of more systematic analysis of models, and in close cooperation with the climate modeling centers, PCMDI assumed responsibility for much of the day-to-

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2018-52
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 22 March 2018

Geoscientific Model Development Discussions

Major ESGF Node Sites

Institution	Gateway URL	Version	Country	Project(s)	Contact
1 CEDA	esgf-index1.ceda.ac.uk	2.4.0	U.K.	CMIP5, CORDEX, Obs4MIPs, SPECS, ESA CCI, EUCLEIA, CLIPC	alan.iwi@stfc.ac.uk
2 DKRZ	esgf-data.dkrz.de	2.4.0	Germany	CMIP5, CORDEX, Obs4MIPs, ISI-MIP	berger@dkrz.de
3 ANU NCI	esgf.nci.org.au	2.4.0	Australia	CMIP5	ben.evans@anu.edu.au
4 NOAA GFDL	esgdata.gfdl.noaa.gov	2.4.0	U.S.	CMIP5, ncpp2013, Obs4MIPs	hans.vahlenkamp@noaa.gov
5 NASA GSFC	esgf.nccs.nasa.gov	2.4.0	U.S.	CMIP5, Obs4MIPs, Ana4MIPs, NEX-GDDP, NEX-DCP30, CREATE-IP	daniel.q.duffy@nasa.gov
6 IPSL	esgf-node.ipsl.upmc.fr	2.4.0	France	CMIP5, CORDEX, Obs4MIPs	sebastien.denvil@ipsl.jussieu.fr
7 NASA JPL	esgf-node.jpl.nasa.gov	2.4.0	U.S.	Obs4MIPs, GASS-YoTC, CMAC	luca.cinquini@jpl.nasa.gov
8 DOE LLNL	esgf-node.llnl.gov	2.4.0	U.S.	CMIP5, CMIP3, input4MIPS, ACME	sasha@llnl.gov
9 LiU	esg-dn1.nsc.liu.se	2.4.0	Sweden	CMIP5, CORDEX, Obs4MIPs	pchengi@nsc.liu.se



Also Shown:

Country	Institute	Gateway URL
United States	DOE Argonne National Laboratory	dev.esgf.anl.gov/projects/esgf-anl
United States	NOAA Earth System Research Laboratory	esgf.esrl.noaa.gov
China	Beijing Normal University	naqpms.iap.ac.cn
India	Indian Institute of Tropical Meteorology	cordex.org
Spain	University of Cantabria	cordex.org

Figure 1. Sites participating in the Earth System Grid Federation in 2017. Figure courtesy Dean Williams, adapted from the ESGF Brochure.

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2018-52
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 22 March 2018

day coordination of CMIP. Until CMIP3, the hosting of datasets from different modeling groups could be managed at a single archival site; PCMDI alone hosted the entire 40 TB archive.

From its earliest phases, CMIP grew in importance, and its results provided a major pillar supporting the periodic Intergovernmental Panel on Climate Change (IPCC) assessment activity. However, the explosive growth in the scope of CMIP, especially between CMIP3 and CMIP5, represented a tipping point in the supporting infrastructure. It became evident that fundamental changes would be needed to address the evolving scientific and operational requirements, which are summarized here:

1. With greater complexity and a globally distributed data resource, it has become clear that in the design of globally coordinated scientific experiments, the global computational and data infrastructure needs to be formally examined as an integrated element.

   – The WIP was formed in response to this observation, with membership drawn from experts in various aspects of the infrastructure. Representatives of modeling centers, infrastructure developers, and stakeholders in the scientific design of CMIP and its output comprise the panel membership.

   – One of the WIP's first acts was to consider three phases in the process of infrastructure development: *requirements*, *implementation*, and *operations*, all informed by the builders of workflows at the modeling centers.

     – The WIP, in consort with the CMIP Panel, takes responsibility to articulate requirements for the infrastructure.

     – The implementation is in the hands of the infrastructure developers, principally ESGF for the federated archive (Williams et al., 2015), but also related projects like Earth System Documentation (ES-DOC, Guilyardi et al., 2013).

     – In 2016 at the WIP's request, the CMIP6 Data Node Operations Team (CDNOT) was formed. It is charged with ensuring that all the infrastructure elements needed by CMIP6 are properly deployed and actually working as intended at the sites hosting CMIP6 data. It is also responsible for the operational aspects of the federation itself, including specifying what versions of the toolchain are run at every site at any given time, and organizing coordinated version upgrades across the federation.

   Although there is now a clear separation of concerns into requirements, implementation, and operations, close links are maintained by cross-membership between the key bodies, including the WIP itself, the CMIP Panel, the ESGF Executive Committee, and the CDNOT.

2. With the basic fact of anthropogenic climate change now well established (see, e.g., Stocker et al., 2013) the scientific communities with an interest in CMIP is expanding. For example, a substantial body of work has begun to emerge to examine climate impacts.

   – In addition to the specialists in Earth system science – who also design and run the experiments and produce the model output – those relying on CMIP output now include those developing and providing climate services, as well

as *consumers* from allied fields studying the impacts of climate change on health, agriculture, natural resources, human migration, and similar issues (Moss et al., 2010). This confronts us with a *scientific scalability* issue (the data during its lifetime will be consumed by a community much larger, both in sheer numbers, and also in breadth of interest and perspective than the Earth system modeling community itself), which needs to be addressed.

- Accordingly, the WIP has promulgated the requirement that infrastructure should ensure maximum transparency and usability for user (consumer) communities at some distance from the modeling (producer) communities.

3. While CMIP and the IPCC are formally independent, the CMIP archive is increasingly a reference in formulating climate policy. Hence the *scientific reproducibility* (Collins and Tabak, 2014) and the underlying *durability* and *provenance* of data have now become matters of central importance: being able to trace, long after the fact, back from model output to the configuration of models and analysis procedures and choices made along the way.

   - This led the IPCC to require data distribution centers (DDCs) to attempt to guarantee the archival and dissemination of this data in perpetuity, and
   - the WIP to promote the importance in the CMIP context of achieving reproducibility. Given the use of multi-model ensembles for both consensus estimates and uncertainty bounds on climate projections, it is important to document – as precisely as possible, given the independent genealogy and structure of many models – the details and differences among model configurations and analysis methods, to deliver both the requisite provenance and the routes to reproduction.

4. With the expectation that CMIP DECK experiment results should be routinely contributed to CMIP, opportunities now exist for engaging in a more systematic and routine evaluation of Earth System Models (ESMs). This has led to community efforts to develop standard metrics of model "quality" (Eyring et al., 2016b; Gleckler et al., 2016).

   - Typical multi-model analysis has hitherto taken the multi-model average, assigning equal weight to each model, as the most likely estimate of climate response. This "model democracy" (Knutti, 2010) has been called into question and there is now a considerable literature exploring the potential of weighting models by quality (Knutti et al., 2017). The development of standard metrics would aid this kind of research.
   - To that end, there is now a requirement to enable through the ESGF a framework for accommodating quasi-operational evaluation tools that could routinely execute a series of standardized evaluation tasks. This would provide data consumers with an increasingly (over time) systematic characterization of models. The WIP recognizes it may be some time before a fully operational system of this kind can be implemented, but planning must start now.

5. As the experimental design of CMIP has grown in complexity, costs both in time and money have become a matter of great concern, particularly for those designing, carrying out, and storing simulations. In order to justify commitment of resources to CMIP, mechanisms to identify costs and benefits in developing new models, performing CMIP simulations, and disseminating the model output need to be developed.

- To quantify the scientific impact of CMIP, measures are needed to *track* the use of model output and its value to consumers.
- In addition to usage quantification, credit and tracing data usage in literature via citation of data is important. Current practice is at best citing large data collections provided by a CMIP participant, or all of CMIP. Accordingly, the WIP has defined and is encouraging use of a mechanism to identify and *cite* data provided by each modeling center.
- Alongside the intellectual contribution to model development, which can be recognized by citation, there is a material cost to centers in computing which is both burdensome and poorly understood by those requesting, designing and using CMIP experiments. To begin documentation of these costs for CMIP6, the "Computational Performance" MIP project (CPMIP) (Balaji et al., 2017) has been established.

6. Experimental specifications have become ever more complex, making it difficult to verify that experiment configurations conform to those specifications.
   - Several modeling centers have encountered this problem in preparing for CMIP6, noting, for example, the challenging intricacies in dealing with input forcing data (see Durack et al., 2017), output variable lists (Juckes et al., 2015), and crossover requirements between the endorsed MIPs and the DECK (Eyring et al., 2016a) . Moreover, these protocols inevitably evolve over time, as errors are discovered or enhancements proposed, and centers needed to be adaptable in their workflows accordingly.
   - The WIP therefore recognized a requirement to encode the protocols to be directly ingested by workflows, in other words, *machine-readable experiment design*. The requirement spans all of the *controlled vocabularies* (CVs: for instance the names assigned to models, experiments, and output variables) used in the CMIP protocols as well as the CMIP6 Data Request (Juckes et al., 2015), which must be stored in version-controlled, machine-readable formats. Precisely documenting the *conformance* of experiments to the protocols (Lawrence et al., 2012) is an additional requirement.

