WGCM Infrastructure Panel Report

WGCM-19
Dubrovnik, CROATIA

V. Balaji and Karl Taylor

on behalf of the WGCM Infrastructure Panel

18 October 2015

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.)

Role of WGCM and its infrastructure panel (from 2013 meeting)

- 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.

WIP: The WGCM Infrastructure Panel formed 2014

- Chaired by V. Balaji (Princeton/GFDL) and K. Taylor (PCMDI).
- Strategy to develop a series of "position papers" on global data infrastructure and its interaction with the scientific design of experiments. These will be presented to WGCM annual meeting.
  - protocol document for the "endorsed MIPs" delivered. Working with CMIP panel and MIP sponsors on CMIP6 data request.
  - data access policies: would open access simplify the technical design of the infrastructure?
  - data citations. Developing and promoting a path to data citations using DOIs and the emerging data journals, such as ESSD, Nature Scientific Data.
  - projected data volumes for CMIP6, strategies for managing the growth path
- Close involvement of the WIP and CMIP panel (e.g. joint papers)
- Interest from other WCRP working groups (WGSIP, WGNE)
- ESGF *and* other tools: ESDOC, CMOR, CF Conventions, ..

Why not carry on as in the past?

- Heavy reliance on a few individuals worked O.K. for CMIP5, but may fail for the distributed management envisioned for CMIP6
- Need a procedure for evolving the infrastructure in a coordinated way so that the many groups and projects developing it can be responsive to the scientific needs.
- A panel with broad expertise may more nimbly respond to future needs than relying on a few individuals to poll community experts and build a consensus.
- Modeling groups are tasked with meeting the MIP requirements and deserve formal input to define them.
- Anything done to ensure that standards are as uniform as possible across all MIPs will reduce the burden.
- Membership on an official panel might help individual members to fund their work in this area.

WIP Mission

“to promote a robust and sustainable global data infrastructure in support of the scientific mission of the WGCM”

- Establish standards and policies for sharing climate model output
- ensure consistency across WGCM activities
- Extend standards as needed to meet evolving needs
- Review and provide guidance on requirements of the infrastructure (e.g. level of service, accessibility, level of security)
- Oversee
    - file formats, structure and metadata
    - controlled vocabularies, name spaces, and naming conventions
    - protocols for interfacing components of the infrastructure
    - URL and catalog standards
    - protocols for data publication (including version identification), node management and data harvesting
    - standardized descriptions of models and simulations
    - security protocol for authentication and authorization query formats.

Covers ESGF, DRS, CMOR, ESDOC, ...

WIP Membership

- V. Balaji (co-chair): GFDL
- Karl Taylor (co-chair): PCMDI
- Luca Cinquini: NASA JPL
- Cecelia DeLuca: NOAA
- Sébastien Denvil: IPSL
- Mark Elkington: MOHC
- Francesca Guglielmo, LSCE
- Eric Guilyardi: IPSL
- Martin Juckes: BADC
- Slava Kharin: CCCma
- Michael Lautenschlager: DKRZ
- Bryan Lawrence : NCAS, BADC
- Dean Williams: PCMDI

a blend of computer and climate scientists representing data centers and modeling groups: rotating membership with overlapping 2-year cycles

Position Paper: Formation of CDNOT

- WIP recommended to the WGCM and CMIP panel the formation of a technical consortium charged with operationalizing the CMIP6 ESGF Federation: the CMIP6 Data Node Operations Team (CDNOT).
- Distinct bodies (with overlapping membership) responsible for requirements (WIP), software development (ESGF, ESDOC, ...), and operations (CDNOT)
- Formation approved by WGCM and CMIP, June 2015.
- Sébastien Denvil appointed Chair of CDNOT.
- Many sites have proposed members: if you are planning to operate a CMIP6 data node, please contact Sébastien right away! CDNOT operations are imminent (as soon as ESGF 2.0 is released).

Position paper: CMIP6 Data Request

Led by Martin Juckes, STFC. See talk by Martin.
Highlights:

- Data request now available in machine-readable formats, including XLS and XML.
- A python API to allow the building of workflow tools that can work directly with the data request (e.g for setting switches in the model or post-processing).
- Endorsed MIPs have provided input on how the data will be used and analyzed.
- Actions needed from MIPs: develop and share analytic capabilities related to data request.
- Actions needed from modeling groups: review data request and provide feedback re feasibility.

