Oversight and Governance of Standards for Model Intercomparison Projects and Related Activities

Prepared for the WGCM
Draft V2.2
March 12, 2014[1]

Background:

The increasingly integrated system for distributing output from Earth system models is also expected to serve observational data, reanalysis output, and so on. This evolving global data infrastructure has become critical, underpinning climate science and policy, and has been recognized as such internationally (e.g. via the US National Research Council[2] and the European Network for Earth System Simulation[3]).

Users of this infrastructure include not just the climate modeling community itself but also climate scientists from the interdisciplinary fields of climate change mitigation and adaptation. These users, from various scientific disciplines, set challenging requirements for climate data documentation, data structure and completeness, as well as the interfaces for data search, access and processing. Meeting the requirements of these users will not be trivial. It must be built on an underpinning governance body, suitable technology and operational services.

Software distributed by the Earth System Grid Federation (ESGF) provides the backbone for this infrastructure, which now with the latest phase of the Coupled Model Intercomparison Project (CMIP5) of the World Climate Research Programme (WCRP) has become a federated global archive. This software leverages and is supplemented by other infrastructure elements with functions such as viewing and comparing metadata, managing the user interface, and analyzing and visualizing data. This software system critically depends on standards that guarantee that users and different data distribution centers can discover, browse, catalog, obtain, analyze, and archive datasets from each other.

These standards range from basic Internet protocols through to community-specific data and metadata conventions. In particular, model intercomparison infrastructure depends on:

- constrained file formats, structures and metadata (netCDF conforming to both the general Climate-Forecast (CF) conventions, and often specific conventions such as

[1] Original proposal prepared by V. Balaji, Karl Taylor, Cecilia DeLuca, Eric Guilyardi, Martin Juckes, Michael Lautenschlager, Bryan Lawrence, and Dean Williams.

[2] "A National Strategy for Advancing Climate Modeling", NRC 2012

[3] "Infrastructure Strategy for the European Earth System Modelling Community, 2012-2022", European Network for Earth Simulation, 2012.

the CMIP5 protocols which can be satisfied using standardized software such as CMOR);

- URL and catalog standards such as OPeNDAP and THREDDS, making data accessible to remote locations;
- a search Application Programming Interface (API) allowing 3rd party software to query the archive catalog;
- data publication, node management and data harvesting protocols;
- a Data Reference Syntax (DRS) (supported by software such as DRS lib) allowing for creation of a uniform URL namespace for the data both within a project and, as far as possible, between projects;
- Standardized identification of data versions to enable users to determine whether downloaded data has be subsequently withdrawn or replaced and to support data replication across the federated archive.
- a Common Information Model (CIM) for the description of models and simulations, and
- a security protocol which gives users from participating identity providers transparent access to resources from all parts of the federated archive.

Together these conventions and standards can enable the high-level of automation necessary to deal with millions of files along with automatic rule-based data replication and persistence system.

The interface and data standards should be supported by software implementations that both minimize the risk of misinterpretation of the standards and provide some insulation from changes in the standards. In addition, formal agreements on rules of operation must be established in order to ensure the interchangeability of climate data entities between data nodes. Furthermore, a process should be established to ensure that as the science needs evolve, extensions and revisions of the current standards (some of which were tailored to the CMIP) do not unduly disrupt the existing infrastructure. Disruptions can be reduced through good architectural and software design and by providing clear advance warning and by striving to constrain extensions to fit within existing capabilities.

Challenges and need for oversight:

While the current conventions and standards have largely been established through grass-root and voluntary efforts, and these have in fact been essential to the successes of CMIP3 and CMIP5, nearly everyone involved is aware of many things that could have worked better. The system is not as scalable and automatic as it should be, and maintaining it is difficult. All the technologies are evolving, and there are new interdisciplinary pressures (such as the potential engagement of the Research Data Alliance). Besides the evolving technology issues, there is also pressure from the climate science community to expand the scope of ESGF; there is now a rapidly growing “cottage industry” promoting model intercomparison projects (MIPs). More and more communities of climate researchers are setting up specialized MIPs for studying specific problems, which sometimes may be of interest to only relatively small groups. Moreover, an infrastructure of the scope envisioned will be an absolute requirement as the demand for "climate services" increases.

With expanding recognition of the value of multi-model ensembles for the overall advancement of climate science, we believe it is imperative that we find ways to ensure that modeling centers can participate in MIPs without substantial extra logistical effort (beyond that with which they are already familiar). At the same time, these procedures should minimize the effort of the groups responsible for storing the data and maintaining the infrastructure. To those ends, a modicum of uniform standard practice has so far been maintained by the personal efforts of a few key players in the MIPs who have attempted to ensure that other MIPs broadly follow the precedents set by CMIP5. These personal efforts, even when supplemented with some publicly available software (such as CMOR, which facilitates both data writing and standards conformance), fall somewhat below the real requirements of all parties.

We believe that as MIPs proliferate, this informal approach will be difficult to "scale up" and will eventually fail. The immediate effect is likely to be felt by the modeling centers, which will be faced with meeting the diverse requests of multiple MIPs. Without enforcement of common standards, special procedures will be required for preparing data for each intercomparison activity, and this will ultimately overwhelm modeling group resources. The secondary effects will be felt by data users (often, but not always, the same people), who despite heroic efforts by the modeling groups, will inevitably be faced with heterogeneities in accessing the data and in the data structures themselves that will ultimately substantially impede scientific progress. Moreover without unified infrastructure approaches, it will be difficult to appropriately credit the data providers and systematically gauge the impact of modeling efforts.