7. The transition from a unitary archive at PCMDI in CMIP3 to a globally federated archive in CMIP5 led to many changes in the way users interact with the archive, which impacts management of information about users and complicates communications with them.
   - In particular, a growing number of data users no longer register or interact directly with the ESGF. Rather they rely on secondary repositories, often "snapshots" of the state of some portion of the ESGF archive created by others at a particular time (see for instance the IPCC CMIP5 Data Factsheet for a discussion of the snapshots and their coverage). This meant that reliance on the ESGF's inventory of registered users for any aspect of the infrastructure – such as tracking usage, compliance with licensing requirements, or informing users about errata or retractions – could at best ensure partial coverage of the user base.

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2018-52
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 22 March 2018

– The WIP therefore committed to a more distributed design for several features outlined below, which devolve many of these features to the datasets themselves rather than the archives. One may think of this as a *dataset-centric rather than system-centric* design (in software terms, a *pull* rather than *push* design): information is made available upon request at the user/dataset level, relieving the ESGF implementation of an impossible burden.

Based upon these considerations, the WIP produced a set of position papers (see Appendix A) encapsulating specifications and recommendations for CMIP6 and beyond. These papers, summarized below, are available from the WIP website. As the WIP continues to develop additional recommendations, they too will be made available. All WIP papers distributed in this way are thought be stable, but should revision be necessary, a modified document will be released with a new version number.

3 A structured approach to data production

The CMIP6 data framework has evolved considerably from CMIP5, and follows the principles of scientific reproducibility (Item 3 in Section 2), and the recognition that the complexity of the experimental design (Item 6) required far greater degrees of automation and embedding in workflows. This requires that all elements in the specification be recorded in structured text formats (XML and JSON, for example), and subject to rigorous version control. *Machine-readable* specification of as many aspects of the model output configuration as possible is a WIP design goal.

The data request spans several elements discussed in sub-sections below.

3.1 CMIP6 Data Request

The data request (Juckes et al., 2015) is now available through the DREQ tool, the associated `dreqPy` Python library, and underlying database. The DREQ combines definitions of variables and their output format with specifications of the objectives they support and the experiments that they are required for. The entire request is encoded in an XML database with rigorous type constraints. Important elements of the request, such as units, cell methods (expressing the subgrid processing implicit in the variable definition), and time slices for required output, are defined as controlled vocabularies within the request to ensure consistency of usage. The request is designed to enable flexibility, allowing modeling centers to make informed decisions about the variables they should submit to the CMIP6 archive from each experiment.

The data request spans several elements.

1. specification of the parameter to be calculated in terms of a CF standard name and units,
2. an output frequency,
3. a structural specification which includes specification of dimensions and of subgrid processing.

In order to facilitate the cross linking between the 2100 variables from 248 experiments, the request database allows MIPs to aggregate variables and experiments into groups. The link between variables and experiments is then made through the following chain:

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2018-52
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 22 March 2018

1. A *variable group*, aggregating variables with priorities specific to the MIP defining the group;

2. A *request link* associating a variable group with an objective and a set of request items;

3. *Request* items associating a particular time slice with a request link and a set of experiments.

This formulation takes into account the complexities that arise when a particular MIP requests that variables needed for their own experiments should also be saved from a DECK experiment or from an experiment proposed by a different MIP.

The data request supports a broad range of users who are provided with a range of different access points.

1. The XML database provides the reference document;

2. Web pages provide a direct representation of the database content;

3. Excel workbooks provide selected overviews for specific MIPs and experiments;

4. A python library provides an interface to the database with some built-in support functions;

5. A command line tool based on the python library allows quick access to simple queries.

The data request's machine-readable database, which is accessible through a simple python API, has been an extraordinary resource for the modeling centers. They can, for example, directly integrate the request specifications with their workflows to ensure that the correct set of variables are saved for each experiment they plan to run. In addition, it has given them a new-found ability to estimate the data volume associated with meeting a MIP's requirements, a feature exploited below in Section 3.4.

3.2 Model inputs

Datasets used by the model for configuration of model inputs (`input4MIPs`, see Durack et al., 2017) as well as observations for comparison with models (`obs4MIPs`, see Teixeira et al., 2014) are both now organized in the same way, and share many of the naming and metadata conventions as the CMIP model output itself. The datasets follow versioning methodologies recommended by the WIP.

3.3 Data Reference Syntax

The organization of the model output follows the Data Reference Syntax (DRS) first used in CMIP5, and now in somewhat modified form in CMIP6. The DRS depends on pre-defined *controlled vocabularies* (CVs) for various terms including: the names of institutions, models, experiments, time frequencies, etc. The CVs are now recorded as a version-controlled set of structured text documents, and the WIP has taken steps to ensure that there is a single authoritative source for any CV, on which all elements in the toolchain will rely. The DRS elements that rely on these controlled vocabularies appear as netCDF attributes and are used in constructing file names, directory names, and unique identifiers of datasets that are essential throughout the CMIP6 infrastructure. These aspects are covered in detail in the CMIP6 Global Attributes, DRS, Filenames, Directory

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2018-52
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 22 March 2018

Structure, and CVs position paper. A new element in the DRS indicates whether data has been stored on a native grid or has been regridded (see discussion below in Section 3.4 on the potentially critical role of regridded output). This element of the DRS will allow us to track the usage of the *regridded subset* of data, and assess the relative popularity of native-grid vs. standard-grid output.

3.4 CMIP6 data volumes

As noted, extrapolations based on CMIP3 and CMIP5 lead to some alarming trends in data volume (see e.g., Overpeck et al., 2011). The WIP has undertaken a rigorous approach to the estimation of future data volumes, rather than simple extrapolation. Contributions to increase in data volume include the systematic increase in model resolution and complexity of the experimental protocol and data request. We consider these separately:

Resolution The median horizontal resolution of a CMIP model tends to grow with time, and is expected to be more typically 100 km in CMIP6, compared to 200 km in CMIP5. The vertical resolution grows in a more controlled fashion, at least as far as the data is concerned, as often the requested output is reported on a standard set of atmospheric levels that has not changed much over the years. Similarly the temporal resolution of the data request does not increase at the same rate as the model timestep: monthly averages remain monthly averages. A doubling of model resolution leads therefore to a quadrupling of the data volume, in principle. But typically the temporal resolution of the model (though not the data) is doubled as well, for reasons of numerical stability. Thus, for an N-fold increase in horizontal resolution, we require an N^3 increase in computational capacity, which will result in an N^2 increase in data volume. We argue therefore, that data volume V and computational capacity C are related as $V \sim C^{\frac{2}{3}}$, purely from the point of view of resolution. The exponent is even smaller if vertical resolution increases are assumed. If we then assume that centers will experience an 8-fold increase in C between CMIPs (which is optimistic in an era of tight budgets), we can expect a 4-fold increase in data volume. However, this is not what we experienced between CMIP3 and CMIP5. What caused that extraordinary 50-fold increase in data volume?

Complexity The answer lies in the complexity of CMIP: the complexity of the data request, and of the experimental protocol. The data request complexity is related to that of the science: the number of processes being studied, and the physical variables required for the study. In CPMIP (Balaji et al., 2017), we have attempted a rigorous definition of this complexity, measured by the number of physical variables simulated by the model. This, we argue, grows not smoothly like resolution, but in very distinct generational step transitions, such as the one from atmosphere-ocean models to Earth system models, which involved a substantial jump in complexity, the number of physical, chemical, and biological species being modeled, as shown in Balaji et al. (2017).

The second component of complexity is the experimental protocol, and the number of experiments themselves when comparing CMIP5 and CMIP6. With the new structure of CMIP6, with a DECK and 21 endorsed MIPs, this would appear to have grown tremendously. We propose as a measure of experimental complexity, the *total number of simulated years (SYs)* conforming to a given protocol. Note that this too is gated by C: modeling centers usually make tradeoffs

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2018-52
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 22 March 2018

between experimental complexity and resolution in deciding their level of participation in CMIP6, discussed in Balaji et al. (2017).

The WIP has recommended two further steps toward ensuring sustainable growth in data volumes.

1. The first of these is the consideration of standard horizontal resolutions for saving data, as is already done for vertical and temporal resolution in the data request. Cross-model analyses already cast all data to a common grid in order to evaluate it as an ensemble, typically at fairly low resolution. The studies of Knutti and colleagues (e.g., Knutti et al. (2017)) are typically performed on relatively coarse grids. We recommend that for most purposes atmospheric data on the ERA-40 grid ($2^\circ \times 2.5^\circ$) would suffice, with of course exceptions for experiments like those called for by HighResMIP (Haarsma et al., 2016). A similar recommendation is made for ocean data (the World Ocean Atlas $1^\circ \times 1^\circ$ grid), with extended discussion of the benefits and losses due to regridding (see Griffies et al., 2014, 2016). Regridding remains a contentious topic, and owing to a lack of consensus, the WIP recommendations on regridding remain in flux. The CMIP6 Output Grid Guidance document outlines a number of possible recommendations, including the provision of "weights" to a target grid. Many of the considerations around regridding, particularly for ocean data in CMIP6, are discussed at length in Griffies et al. (2016). A similar lack of consensus has made the WIP drop a recommendation of a common *calendar* for particular experiments: a wide variety of calendars are in use – Gregorian, Julian, 365-day, and equal-month (360-day) all remain popular options – and the onus of converting data across the multi-model ensemble (MME) to a common one for analysis remains upon the end-user.

