Status of Infrastructure for CMIP6

WGCM-20
Lewis Library
Princeton University

V. Balaji, on behalf of the WIP

NOAA/GFDL and Princeton University

1 November 2016

Outline

1. WIP Documents on Infrastructure Needs
2. input4MIPs
3. ESGF
4. ESDOC
5. Data Volume Estimates
6. Items for WGCM Consideration

WIP Position Papers

https://earthsystemcog.org/projects/wip/

- CDNOT Terms of Reference
- CMIP6 Licensing and Access Control
- CMIP6 Persistent Identifiers Implementation Plan
- CMIP6 Replication and Versioning
- CMIP6 Data Citation and Long Term Archival
- CMIP6 Quality Assurance
- CMIP6 ESGF Publication Requirements
- CMIP6 Global Attributes, DRS, Filenames, Directory Structure, and CVs

... a few others not in final status (data volume, errata, ...)

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to know if they are working with the right dataset in a federation where data is replicated multiple times, may have been retracted or superseded.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 (PCMDI repository of published papers to collect these lists)
- PID-based query system to see if errata have been reported, or data have been superseded.
- Proposal to ESGF working teams 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 WGCM: 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

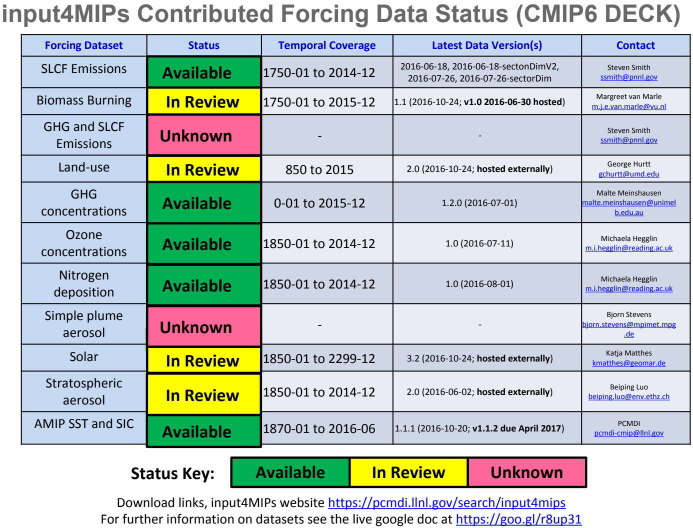
- Main requirement: simplified access control on ESGF, data license applicable even when data is found in non-ESGF repositories.
- For CMIP6 data licenses will be embedded in the data files (netCDF global attribute). There will be choice of two different licenses (Creative Commons “share-alike” and “non-commercial 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 WGCM: 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.

Outline

1. WIP Documents on Infrastructure Needs
2. input4MIPs
3. ESGF
4. ESDOC
5. Data Volume Estimates
6. Items for WGCM Consideration

input4MIPs status

https://pcmdi.llnl.gov/search/input4MIPs/



input4MIPs Contributed Forcing Data Status (CMIP6 DECK)

