WGCM's global data infrastructure

WGCM Meeting
Victoria CANADA

V. Balaji

NOAA/GFDL and Princeton University

2 October 2013

Data consumers

Scientists perform sequences of computations (e.g *"poleward heat transport"*, *"length of growing season"*) on datasets. Typically this is scripted in some data analysis language, and ideally it should be possible to apply the script to diverse datasets.

Data producers

Observational and model output data in the climate-ocean-weather (COW) community is initially generated in some “native” non-standard format, and any subsequent relative analyses requires considerable effort to systematise. Issues include moving and transient data sources, lossy data formats, curvilinear and other “exotic” coordinates.

Data organizers

Data organizers are the community within this ecosystem that facilitates the transformation of source dependent data to a neutral and readily consumable form. They maintain the standards for describing data in a manner that permits these transformations, and develop tools to perform them.

The global data infrastructure underpinning MIPs

- MIPs, and in general any science involving cross-model comparisons, critically depend on the global data infrastructure – the “vast machine” (Edwards 2010) – making this sort of data-sharing possible.
- Infrastructure should not be a research project.
- Infrastructure should be treated as such by the national and international research agencies, but it is instead funded piecemeal, as a soft-money afterthought. This places the system at risk (NRC 2012: “A National Strategy for Advancing Climate Modeling”, ISENES-2 Infrastructure Strategy document, 2012.)

The Earth System Grid Federation

- The Earth System Grid Federation (ESGF; comprising large funded efforts at PCMDI, BADC, DKRZ, NCDC, and many modeling centers) designs, operates and maintains server software and hardware for the distribution of model data
- (and model-related observations including reanalysis).
- Software allows for archiving, browsing, cataloguing and discovering datasets.
- Services include search, download, replication, versioning, server-side analysis.
- Critically depends on standards!

Standards underpinning the GDI

- Data formats (netCDF) conforming to both the general Climate-Forecast (CF) conventions, and specific conventions such as the CMIP5 standards (satisfied using CMOR);
- URL and catalog standards such as OPeNDAP and THREDDS, making data accessible to remote locations regardless of local storage format;
- ESGF software: custom data publication, node management and data harvesting protocols developed by the ESG and the ESG Federation;
- the CMIP5 Data Reference Syntax (DRS) allowing for creation of a uniform URL namespace for CMIP5 data, and
- the Common Information Model (CIM) for the description of models and simulations. (Includes Gridspec, but not much used.)

Overseen by piecemeal volunteer efforts such as ESGF, GO-ESSP, CF Conventions and Variables Committees, ES-DOC, ... Standard extensions (e.g downscaling) may not have been adequately reviewed.

Data provenance and citation

- Datasets are quite often used without proper acknowledgement or record of provenance.
- The effort to issue Document Object Identifiers (DOIs) in CMIP5 was immature, and probably requires some extra steps.
- Quality control (QC-L2) was designed to provide several levels of data and metadata quality checking in CMIP5, but in addition proper peer review of metadata is needed.
- There are journals (e.g ESDD) that provide a mechanism for citable entities around data, and can be used as a vehicle for quality control, peer review, and credit.
- Record of data provenance likely to become a requirement at NOAA; many journals starting to require permanent record of datasets and methods.

Requirements for a robust and agile MIP infrastructure

We expect (and this workshop confirms) more specialized MIPs (and maybe ARs... see Nature editorial 18 September 2013, “The final assessment”). Current approach is not scalable!

- Recognition by funding agencies that the science critically depends on a GDI currently financed and operated on a risky ad-hoc basis.
- ESGF servers to be continually available and operated (new data will not appear at 6-yearly interval).
- Modeling centers will be unable to comply unless all MIPs to follow consistent standards established by a WGCM Infrastructure Panel. (COOKIE is a good example to follow...)
- WGCM Infrastructure Panel to act as data quality review body for new experiments.

Role of WGCM and its infrastructure panel

- Provide scientific guidance and requirements for the GDI; exert greater influence over its design and features.
- Provide standards governance allowing for orderly evolution of standards.
- Provide design templates (e.g CMOR extensions) for groups designing MIPs and work to ensure their conformance to standards.
- Work with academies and publishers to require adequate data citation and recognition for data providers.
- Intercede with national agencies to provision data infrastructure with adequate and stable long-term funding.

We expect this to be a non-trivial commitment of time and effort by Panel members.

Acknowledgements: Proposal initially prepared by V. Balaji and Karl Taylor, with input and revisions made by co-authors Eric Guilyardi, Michael Lautenschlager, Bryan Lawrence, and Dean Williams.