   As outlined below in Section 6, both ESGF data nodes and the creators of secondary repositories are given considerable leeway in choosing data subsets for replication, based on their own interests. The tracking mechanisms outlined in Section 5.2 below will allow us to ascertain, after the fact, how widely used the native grid data may be *vis-à-vis* the regridded subset, and allow us to recalibrate the replicas, as usage data becomes available. We note also that the providers of at least one of the standard metrics packages (ESMValTool, Eyring et al., 2016a) have expressed a preference of standard grid data for their analysis, as regridding from disparate grids increases the complexity of their already overburdened infrastructure.

2. The second is the issue of data compression. netCDF4, which is the WIP's required standard for CMIP6 data, includes an option for lossless compression or deflation (Ziv and Lempel, 1977) that relies on the same technique used in standard tools such as `gzip`. In practice, the reduction in data volume will depend upon the "entropy" or randomness in the data, with smoother data being compressed more.

   Deflation entails computational costs, not only during creation of the compressed data, but also every time the data are re-inflated. There is also a subtle interplay with precision: for instance temperatures usually seen in climate models appear to deflate better when expressed in Kelvin, rather than Celsius, but that is due to the fact that the leading order bits are always the same, and thus the data is actually less precise. Deflation is also enhanced by reorganizing ("shuffling") the data internally into chunks that have spatial and temporal coherence.

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2018-52
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 22 March 2018

Some in the community argue for the use of more aggressive *lossy* compression methods (Baker et al., 2016), but the WIP, after consideration, believes the loss of precision entailed by such methods, and the consequences for scientific results, require considerably more evaluation by the community before such methods can be accepted as common practice.

Given the options above, we undertook a systematic study of the behavior of typical model output files under lossless compression, the results of which are publicly available. The study indicates that standard `zlib` compression in the netCDF4 library with the settings of `deflate=2` (relatively modest, and computationally inexpensive), and `shuffle` (which ensures better spatiotemporal homogeneity) ensures the best compromise between increased computational cost and reduced data volume. For a coupled model, we expect a total savings of about 50%, with ocean, ice, land realms getting the most savings (owing to large areas of the globe that are masked), and atmospheric data the least. This 50% estimate has been verified with sample output from some models preparing for CMIP6.

The DREQ alluded to above in Section 3 allows us to make a systematic assessment of these considerations. The tool expects one to input a model's resolution along with the experiments that will be performed and the data one intends to save (using DREQ's *priority* attribute). With this information dreqDataVol.py, which is a tool built atop DREQ available from the WIP website calculates the data volume that will be produced. While similar analyses were undertaken at PCMDI for CMIP5, this tool puts this capability in the hands of the modeling centers themselves.

To make a preliminary estimate of total data volume, the WIP carried out a survey of modeling centers in 2016, asking them for their expected model resolutions, and intentions of participating in various experiments. Based on that survey, we initially have forecast a data volume of 18 PB for CMIP6. This assumes an overall 50% compression rate, which has been approximately verified for at least one CMIP6 model, and whose compression rates should be quite typical. This number, 18 PB, is about 6 times the CMIP archive size, and can be explained in terms of the compounding of modest increases in resolution and complexity, as explained above. The more dramatic increase in data volume between CMIP3 and CMIP5 was also due to these same causes, but with a much larger change. Many models of the CMIP5 era added atmospheric chemistry and aerosol-cloud feedbacks, sometimes with $\mathcal{O}(100)$ species. CMIP5 also marked the first time in CMIP that ESMs were used to simulate changes in the carbon cycle and modeling groups performed many more simulations than in CMIP3 with a corresponding increase in years simulated. There is no comparable jump between CMIP5 and CMIP6. CMIP6's innovative DECK/endorsed-MIP structure should thus be seen as an extension and an attempt to impose a rational order on CMIP5, rather than a qualitative leap.

It should be noted that reporting output on a lower resolution standard grid (rather than the native model grid) could shrink this volume 10-fold, to 1.8 PB. This is an important number, as will be seen below in Section 6: the managers of Tier 1 nodes have indicated that 2 PB is about the practical limit for replicated storage of combined data from all models. The WIP believes this target is achievable based on compression and the use of standard grids. Both of these (the use of netCDF4 compression and regridding) remain merely recommendations, and the centers are free to choose whether or not to compress and regrid.

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2018-52
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 22 March 2018

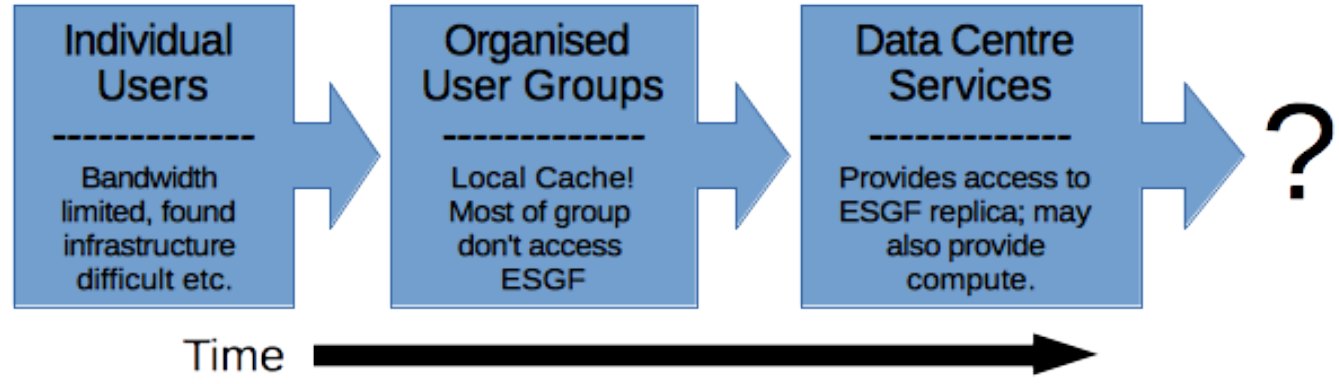


Figure 2. Typical data usage pattern in CMIP5 involved users making local copies, and user groups making institutional-scale caches from ESGF. Figure courtesy Stephan Kindermann, DKRZ, adapted from WIP Licensing White Paper.

4 Licensing

The WIP's recommended licensing policy is based on an examination of data usage patterns in CMIP5. First, while the licensing policy called for registration and acceptance of the terms of use, a large fraction, perhaps a majority of users, actually obtained their data not directly from ESGF, but from other copies, such as the "snapshots" alluded to above in Item 7, Section 2. Those users accessing the data indirectly, as shown in Figure 2, relied on user groups or their home institutions to make secondary repositories that could be more conveniently accessed. The WIP CMIP6 Licensing and Access Control position paper refers to the secondary repositories as "dark" and those obtaining CMIP data from those reposistories as "dark users" who are invisible to the ESGF system. While this appears to subvert the licensing and registration policy put in place for CMIP5, this should not be seen as a "bootleg" process: it is in fact the most efficient use of limited network bandwidth at the user sites. However, this also removes the ability for users of these "dark" repositories to benefit from the augmented provenance provided by infrastructure updates, such as being notified of data retractions or replacements in the case that contributed datasets are found to be erroneous and replaced.

The WIP therefore recommends a licensing policy that inverts this and removes the impossible task of license enforcement from the distribution system, and embraces the "dark" repositories and users. To quote the WIP position paper:

> The proposal is that (1) a data license be embedded in the data files, making it impossible for users to avoid having a copy of the license, and (2) the onus on defending the provisions of the license be on the original modeling center...

The data archive snapshots and emerging resources that combine archival and analysis capabilities (e.g., NCAR's CMIP Analysis Platform) will host data and offload some of the network provisioning requirements from ESGF nodes themselves.

Modeling centers are offered two choices of *Creative Commons* licenses: data covered by the Creative Commons Attribution "Share Alike" 4.0 International License will be freely available; for centers with more restrictive policies, the Creative Com-

mons Attribution "NonCommercial Share Alike" 4.0 International License, which restricts the data to non-commercial use. Further sharing of the data is allowed, as the license travels with the data. The PCMDI website provides a link to the current CMIP6 Terms of Use webpage.

5 Citation and provenance

As noted in Section 2, the WIP's position on citation flows from two underlying considerations: one, to provide proper credit and formal acknowledgment of the authors of datasets; and the other, to enable rigorous tracking of data provenance and data usage. The tracking facilitates scientific reproducibility and traceability, as well as enabling statistical analyses of dataset utility.