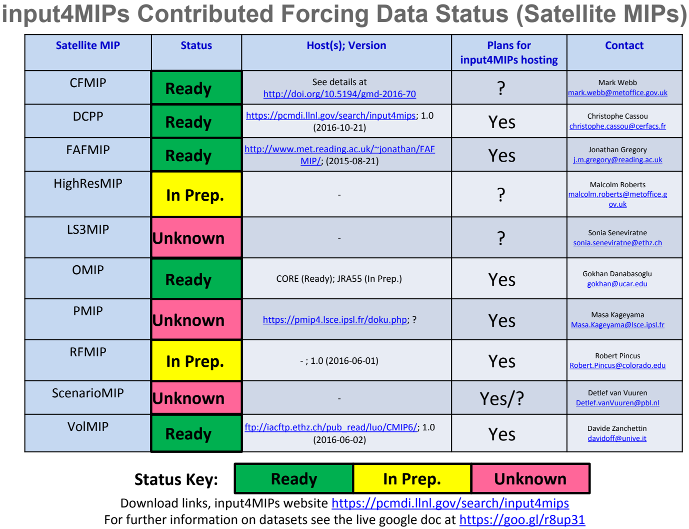
Forcing Dataset	Status	Temporal Coverage	Latest Data Version(s)	Contact
SLCF Emissions	Available	1750-01 to 2014-12	2016-06-18, 2016-06-18-sectonDimV2, 2016-07-26, 2016-07-26-sectorDim	Steven Smith ssmith@pnnl.gov
Biomass Burning	In Review	1750-01 to 2015-12	1.1 (2016-10-24; **v1.0 2016-06-30 hosted**)	Margreet van Marle m.j.e.van.marle@vu.nl
GHG and SLCF Emissions	Unknown	-	-	Steven Smith ssmith@pnnl.gov
Land-use	In Review	850 to 2015	2.0 (2016-10-24; **hosted externally**)	George Hurtt gchurtt@umd.edu
GHG concentrations	Available	0-01 to 2015-12	1.2.0 (2016-07-01)	Malte Meinshausen malte.meinshausen@unimelb.edu.au
Ozone concentrations	Available	1850-01 to 2014-12	1.0 (2016-07-11)	Michaela Hegglin m.i.hegglin@reading.ac.uk
Nitrogen deposition	Available	1850-01 to 2014-12	1.0 (2016-08-01)	Michaela Hegglin m.i.hegglin@reading.ac.uk
Simple plume aerosol	Unknown	-	-	Bjorn Stevens bjorn.stevens@mpimet.mpg.de
Solar	In Review	1850-01 to 2299-12	3.2 (2016-10-24; **hosted externally**)	Katja Matthes kmatthes@geomar.de
Stratospheric aerosol	In Review	1850-01 to 2014-12	2.0 (2016-06-02; **hosted externally**)	Beiping Luo beiping.luo@env.ethz.ch
AMIP SST and SIC	Available	1870-01 to 2016-06	1.1.1 (2016-10-20; **v1.1.2 due April 2017**)	PCMDI pcmdi-cmip@llnl.gov

Status Key: Available | In Review | Unknown

Download links, input4MIPs website https://pcmdi.llnl.gov/search/input4mips
For further information on datasets see the live google doc at https://goo.gl/r8up31

input4MIPs status

https://pcmdi.llnl.gov/search/input4MIPs/



input4MIPs Contributed Forcing Data Status (Satellite MIPs)

Satellite MIP	Status	Host(s); Version	Plans for input4MIPs hosting	Contact
CFMIP	Ready	See details at http://doi.org/10.5194/gmd-2016-70	?	Mark Webb mark.webb@metoffice.gov.uk
DCPP	Ready	https://pcmdi.llnl.gov/search/input4mips; 1.0 (2016-10-21)	Yes	Christophe Cassou christophe.cassou@cerfacs.fr
FAFMIP	Ready	http://www.met.reading.ac.uk/~jonathan/FAFMIP/; (2015-08-21)	Yes	Jonathan Gregory j.m.gregory@reading.ac.uk
HighResMIP	In Prep.	-	?	Malcolm Roberts malcolm.roberts@metoffice.gov.uk
LS3MIP	Unknown	-	?	Sonia Seneviratne sonia.seneviratne@ethz.ch
OMIP	Ready	CORE (Ready); JRA55 (In Prep.)	Yes	Gokhan Danabasoglu gokhan@ucar.edu
PMIP	Unknown	https://pmip4.lsce.ipsl.fr/doku.php; ?	Yes	Masa Kageyama Masa.Kageyama@lsce.ipsl.fr
RFMIP	In Prep.	- ; 1.0 (2016-06-01)	Yes	Robert Pincus Robert.Pincus@colorado.edu
ScenarioMIP	Unknown	-	Yes/?	Detlef van Vuuren Detlef.vanVuuren@pbl.nl
VolMIP	Ready	ftp://iacftp.ethz.ch/pub_read/luo/CMIP6/; 1.0 (2016-06-02)	Yes	Davide Zanchettin davidoff@unive.it

Status Key: Ready | In Prep. | Unknown

Download links, input4MIPs website https://pcmdi.llnl.gov/search/input4mips

For further information on datasets see the live google doc at https://goo.gl/r8up31

Outline

1 WIP Documents on Infrastructure Needs

2 input4MIPs

3 ESGF

4 ESDOC

5 Data Volume Estimates

6 Items for WGCM Consideration

ESGF Readiness for CMIP6

Courtesy Dean Williams, LLNL.