Position paper: Data reference structure: syntax, vocabularies, filenames and global attributes

Highlights: mostly follows CMIP5 with some additional items:

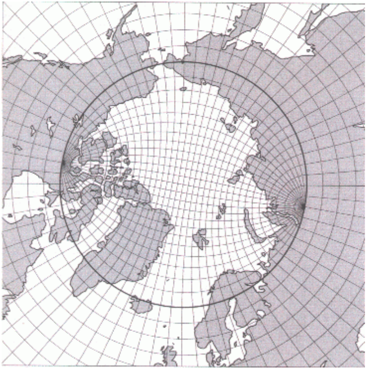
- Allows easier grouping and selection. For instance, runs distinguished only by forcings will now be seen as an ensemble (extension of **`rip`** to **`ripf`**).
- Notation for data regridded on standard grids (e.g 1x1, see below).
- Improved association of data across multiple files, e.g auxiliary **`cell_measures`** such as **`volcello`** (ocean cell volume).
  - **`associated_files`** now proposed for adoption in the CF convention as a general mechanism. We use this to point to a URL for tracking associations.
- DCPP extensions to allow additional forecast lead time coordinate.
- More sophisticated tracking of datasets (see discussion of PIDs below, **`tracking_id`** is now deprecated)
- When these papers are finalized and released, modeling groups can incorporate these into their workflows.

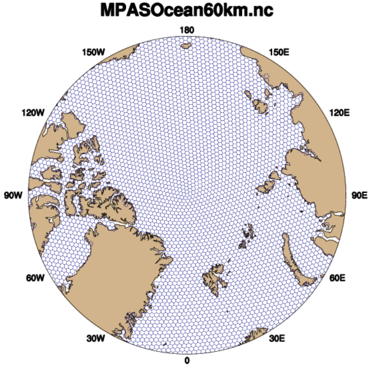
Other data and metadata standards

(Not in any WIP position paper as of now, but are WIP recommendations).

- The WIP recommends the use of netCDF4 with lossless compression as the data format for CMIP6.
  - Lossless compression from **`zlib`** (settings **`deflate=2`** and **`shuffle`**) expected to generate roughly 2X decrease in data volumes (varies depending on data entropy or noisiness).
  - Requires upgrading entire toolchain (data production and consumption) to netCDF4.
  - Potential loss in performance during data creation.
- The WIP recommends the use of standard grids for datasets where native-grid data is not strictly required. For example: the Clivar OMDP may request the use of WOA standard grids ($1^\circ \times 1^\circ$, $0.25^\circ \times 0.25^\circ$) as the target grid of choice.
- No progress on adoption of standard calendars.

Grid diversity may increase in CMIP6





Downstream communities may not wish to deal with novel grids, but specialist communities are likely to insist on it for their own research.

Model metadata

(Not in any WIP position paper as of now, but are WIP recommendations).

- ESDOC documents of model metadata are a required element in quality control and DOI generation.
- Considerably simplified questionnaire relative to CMIP5.
- Command-line tools (e.g `py-esdoc`) will be made available to make it easy to generate, clone, share CIM documents.
- Forcing documentation in the works.
- Should we include tuning documentation? cf. Hourdin et al BAMS paper, “The Art and Science of Model Tuning”, in final stages of preparation. Outcome of 2014 tuning workshop.

Position papers: Replication, versioning and errata

Main requirement is for end users to know if they are working with the right dataset in a federation where data is replicated multiple times, may have been retracted or superseded. Highlights:

- Extend use of persistent identifiers (PIDs) for dataset tracking (replaces **`tracking_id`** from CMIP5).
- Lists of PIDs can be used as supplementary citation information in papers.
- PID-based query system to see if errata have been reported, or data have been superseded.
- Proposal to ESGF working teams (see Dean's talk) on how PIDs can be incorporated into replication workflow.

Position paper: Data citation and long-term access

Highlights:

- Main requirement: ensure proper citation of data used in a study to acknowledge contributions by modeling groups.
- Automated QC mechanisms to ensure adherence to metadata and data quality standards.
- Commitment to long-term archival by at least some data centers.
- Links connecting datasets to model and experiment documentation (ESDOC/CIM)
- DOI generation at the granularity of *model* and *simulation*.
- Action needed from CMIP Panel: endorse the requirement of data citation as part of the terms of use of CMIP6 model output.
- Recommendation to modeling groups: generate citations in the emerging data science journals e.g., Nature Scientific Data or ESSD. Possibly approach for special issue?