In addition to clearly identifying what data have been used in research studies and who deserves credit for providing that data, it is essential that the data be examined for quality and that documentation be made available describing the model and experiment conditions under which it was generated. These subjects are addressed in the four position papers summarized in this section.

The principles outlined above are well-aligned with the Joint Declaration of Data Citation Principles formulated by the Force11 (The Future of Research Communications and e-Scholarship) Consortium, which has acknowledged the rapid evolution of digital scholarship and archival, as well as the need to update the rules of scholarly publication for the digital age. We are convinced that not only peer-reviewed publications but also the data itself should now be considered a first-class product of the research enterprise. This means that data requires curation and should be treated with the same care as journal articles. Moreover, most journals and academies now insist that data used in the literature be made publicly available for independent inquiry and reproduction of results. New services like Scholix are evolving to support the exchange and access of such data-data and data-literature interlinking.

Given the complexity of the CMIP6 data request, we expect, as shown in Section 3.4, a total dataset count of $\mathcal{O}(10^6)$. Because dozens of datasets are typically used in a single scientific study, it is impractical to cite each dataset individually in the same way as individual research publications are acknowledged. The WIP therefore offers an option of citing data and giving credit to data providers that relies on a rather coarse granularity, while at the same time offering another option at a much finer granularity for recording the specific files and datasets used in a study.

In the following, two distinct types of persistent identifiers (PIDs) are discussed: DOIs, which can only be assigned to data that comply with certain standards for citation metadata and curation, and the more generic "Handles"that have fewer constraints and may be more easily adapted for a particular use. Technically both types of PIDs rely on the underlying global Handle System to provide services (e.g., to resolve the PIDs and provide associated metadata, such as the location of the data itself).

5.1 Persistent identifiers for acknowledgment and citation

Based on earlier phases of CMIP, some datasets initially contributed to the CMIP6 archive will be flawed (due, for example, to errors in processing) and therefore will not accurately represent a model's behavior. When errors are uncovered in the datasets,

they may be replaced with corrected versions. Similarly, additional datasets may be added to an initially incomplete collection of datasets. Thus, initially at least, the DOIs assigned for the purposes of citation and acknowledgement will represent an evolving underlying collection of datasets.

The recommendations, detailed in the CMIP6 Data Citation and Long Term Archival position paper, recognize two phases to the process of assigning DOI's to collections of datasets: an initial phase, when the data have been released and preliminary community analysis is still underway and a second stage when most errors in the data have been identified and corrected. Upon reaching stage two, the data will be transferred to long-term archival (LTA) of the IPCC Data Distribution Centre (IPCC DDC) and deemed appropriate for interdisciplinary use (e.g., in policy studies). The timing of the planned DDC snapshot is linked to the IPCC AR6 schedule.

For evolving dataset aggregations, the data citation infrastructure relies on information collected from the data providers and uses the DataCite data infrastructure to assign DOIs and record associated metadata. DataCite is a leading global non-profit organisation that provides persistent identifiers (DOIs) for research data. The DOIs will be assigned to:

1. aggregations that include all the datasets contributed by one model from one institution from all of a single MIP's experiments, and

2. aggregations that include all datasets contributed by one model from one institution generated in performing one experiment (which might include one or more simulations).

These aggregations are dynamic as far as the PID infrastructure is concerned: new elements can be added to the aggregation without modifying the PID. As an example, for the coarser of the two aggregations defined above, the same PID will apply to an evolving number of simulations as new experiments are performed with the model. This PID architecture is shown in Figure 3. Since these collections are dynamic, citation requires authors to provide a version reference.

For the stable dataset collections, the data citation infrastructure requires some additional steps to meet formal requirements. First, we ensure that there has been sufficient community examination of the data to qualify it as having been informally peer reviewed. Second, further steps are undertaken to assure important information exists in ancillary metadata repositories, including, for example, documentation (ES-DOC, errata and citation) and to provide quality assurance of data and metadata consistency and completeness (see Section 5.3). Once these criteria have been satisfied, a DOI will be issued by the IPCC DDC hosted by DKRZ. These dataset collections will meet the stringent metadata and documentation requirements of the IPCC DDC. Since these collections are static, no version reference is required in a citation.

The WIP's position is that for CMIP6, the initially assigned DOIs (associated with evolving collections of data) must be used in research papers to properly give credit to each of the modeling groups providing the data. Once a stable collection of datasets has met the higher standards for long-term curation and quality, the DOI assigned by the IPCC DDC should be used instead.

The data citation approach is described in greater detail in Stockhause and Lautenschlager (2017).

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2018-52
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 22 March 2018

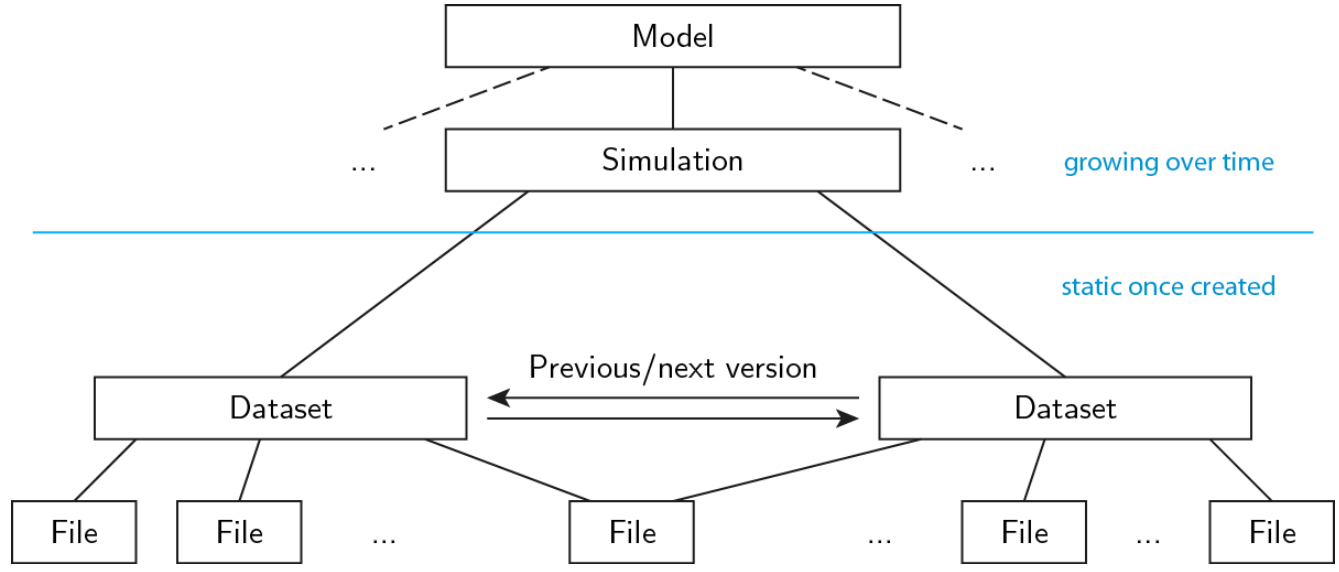


Figure 3. PID architecture, showing layers in the PID hierarchy. In the lower layers of the hierarchy, PIDs are static once generated, and new datasets generate new versions with new PIDs.

5.2 Persistent identifiers for tracking, provenance, and curation

Although the DOIs assigned to relatively large aggregations of datasets are well suited for citation and acknowledgment purposes, they are not issued at fine enough granularity to meet the scientific imperative that published results should be traceable and verifiable. Furthermore, management of the CMIP6 archive requires that PIDs be assigned at a much finer granularity than the DOIs. For these purposes, PIDs recognized by the global Handle registry will be assigned at two different levels of granularity:

A unique Handle will be generated each time a new CMIP6 data file is created, and the Handle will be recorded in the file's metadata (in the form of a netCDF global attribute named `tracking_id`). At the time the data is published, the `tracking_id` will be processed by the CMIP6 Handle service infrastructure and recorded in the ESGF metadata catalog. Another Handle will subsequently be assigned at somewhat coarser granularity to each aggregation of files containing the data from a single variable sampled at a single frequency from a single model running a single experiment. In ESGF terminology, this collection of files is referred to as an *atomic dataset*.