- ESGF Governance Policy **https://goo.gl/7KnUhh**
- Logo Requirement and Usage Guidelines **https://goo.gl/3jVx9n**
- ESGF Strategic Roadmap **https://goo.gl/D2Rw0z**
- Software Security Plan **https://goo.gl/mFBfGt**
- ESGF Federation Policies and Guidelines **https://goo.gl/dY339I**
- Root Certificate Authorities Policy DRAFT: **https://goo.gl/kPXfFU**
- ESGF Tier 1 and Tier 2 Node Requirements (under development);
- Data Storage and Replication Plan (under development);
- User Training Plan (under development); and
- ESGF CMIP6 Readiness Document (under development).

Williams et al 2016, BAMS: **https://goo.gl/AYU7Kj**

Outline

1. WIP Documents on Infrastructure Needs
2. input4MIPs
3. ESGF
4. ESDOC
5. Data Volume Estimates
6. Items for WGCM Consideration

ESDOC status for CMIP6

- Project to document CMIP6 well underway
- Building on CMIP5 experience (both good and bad!)
- Clear set of use cases
- Community review formalised (internal, WIP/WGCM, wider)
- Designed so that process less painful for groups:
  - Large fraction is automated
  - Starting model description from CMIP5 version
  - Beta testing of 5 months (Oct 2016 – Feb 2017, UKMO, GFDL, IPSL)
    - Would like to bring in 2 new groups (suggestions?)
  - Documentation for all steps (+ overview as WIP white paper)
- Full community release: March 2017
- Looking ahead (posts CMIP6) to include other “realms”:
  - Regional models, downscaling
  - Evaluation and metrics, obs4MIPs

ESDOC status for CMIP6

- About half of the documents (experiment, simulation, ensemble,...) automated (following ESGF publishing)
- Others (model, conformance to protocol, forcings, responsible party,...) produced by groups when ready – linked via the **`further_info_URL`** attribute
- Multiple tools to create these documents (python library or notebooks, questionnaire, ...) currently in internal review.
- Realm descriptions: Ocean, atmosphere, sea-ice further along, others still to come. Update science contents of other realms (with the community/WGCM) – Feb 2017
- Working on:
  - Forcings description (with Tim Johns et al. e.g. IPCC Table 12.1) – timeline: Nov 2016
  - Short model tables for papers (draft for ocean available) – Jan 2017

Outline

1. WIP Documents on Infrastructure Needs
2. input4MIPs
3. ESGF
4. ESDOC
5. Data Volume Estimates
6. Items for WGCM Consideration

Data Volume Estimate

- Lower bound on the total volume estimate is 18.3 PB. Updated figures shortly.
- 50% compression from the use of netCDF4 lossless compression.
- Bulk of volume comes from high-frequency (sub-monthly) data.
- Use of standard grids (ERA-Interim grid for atmosphere, 2.5°; WOA grid for ocean, 1°, 33 levels) can reduce the volume to about 2 PB.
- Limited enthusiasm for standard grids among modeling centers, excepting OMDP recommendations.
- No acceptable common calendar among modeling groups.
- Proposal to supply "interpolation weights" to standard grid for regridding post-facto.
- ESGF-XC unwilling to offer regridding service.
- ESGF-XC does not believe 20 PB can be successfully replicated.
- Status quo: "snapshot" repositories will cull subset, onus of regridding and moving to common calendar on end-user.

Outline

1. WIP Documents on Infrastructure Needs
2. input4MIPs
3. ESGF
4. ESDOC
5. Data Volume Estimates
6. Items for WGCM Consideration

Items for WGCM Consideration

- Insist on WIP-recommended data citation methods. The citation requirement is encoded in the “terms of use”.
- The **`drq`** is the single authoritative source for the data request: all revisions and corrections to begin there. Other tools (e.g CMOR3, modeling center workflows e.g at UKMO and GFDL, ...) begin there.
- Data volume estimates: 20 PB.
- Dataset standardization: my personal opinion is that modeling centers are not deeply invested in making their datasets widely used by non-specialists. Can