Proposed Oversight and Guidance:

Motivated by the above considerations, we hereby propose that without undue delay the WGCM appoint a small panel (perhaps named the WGCM Infrastructure Panel (WIP)) tasked with establishing and maintaining standards and policies for model data sharing. They would be expected to endorse software implementations that support the standards. This group would serve as a counterpart to the CMIP Modeling Panel and would allow the modeling groups, through the WGCM, to maintain some internal control over the technical requirements imposed by the increasingly burdensome MIPs. The membership would also include representation of those responsible for the infrastructure underpinning the MIPS.

This new working group will create and maintain a document outlining the technologies necessary for operation of a global data infrastructure, along with the standards necessary for maintaining these technologies. The document will outline a protocol for creating and running a MIP. It will also identify gaps in the underpinning software needed to support the standards and be expected to identify the resources necessary to support the standards. The working group will also be tasked with drawing a broader community into a discussion of these standards, such as by hosting sessions at AGU/EGU and other meetings. Finally they will work towards ensuring proper credit for the providers of model output by helping establish standard ways of citing data.

It is anticipated that there will be more than modest time spent by members of the panel in carrying out its charge. Thus, it will be important that funding for their work be secured. The panel members might initially best be drawn from groups with ongoing support for climate modeling infrastructure work (e.g., PCMDI, the IS-ENES project, ES-DOC, NESII, and certain modeling groups), but also from unfunded, ongoing volunteer efforts (e.g., CF conventions leaders, GO-ESSP), and major data centers (e.g., BADC and DKRZ). Ideally, members will be familiar with what is practical in terms of software and what is needed as far as the typical climate researcher.

The development and documentation of standards is made much more effective when the standards are supported by software. This was one of the motivations for the development of CMOR. The work of the WIP will identify the requirement for new software. It is important that the standards supporting software development are properly managed and adequately funded.

It is the fervent hope of the authors of this document that with due speed the WGCM will establish an oversight panel along the lines suggested above to operate under the Terms of Reference found in the Appendix to this document.

Links:

BADC: http://badc.nerc.ac.uk/home/index.html
CIM: http://www.earthsystemcog.org/projects/es-doc-models/cim
CF Conventions: http://cf-pcmdi.llnl.gov/
CMIP5: http://cmip-pcmdi.llnl.gov/cmip5/
CMIP5 Model Output Requirements: http://cmip-pcmdi.llnl.gov/cmip5/docs/CMIP5_output_metadata_requirements.pdf
CMOR: http://www2-pcmdi.llnl.gov/cmor
DataCite: http://datacite.org
DKRZ: http://www.dkrz.de/?set_language=en
DRS: http://cmip-pcmdi.llnl.gov/cmip5/docs/cmip5_data_reference_syntax.pdf
ES-DOC: http://earthsystemcog.org/projects/es-doc-models/
ESGF: http://esgf.org/
GO-ESSP: http://go-essp.gfdl.noaa.gov/
IS-ENES: https://is.enes.org/
NESII: https://earthsystemcog.org/projects/nesii/
netCDF: http://www.unidata.ucar.edu/software/netcdf/
PCMDI: http://www-pcmdi.llnl.gov/
Research Data Alliance: https://rd-alliance.org/node
WCRP: http://www.wcrp-climate.org/

Appendix

Mission of the WGCM Infrastructure Panel (WIP)

The mission of the WGCM Infrastructure Panel (WIP) is to promote a robust and sustainable global data infrastructure in support of the scientific mission of the WGCM. Drawing on experts intimately familiar with the scientific goals of the WGCM and aware of the promises and limitations of infrastructural technologies, the WIP will formulate achievable goals for global data infrastructure, ensure coordination of the various groups building components of the system, and advise the relevant institutions on the requirements and commitments needed to maintain its long term vitality.

Terms of Reference for the WGCM Infrastructure Panel (WIP)

1. Serve the interests of the WGCM in establishing and maintaining standards and policies for sharing climate model output and derived products.
2. Encourage, when needed, proposals for extensions or modifications of established standards to meet new needs for sharing climate data. Review proposals and suggest modifications to achieve better consistency with existing standards and to minimize disruption of existing infrastructure. Endorse proposed changes that serve the interests of the WGCM.
3. Review for consistency with existing standards and infrastructure all specifications defined by model intercomparison projects and related efforts; endorse specifications that qualify.
4. Review and provide guidance on requirements of the infrastructure (e.g. level of service, accessibility, level of security);
5. Encourage development of and compose content for a website providing information on standards, policies, infrastructure, and controlled vocabularies endorsed by the WIP; provide clear guidance on infrastructure requirements for creating new community modeling efforts.
6. Collaborate with and rely on the ideas and leadership of other groups with interests in standards and infrastructure for climate data (e.g., CMIP, obs4MIPs, CORDEX, ESGF, ES-DOC, CF conventions), with the understanding that the WGCM expects the WIP to provide oversight.
7. Include in the scope of WIP oversight: a) file formats, structure and metadata, b) controlled vocabularies, name spaces, and naming conventions, c) protocols for interfacing components of the infrastructure, d) URL and catalog standards making data accessible regardless of local storage format, e) protocols for data publication (including version identification), node management and data harvesting, f) standardized descriptions of models and simulations, g) security protocol for authentication and authorization, and (h) query formats.
8. Report to the WGCM and seek their input as needed.