As described in the CMIP6 Persistent Identifiers Implementation Plan position paper, a Handle assigned at either of these two levels of the PID hierarchy identifies a static entity; if any file associated with a Handle is altered in any way a new Handle must be created. The PID infrastructure is also central to the replication and versioning strategies, as described in Section 6 and Section 7 below. Furthermore, as a means of recording provenance and enabling tracking of dataset usage, the

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2018-52
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 22 March 2018

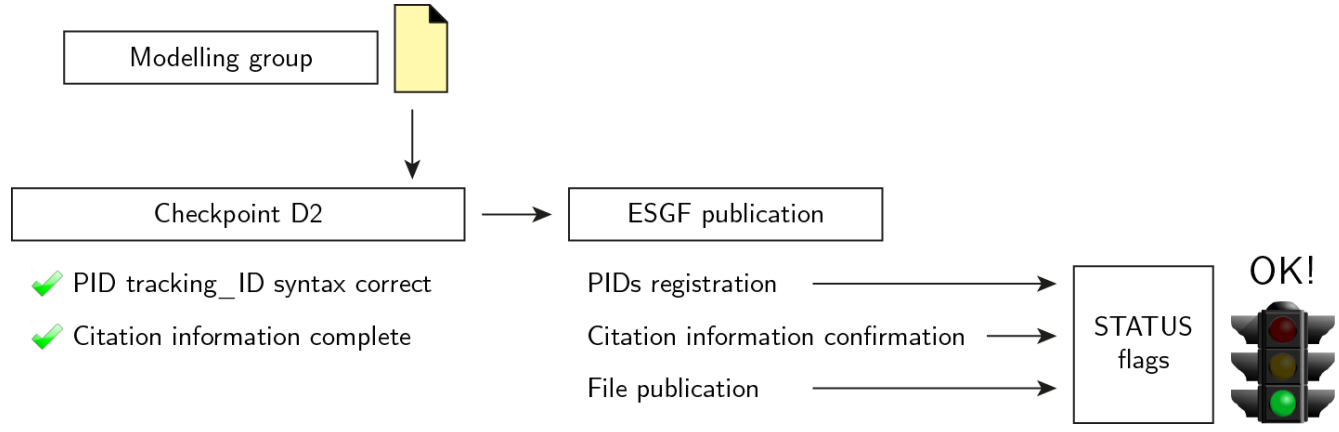


Figure 4. PID workflow, showing the generation and registry of PIDs, with checkpoints where compliance is assured.

WIP urges authors to include as supplementary material attached to each CMIP6-based publication a PID list (a flat list of all PIDs referenced).

The implementation plan describes methods for generating and registering Handles using an asynchronous messaging system known as RabbitMQ. This system, designed in collaboration with ESGF developers and shown in Figure 4, guarantees, for example, that PIDs are correctly generated in accordance with the versioning guidelines. The CMIP6 handle system builds on the idea of tracking-ids used in CMIP5, but with a more rigorous quality control to ensure that new PIDs are generated when data are modified. The dataset and file Handles are also associated with basic metadata, called PID Kernel information (Zhou et al., 2018), which facilitate the recording of basic provenance information. Datasets and files point to each other to bind the granularities together. In addition, dataset kernel information refers to previous and later versions, errata information and replicas, explained in more detail in the position paper.

5.3 Quality Assurance

The WIP's perspective on quality assurance (QA) encompasses the entire data lifecycle, as depicted in Figure 5. At all stages, a goal is to capture provenance information that will enable scientific reproducibility. Further, as noted in Item 2 in Section 2, the QA procedures should uncover issues that might undermine trust in the data by those outside the Earth system modeling community if errors were left unreported.

QA must ensure that the data and metadata correctly reflect a model's simulation, so that it can be reliably used for scientific purposes. As depicted in Figure 5, the first stage of QA is the responsibility of the data producer: in fact the cycle of model development and diagnosis is the most critical element of QA. The second aspect is ensuring that disseminated data include common metadata based on common CVs, which will enable consistent treatment of data from different groups and institutions. These requirements are directly embedded in the ESGF publishing process and in tools such as CMOR (and its validation component, PrePARE). These checks (the D1 and M1 phases of QA in Figure 5) ensure that the data conform to the CMIP6

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2018-52
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 22 March 2018

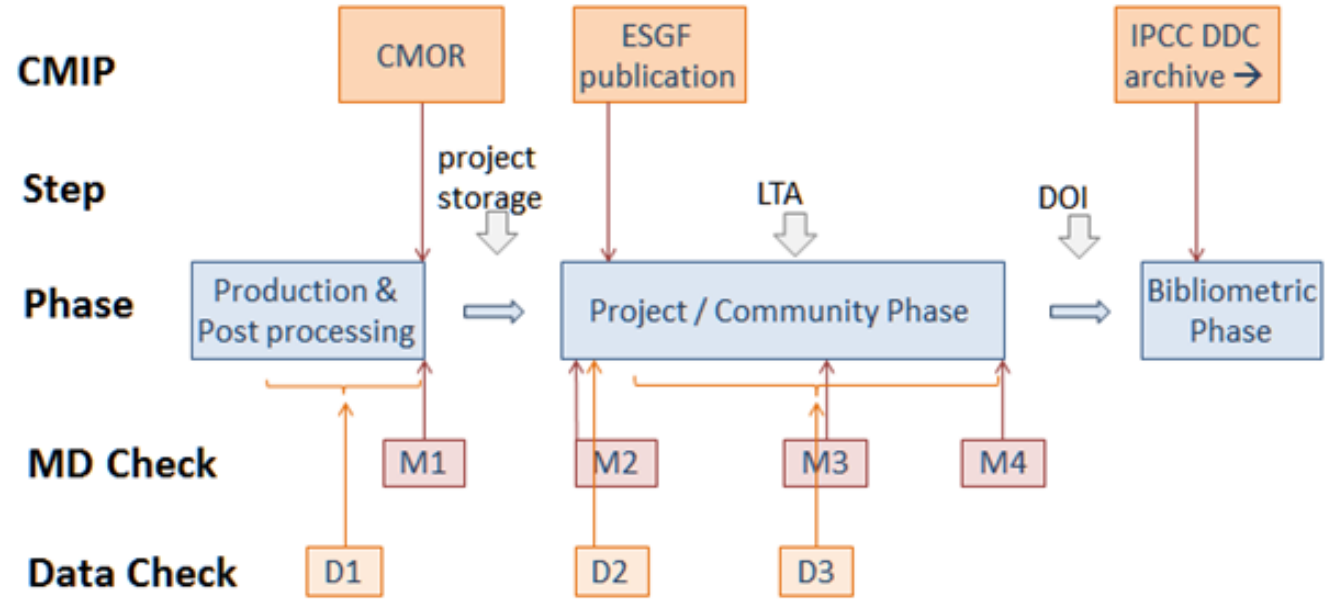


Figure 5. Schematic of the phases of quality assurance, with earlier stages in the hands of modeling centers (left), and more formal long-term data curation stages at right. Quality assurance is applied both to the data (D, above) as well as the metadata (M) describing the data. Figure drawn from the WIP's Quality Assurance position paper.

Data Request specifications, conform to all naming conventions and CVs, and follow the mandated structure for organization into a common directory structure. As noted in Section 3, many modeling centers have chosen to embed these steps directly in their workflows to ensure conformance with the CMIP6 as the models are being run and their output processed.

At this point, as noted in Figure 5, control is ceded to the ESGF system, where designated QA nodes perform further QA checks. A critical step is the assignment of PIDs (Section 5.2, the D2 stage of Figure 4), which is more controlled than in CMIP5 and guarantees that across the data lifecycle, the PIDs will be reliably useful as unique labels of datasets.

Beyond this, further stages of QA will be handled within the ESGF system following procedures outlined in the CMIP6 Quality Assurance position paper. As described before, once data have been published, the data will be scrutinized by researchers in what can be considered an ongoing period of community-wide scientific QA of the data. During this period, modeling centers may correct errors and provide new versions of datasets. In the final stage, the data pass into long term archival (LTA) status, described as the "bibliometric" phase in Figure 5. Just prior to LTA, the system will verify minimum standards of provenance documentation. This is described in the next section.

5.4 Documentation of provenance

As noted earlier in Section 3, for data to become a first-class scientific resource, the methods of their production must be documented to the fullest extent possible. For CMIP6, this includes documenting both the models and the experiments. While

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2018-52
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 22 March 2018

Geoscientific Model Development Discussions

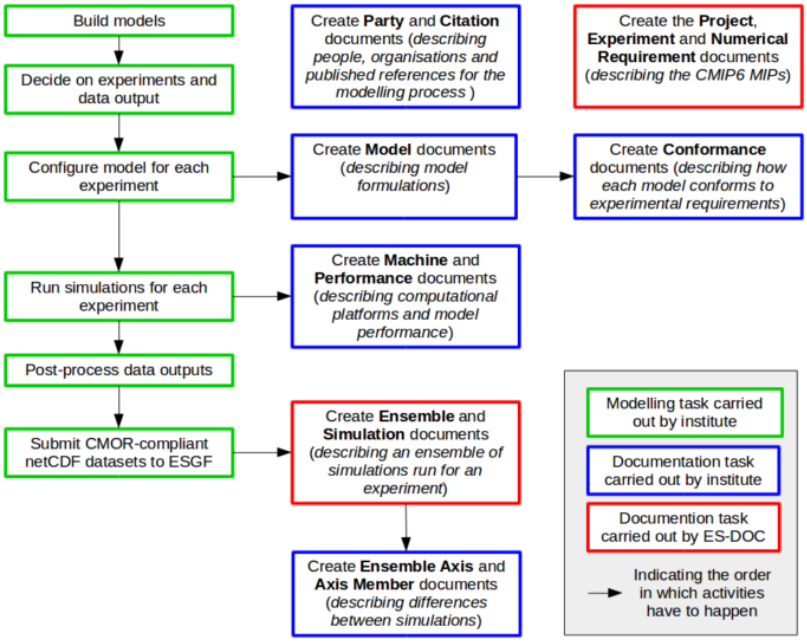


Figure 6. Flowchart of ES-DOC documentation process, delineating sequence of events and indicating the parties responsible for producing the documentation. Figure courtesy Eric Guilyardi and Mark Greenslade.

traditionally this is done through peer-reviewed literature, which remains essential, we note that to facilitate various aspects of search, discovery and tracking of datasets, there is an additional need for structured documentation in machine readable form.