Position paper: Data licensing and access control

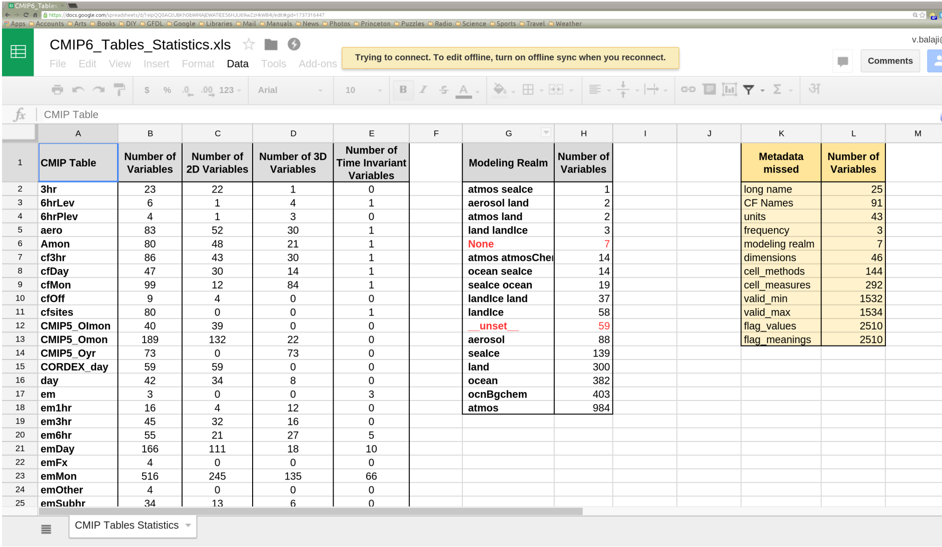
- Main requirement: simplified access control on ESGF, data license applicable even when data is found in non-ESGF repositories.
- For CMIP6 data licenses will be embedded in the data files (netCDF global attribute). There will be choice of two different licenses (Creative Commons "share-alike" and "non-commercial share-alike")
- Recognition that many users will (and did) use data from secondary ("dark") repositories. Embedded license implies that user is subject to the terms of use no matter where they retrieved the data.
- Required action from CMIP Panel: endorse the new WIP-recommended licensing policy.
- Required action from modeling groups: choose a license consistent with your own institutional policies and record in global file attributes. Let us know if the two recommended licenses are both unacceptable.

Position paper: CMIP6 Data Volume

CDNOT member institutions and ESGF require realistic data volume estimates for hardware planning.

- A number of current estimates are based on an assumption of geometric progression (straight line on a log scale!) drawn through CMIP3 and CMIP5.
- Based on known growth in number of models, years simulated, and increase in resolution, the actual growth will likely be less.
- Some centres (e.g UKMO and GFDL) are developing tools to allow us (and possibly others) to make accurate data volume estimates based on Martin's data request documents, model resolution, experiment planning.
- WIP will release in early 2016 best estimates based on information acquired at this workshop.

CMIP6 Data Request: preliminary analysis



CMIP6_Tables_Statistics.xls

Trying to connect. To edit offline, turn on offline sync when you reconnect.

CMIP Table

CMIP Table	Number of 2D Variables	Number of 3D Variables	Number of Time Invariant Variables	
3hr	23	22	1	0
6hrLev	6	1	4	1
6hrPlev	4	1	3	0
aero	83	52	30	1
Amon	80	48	21	1
cf3hr	86	43	30	1
cfDay	47	30	14	1
cfMon	99	12	84	1
cfOff	9	4	0	0
cfsites	80	0	0	1
CMIP5_OImon	40	39	0	0
CMIP5_Omon	189	132	22	0
CMIP5_Oyr	73	0	73	0
CORDEX_day	59	59	0	0
day	42	34	8	0
em	3	0	0	3
em1hr	16	4	12	0
em3hr	45	32	16	0
em6hr	55	21	27	5
emDay	166	111	18	10
emFx	4	0	0	0
emMon	516	245	135	66
emOther	4	0	0	0
emSubhr	34	13	6	0

Modeling Realm	Number of Variables
atmos sealce	1
aerosol land	2
atmos land	2
land landIce	3
None	7
atmos atmosChei	14
ocean sealce	14
sealce ocean	19
landIce land	37
landIce	58
__unset__	59
aerosol	88
sealce	139
land	300
ocean	382
ocnBgchem	403
atmos	984

Metadata missed	Number of Variables
long name	25
CF Names	91
units	43
frequency	3
modeling realm	7
dimensions	46
cell_methods	144
cell_measures	292
valid_min	1532
valid_max	1534
flag_values	2510
flag_meanings	2510

CMIP Tables Statistics

WIP Position Papers: Current Status

- Recommended formation of CMIP6 Data Node Operations Team (CDNOT: Sébastien Denvil, Chair).
- Recommended use of netCDF4 lossless compression for CMIP6
- Data Citation and Long Term Access: DOIs issued for quality-controlled data at the granularity of model and simulation.
- Recommended use of Persistent Identifier (PID) at the dataset level. Allows tracking for datasets for replication, versioning, and errata.
- Simplified licensing and data access: licenses embedded in files (two options: open access and non-commercial use)
- Recommended use of standard grids (e.g 1x1) of limited set of high-value data.
- Standard format machine-readable data request for DECK and MIPs.
- Finalizing Data Reference Structure and Syntax (paths and controlled vocabularies) and netCDF attributes.
- Data volume estimates to be released after data request finalized.

Conclusions

- WGCM Infrastructure Panel translates CMIP experimental design into requirements for the global data infrastructure
- Governance at different stages of infrastructure: requirements (WIP), software development (ESGF, ESDOC, CMOR, ...), CMIP6 implementation and operations (CDNOT).
- Close involvement of WIP with ESGF-XC and CDNOT (overlapping membership)
- WIP has produced 7 position papers (out of the promised 4), available on the WIP website. Data volume estimate paper soon to follow.

`https://www.earthsystemcog.org/project/wip/resources/`