In CMIP6, the documentation of *experiments*, *models* and *simulations* is done through the Earth System Documentation (ES-DOC, Guilyardi et al., 2013) Project. The various aspects of model documentation are shown in Figure 6, and in greater detail in the WIP position paper on ES-DOC. The CMIP6 experimental design has been translated into structured text documents, already available from ES-DOC. ES-DOC has constructed CVs for the description of the CMIP6 standard model realms, including a set of short tables (*specialisations*, in ES-DOC terminology) for each realm. The WIP, and the CMIP Panel, recommend that the modeling groups integrate with their model development process their provision of documentation to ES-DOC. This will better ensure the accuracy and consistency of the documentation. ES-DOC provides a variety of user interfaces to read and write structured documentation that conforms with the Common Information Model (CIM) of Lawrence et al. (2012). As models evolve or differentiate (for example, an Earth system model derived from a particular general circulation model), branches and new versions of the documentation can be produced in a manner familiar to anyone who works with version-controlled code.

A critical element in the ES-DOC process is the documentation of *conformances*: steps undertaken by the modeling centers to ensure that the simulation was conducted as called for by the experiment design. It is here that we rigorously document

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2018-52
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 22 March 2018

which input datasets were used in a simulation (e.g., the version of each of the forcing datasets, see Durack et al., 2017). The conformances will be an important element in guiding selection of subsets of CMIP6 model results for particular research studies. A researcher might, for example, choose to subselect only those models that used a particular version of the forcing datasets that are imposed as part of the experimental protocol. The conformances will continue to grow in importance under the CMIP vision that the DECK will provide an ongoing foundation on which to build a series of future CMIP phases (shown schematically in Figure 1 of Eyring et al., 2016a). The conformances will be essential in enabling studies across model generations.

The method of capturing the conformance documentation is a two-stage process that has been designed to minimize the amount of work required by a modeling center. The first stage is to capture the many conformances common to all simulations. ES-DOC will then automatically copy these common conformances to multiple simulations thereby eliminating duplicated effort. This is followed by a second stage in which those conformances that are specific to individual experiments or simulations are collected.

While this method of documentation is unfamiliar to many, the WIP emphasizes how important it is destined to become in the maturing digital age as part of best scientific practices. Documentation of software validation (see e.g Peng, 2011) and structured documentation of complete scientific workflows that can be independently read and processed are both becoming more common (see the special issue on the "Geoscience Paper of the Future", David et al., 2016). We have noted earlier (see Item 3 in Section 2 the special importance in climate research today of documenting how results have been obtained and enabling results to be reproduced by others. Rigorous documentation remains a hardy bulwark against challenges to the scientific process.

In keeping with the WIP's "dataset-centric rather than system-centric" approach (Item 7 in Section 2), a user will be directly linked to documentation from each dataset. This is done in CMIP6 by embedding a global attribute `further_info_url` in file headers pointing to the associated CIM document, which will serve as the landing page for documentation from which further exploration (by humans or software) will take place. The existence and functioning of the landing page is assured in Stage M3 of Figure 5.

6 Replication

The WIP's replication strategy is covered in the CMIP6 Replication and Versioning position paper. The recommendations therein are based on the following *primary* goal:

- Ensuring at least one copy of a dataset is present at a stable ESGF node with a mission of long-term maintenance and curation of data. The total data storage resources planned across the Tier 1 nodes in the CMIP6 era is adequate to support this requirement, though some data will likely be held on accessible tape storage rather than spinning disk.

In addition, we have articulated a number of secondary goals:

- Enhancing data accessibility across the ESGF (e.g. Australian data easily accessible to the European continent despite the long distance);
- Enabling each Tier 1 data node to enact specific policies to support their local objectives;
- Ensuring that the most widely requested data is the most accessible across the ESGF federation;
- Enabling large-scale data analysis across the federation (see Item 4 in Section 2);
- Ensuring continuity of data access in the event of individual node failures;
- Enabling network load-balancing and enhanced performance;
- Reducing the manual workload related to replication;
- Building a reliable replication mechanism that can be used not only within the federation, but by the secondary repositories created by user groups (see discussion in Section 4 around Figure 2).

In conjunction with the ESGF and the International Climate Networking Working Group (ICNWG), these recommendations have been translated to a two-pronged strategy.

The basic toolchain for replication is built on updated versions of the software layers used in CMIP5 including: synda (formerly `synchrodata`) and Globus Online (Chard et al., 2015), which are based on underlying data transport mechanisms such as gridftp and the older and now deprecated protocols like `wget` and `ftp`.

As before, these layers can be used for *ad hoc* replication by sites or user groups. For *ad hoc* replication, there is no obvious mechanism for triggering updates or replication when new data are published (or retracted, see Section 7 below). Therefore, the WIP recommends that designated *replica nodes* maintain a protocol for automatic replication, shown in Figure 7.

Given the nature of some of the secondary goals listed above, it would not be appropriate for the WIP to prescribe which data should be replicated by each center. Rather, the plan should be flexible to accommodate changing data use profiles and resource availability. The WIP consider the CDNOT group to be the appropriate organisation to coordinate the replication activities of the CMIP6 data nodes such that the primary goal is achieved and an effective compromise for the secondary goals is established.

The International Climate Network Working Group (ICNWG), formed under the Earth System Grid Federation (ESGF), helps set up and optimize network infrastructures for ESGF climate data sites located around the world. For example prioritising the most widely requested data for replication can best be done based on operational experience and will of course change over time. To ensure that the replication strategy is responding to user need and data node capabilities, the replication team will maintain and run a set of monitoring and notification tools assuring that replicas are up-to-date. The CDNOT is tasked with ensuring the deployment and smooth functioning of replica nodes.

A key issue that emerged from discussions with node managers is that the replication target has to be of sustainable size. The WIP has concluded from the discussions that a replication target about 2 PB in size is the practical (technical and financial)

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2018-52
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 22 March 2018

Geoscientific Model Development Discussions

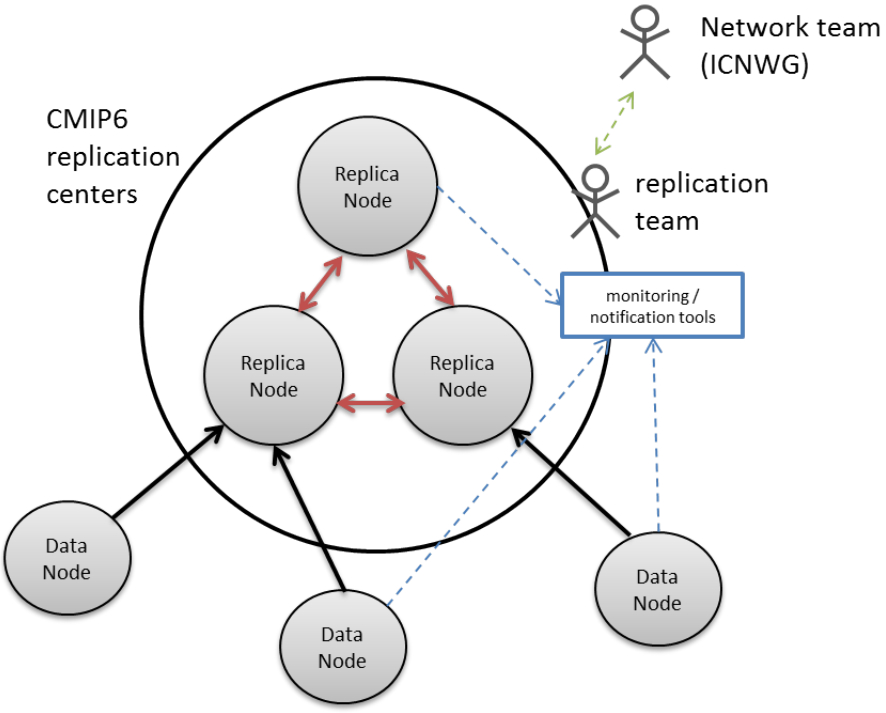


Figure 7. CMIP6 replication from data nodes to replica centers and between replica centers coordinated by a CMIP6 replication team.

limit for CMIP6 online (disk) storage at any single location. Replication beyond this may involve offline storage (tape) for disaster recovery.

Based on experience in CMIP5, it is expected that a number of "special interest" secondary repositories will hold selected subsets of CMIP6 data outside of the ESGF federation. This will have the effect of widening data accessibility geographically, and by user communities, with obvious benefit to the CMIP6 program. The WIP encourages the support of these secondary repositories where it does not undermine CMIP6 data management and integrity objectives.

In CMIP5 a significant issue for users of some third-party archives was that their replicated data was taken as a one-time snapshot (see discussion above in Item 7 in Section 2), and not updated as new versions of the data were submitted to the source ESGF node. Tools have been developed by a number of organisations to maintain locally synchronized archives of CMIP5 data and third party providers should be encouraged to make use of these types of tools to keep the local archives up to date.

In summary, the WIP requirements for replication are limited to ensuring:

- that there is at least one instance of each submitted dataset stored at a Tier 1 node (in addition to its primary residence) within a reasonably short time period following submission;
- that subsequent versions of submitted datasets are also replicated by at least one Tier 1 node (see versioning discussion below in Section 7);

- that creators of secondary repositories take advantage of the replication toolchain described here, to maintain replicas that can be kept up to date, rather than one-time snapshots
- that the CDNOT is the recognized body to manage the operational replication strategy for CMIP6.

7 Versioning

The WIP position on versioning is based on the principle (Section 2) of scientific reproducibility. Recognizing that errors may be found after datasets have been distributed, the WIP insists that erroneous datasets that may have been used downstream continue to be publicly available, but marked as superseded. This will allow users to trace the provenance of published results, even if those point to retracted data; and further allow the possibility of *a posteriori* correction of such results.

The WIP requires a consistent versioning methodology across all the ESGF data nodes. We note that inconsistent or informal versioning practices at individual nodes would likely be invisible to the ESGF infrastructure (e.g., yielding files that look like replicas, but with inconsistent data and checksums), which would inhibit traceability across versions.

In close consultation with the ESGF implementation teams, the WIP has made the following recommendations, described in greater depth in the CMIP6 Replication and Versioning position paper:

- the PID infrastructure of Section 5 is the basis of creating versions of datasets. PIDs are permanently associated with a dataset, and new versions will get a new PID. When new versions are published, there will be two-way links created within the PID kernel information so that one may query a PID for prior or subsequent versions.
- we recommend the unit of versioning be an *atomic dataset*: a complete timeseries of one variable from one experiment and one model. The implication is that other variables need not be republished, if the error is found in a single variable. If an entire experiment is retracted and republished, all variables will get a consistent version number.
- the CDNOT will ensure consistent versioning practices at all participating data nodes.

7.1 Errata

It is worth highlighting in particular the new recommendations regarding errata. Until CMIP5, we have relied on the ESGF system to push notifications to registered users regarding retractions and reported errors. This was found to result in imperfect coverage: as noted in Section 4, a substantial fraction of users are invisible to the ESGF system. Therefore, following the discussion in Section 2 (see Item 7), we have recommended a design which is dataset-centric rather than system-centric. Notifications are no longer pushed to users; rather they will be able to query the status of a dataset they are working with. An *errata client* will allow the user to enter a PID to query its status; and an *errata server* will return the PIDs associated with prior or posterior versions of that dataset, if any. Details are to be found in the Errata position paper.

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2018-52
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 22 March 2018

8 The future of the global data infrastructure

The WIP was formed in response to the explosive growth of CMIP between CMIP3 and CMIP5, and charged with studying and making recommendations about the global data infrastructure needed to support CMIP6 and the future evolution of inter-comparison projects. Our findings reflect the fact that CMIP is no longer a cottage industry, and a more formal approach is needed. The resulting recommendations stop well short of any sort of global governance of this "vast machine", but list many areas where, with a relatively light touch, beneficial order and control result. We emphasize here again some of the key aspects of the design:

- The design is now dataset-centric rather than system-centric: see for example the discussion of licensing (Section 4) and dataset tracking (Section 5.2). This relieves a considerable design burden from the ESGF software stack, and further, recognizes that the data ecosystem extends well beyond the reach of any software system and that data will be used and reused in myriad ways outside anyone's control.

- Standards, conventions, and vocabularies are now stored in machine-readable structured text formats like XML and JSON, thereby enabling software to automate aspects of the process. We believe this meets an existing urgent need, with some modeling centers already exploiting this structured information to mitigate against the overwhelming complexity of experimental protocols. Moreover, we believe this will also enable and encourage unanticipated future use of the information in developing new software tools for exploiting it as technologies evolve. Our ability to predict (whether correctly or not remains to be seen) the expected CMIP6 data volume is one such unexpected outcome.

- The infrastructure allows user communities to assess the costs of participation as well as the benefits. For example, we believe the new PID-based methods of dataset tracking will allow centers to measure which data has value downstream. The importance of citations and fair credit for data providers is recognized, with a design that facilitates and encourages proper citation practices.

Certainly not all issues are resolved, and the validation of some of our findings will have to await the outcome of CMIP6. Nevertheless, we believe the discussion in this article provides a sound basis for beginning to think about the future.

- There is an increasing blurring of the boundary between weather and climate as time and space scales merge (Hoskins, 2013). This will increasingly entrain new communities into our data ecosystems, each with their own modeling and analysis practices, standards and conventions, and other issues. The establishment of the WIP was a crucial step in enhancing the capabilities, standards, protocols and policies around the CMIP enterprise. Earlier discussions on the scope of the WIP also suggested a broader scope for the panel on the longer-term, to coordinate not only the CMIP data aspects (including for example, the CORDEX project (Lake et al., 2017), which also relies upon ESGF for data dissemination, see Figure 1) but also the climate prediction (seasonal to decadal) issues and corresponding observational and reanalysis aspects.We would recommend a closer engagement between these communities in planning the future of global data infrastructure.

– As we have noted, the nature of publication is changing (see e.g David et al., 2016). In the future, datasets and software with provenance information will be first-class entities of scientific publication, alongside the traditional peer-reviewed article. In fact it is likely that those will increasingly feature in the grey literature and scientific social media: one can imagine blog posts and direct annotations on the published literature using analysis directly performed on datasets using their PIDs. Data analytics at large scale is increasingly moving toward machine learning and other directly data-driven methods of analysis, which will also be dependent on data with provenance tracking. We believe our community needs to pay increasing heed to the status of their data and software.

The WIP is well-positioned to extend its activities as these developments continue.

Appendix A: List of WIP position papers

– CDNOT Terms of Reference: a charter for the CMIP6 Data Node Operations Team. Authorship: WIP.

– CMIP6 Global Attributes, DRS, Filenames, Directory Structure, and CVs: conventions and controlled vocabularies for consistent naming of files and variables. Authorship: Karl E. Taylor, Martin Juckes, V. Balaji, Luca Cinquini, Sébastien Denvil, Paul J. Durack, Mark Elkington, Eric Guilyardi, Slava Kharin, Michael Lautenschlager, Bryan Lawrence, Denis Nadeau, and Martina Stockhause, and the WIP.

– CMIP6 Persistent Identifiers Implementation Plan: a system of identifying and citing datasets used in studies, at a fine grain. Authorship: Tobias Weigel, Michael Lautenschlager, Martin Juckes and the WIP.

– CMIP6 Replication and Versioning: a system for ensuring reliable and verifiable replication; tracking of dataset versions, retractions and errata. Authors: Stephan Kindermann, Sebastien Denvil and the WIP.

– CMIP6 Quality Assurance: systems for ensuring data compliance with rules and conventions listed above. Authorship: Frank Toussaint, Martina Stockhause, Michael Lautenschlager and the WIP.

– CMIP6 Data Citation and Long Term Archival: a system for generating Document Object Identifies (DOIs) to ensure long-term data curation. Authorship: Martina Stockhause, Frank Toussaint, Michael Lautenschlager, Bryan Lawrence and the WIP.

– CMIP6 Licensing and Access Control: terms of use and licenses to use data. Authorship: Bryan Lawrence and the WIP.

– CMIP6 ESGF Publication Requirements: linking WIP specifications to the ESGF software stack, conventions that software developers can build against. Authorship: Martin Juckes and the WIP.

– Errata System for CMIP6: a system for tracking and discovery of reported errata in the CMIP6 system. Authorship: Guillaume Levavasseur, Sébastien Denvil, Atef Ben Nasser, and the WIP.

– ESDOC Documentation: An overview of the process for providing structured documentation of the models, experiments and simulations that produce the CMIP6 output datasets, by the ES-DOC Team.

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2018-52
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 22 March 2018

Appendix B: Data and code availability

- The software and data used for the study of data compression are available at https://goo.gl/qkdDnn, courtesy Garrett Wright.
- The software and data used for the prediction of data volumes are available at https://goo.gl/Ezz5v3, courtesy Nalanda Sharadjaya.

Most of the software referenced here for which the WIP is providing design guidelines and requirements, but not implementation, including the ESGF, ESDOC, DREQ software stacks are open source and freely available. They are autonomous projects and therefore not listed here.

Acknowledgements. We thank Michel Rixen, Stephen Griffies, and John Krasting for their close reading and comments on early drafts of this manuscript. Colleen McHugh aided with the analysis of data volumes.

The research leading to these results has received funding from the European Union Seventh Framework program under the IS-ENES2 project (grant agreement No. 312979).

V. Balaji is supported by the Cooperative Institute for Climate Science, Princeton University, Award NA08OAR4320752 from the National Oceanic and Atmospheric Administration, U.S. Department of Commerce. The statements, findings, conclusions, and recommendations are those of the authors and do not necessarily reflect the views of Princeton University, the National Oceanic and Atmospheric Administration, or the U.S. Department of Commerce.

B.N. Lawrence acknowledges additional support from the UK Natural Environment Research Council.

K.E. Taylor and P.J. Durack are supported by the Regional and Global Model Analysis Program of the United States Department of Energy's Office of Science, and their work was performed under the auspices of Lawrence Livermore National Laboratory's Contract DE-AC52-07NA27344.

References

Baker, A. H., Hammerling, D. M., Mickelson, S. A., Xu, H., Stolpe, M. B., Naveau, P., Sanderson, B., Ebert-Uphoff, I., Samarasinghe, S., Simone, F. D., et al.: Evaluating lossy data compression on climate simulation data within a large ensemble, Geoscientific Model Development, 9, 4381–4403, 2016.

Balaji, V., Maisonnave, E., Zadeh, N., Lawrence, B. N., Biercamp, J., Fladrich, U., Aloisio, G., Benson, R., Caubel, A., Durachta, J., Foujols, M.-A., Lister, G., Mocavero, S., Underwood, S., and Wright, G.: CPMIP: measurements of real computational performance of Earth system models in CMIP6, Geoscientific Model Development, 10, 19–34, doi:10.5194/gmd-10-19-2017, http://www.geosci-model-dev.net/10/19/2017/, 2017.

Bony, S., Stevens, B., Held, I. H., Mitchell, J. F., Dufresne, J.-L., Emanuel, K. A., Friedlingstein, P., Griffies, S., and Senior, C.: Carbon dioxide and climate: perspectives on a scientific assessment, in: Climate Science for Serving Society, pp. 391–413, Springer, 2013.

Chard, K., Pruyne, J., Blaiszik, B., Ananthakrishnan, R., Tuecke, S., and Foster, I.: Globus data publication as a service: Lowering barriers to reproducible science, in: e-Science (e-Science), 2015 IEEE 11th International Conference on, pp. 401–410, IEEE, 2015.

Charney, J. G., Arakawa, A., Baker, D. J., Bolin, B., Dickinson, R. E., Goody, R. M., Leith, C. E., Stommel, H. M., and Wunsch, C. I.: Carbon dioxide and climate: a scientific assessment, 1979.

Collins, F. S. and Tabak, L. A.: NIH plans to enhance reproducibility, Nature, 505, 612, 2014.

David, C. H., Gil, Y., Duffy, C. J., Peckham, S. D., and Venayagamoorthy, S. K.: An introduction to the special issue on Geoscience Papers of the Future, Earth and Space Science, 3, 441–444, doi:10.1002/2016EA000201, http://dx.doi.org/10.1002/2016EA000201, 2016EA000201, 2016.

Durack, P. J., Taylor, K. E., Eyring, V., Ames, S. K., Hoang, T., Nadeau, D., Doutriaux, C., Stockhause, M., and Gleckler, P. J.: Input4MIPS: Making CMIP Model Forcing More Transparent, Eos Trans. AGU, submitted., 2017.

Edwards, P.: A vast machine: computer models, climate data, and the politics of global warming, The MIT Press, https://goo.gl/rMHdZk, 2010.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geoscientific Model Development, 9, 1937–1958, 2016a.

Eyring, V., Gleckler, P. J., Heinze, C., Stouffer, R. J., Taylor, K. E., Balaji, V., Guilyardi, E., Joussaume, S., Kindermann, S., Lawrence, B. N., Meehl, G. A., Righi, M., and Williams, D. N.: Towards improved and more routine Earth system model evaluation in CMIP, Earth System Dynamics, 7, 813–830, doi:10.5194/esd-7-813-2016, http://www.earth-syst-dynam.net/7/813/2016/, 2016b.

Gleckler, P., Doutriaux, C., Durack, P., Taylor, K., Zhang, Y., Williams, D., Mason, E., and Servonnat, J.: A more powerful reality test for climate models, Eos Trans. AGU, 97, 2016.

Griffies, S. M., Adcroft, A. J., Balaji, V., Danabasoglu, G., Durack, P. J., Gleckler, P. J., Gregory, J. M., Krasting, J. P., McDougall, T. J., Stouffer, R. J., et al.: Sampling the Physical Ocean in CMIP6 Simulations, CLIVAR Report, 2014.

Griffies, S. M., Danabasoglu, G., Durack, P. J., Adcroft, A. J., Balaji, V., Böning, C. W., Chassignet, E. P., Curchitser, E., Deshayes, J., Drange, H., et al.: OMIP contribution to CMIP6: experimental and diagnostic protocol for the physical component of the Ocean Model Intercomparison Project, Geoscientific Model Development, 9, 3231–3296, 2016.

Guilyardi, E., Balaji, V., Lawrence, B., Callaghan, S., Deluca, C., Denvil, S., Lautenschlager, M., Morgan, M., Murphy, S., and Taylor, K. E.: Documenting Climate Models and Their Simulations, Bull. Amer. Met. Soc., 94, 623–627, http://journals.ametsoc.org/doi/pdf/10.1175/BAMS-D-11-00035.1, 2013.

Haarsma, R. J., Roberts, M. J., Vidale, P. L., Senior, C. A., Bellucci, A., Bao, Q., Chang, P., Corti, S., Fučkar, N. S., Guemas, V., et al.: High Resolution Model Intercomparison Project (HighResMIP v1. 0) for CMIP6, Geoscientific Model Development, 9, 4185–4208, 2016.

Hoskins, B.: The potential for skill across the range of the seamless weather-climate prediction problem: a stimulus for our science, Quarterly Journal of the Royal Meteorological Society, 139, 573–584, 2013.

Juckes, M., Eyring, V., Taylor, K., Balaji, V., and Stouffer, R.: The CMIP6 Data Request: the next generation climate archive, in: EGU General Assembly Conference Abstracts, vol. 17, p. 13112, 2015.

Knutti, R.: The end of model democracy?, Climatic change, 102, 395–404, 2010.

Knutti, R., Sedláček, J., Sanderson, B. M., Lorenz, R., Fischer, E. M., and Eyring, V.: A climate model projection weighting scheme accounting for performance and interdependence, Geophysical Research Letters, 44, 1909–1918, 2017.

Lake, I., Gutowski, W., Giorgi, F., and Lee, B.: CORDEX: Climate Research and Information for Regions, Bulletin of the American Meteorological Society, 98, ES189–ES192, 2017.

Lawrence, B. N., Balaji, V., Bentley, P., Callaghan, S., DeLuca, C., Denvil, S., Devine, G., Elkington, M., Ford, R. W., Guilyardi, E., Lautenschlager, M., Morgan, M., Moine, M.-P., Murphy, S., Pascoe, C., Ramthun, H., Slavin, P., Steenman-Clark, L., Toussaint, F., Treshansky, A., and Valcke, S.: Describing Earth system simulations with the Metafor CIM, Geoscientific Model Development, 5, 1493–1500, https://www.geosci-model-dev.net/5/1493/2012/, 2012.

Moss, R. H., Edmonds, J. A., Hibbard, K. A., Manning, M. R., Rose, S. K., Van Vuuren, D. P., Carter, T. R., Emori, S., Kainuma, M., Kram, T., et al.: The next generation of scenarios for climate change research and assessment, Nature, 463, 747–756, 2010.

Overpeck, J., Meehl, G., Bony, S., and Easterling, D.: Climate data challenges in the 21st century, Science, 331, 700, http://science.sciencemag.org/content/331/6018/700, 2011.

Peng, R. D.: Reproducible Research in Computational Science, Science, 334, 1226–1227, doi:10.1126/science.1213847, 2011.

Stocker, T., Qin, D., Plattner, G., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, B., and Midgley, B.: IPCC, 2013: climate change 2013: the physical science basis. Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change, Cambridge University Press, 2013.

Stockhause, M. and Lautenschlager, M.: CMIP6 Data Citation of Evolving Data, Data Science Journal, 16, 2017.

Teixeira, J., Waliser, D., Ferraro, R., Gleckler, P., Lee, T., and Potter, G.: Satellite observations for CMIP5: The genesis of Obs4MIPs, Bulletin of the American Meteorological Society, 95, 1329–1334, 2014.

Williams, D. N., Balaji, V., Cinquini, L., Denvil, S., Duffy, D., Evans, B., Ferraro, R., Hansen, R., Lautenschlager, M., and Trenham, C.: A Global Repository for Planet-Sized Experiments and Observations, Bulletin of the American Meteorological Society, doi:10.1175/BAMS-D-15-00132.1, 2015.

Zhou, G., Weigel, T., and Plale, B.: Persistent Identifier Kernel Information for Machine Discovery, in: Joint Conference on Digital Libraries, 2018.

Ziv, J. and Lempel, A.: A universal algorithm for sequential data compression, IEEE Transactions on information theory, 23, 337–343, 1977.