Report from the WGCM Infrastructure Panel (WIP)

V. Balaji (GFDL/Princeton)
and
Karl E. Taylor (PCMDI)

18th Session of the WGCM
Grainau, Germany
6-8 October 2014

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 and 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.

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.

Outline

- WIP overview
- Strategy and progress
- White papers in development
- Tomorrow:
  - Regrid some output fields in CMIP6?
  - Adopt common calendar in CMIP6?
  - CMIP data request (timeline, responsibilities)
  - ESGF status and plans
  - ES-doc model documentation status and plans

WIP progress

- Established following 17th Session of WGCM
- March 2014: Terms of Reference written
- May 2014: Members invited
- June 2014: Plan presented to the WCRP and endorsed
- Panel has met via telecon four times
- Web site established:

  http://cog-esgf.esrl.noaa.gov/projects/wip/
- 4 white papers are under preparation

WIP members: a blend of computer and climate scientists representing data centers and modeling groups

V. Balaji (co-chair): GFDL

Karl Taylor (co-chair): PCMDI

Luca Cinquini: NASA JPL

Cecelia DeLuca: NOAA

Sebastien Denvil: IPSL

Mark Elkington: MOHC

Eric Guilyardi: IPSL

Martin Juckes: BADC

Slava Kharin: CCCma

Michael Lautenschlager: DKRZ

Bryan Lawrence : NCAS, BADC

Dean Williams: PCMDI

Activities that WIP will help keep coordinated

- Major activities:
  - ESGF (data archive and delivery)
  - COG (Web interface to MIPs and MIP data)
  - ES-DOC (Model and experiment documentation)
  - CMOR (code to rewrite model output)
- Other activities:
  - Liaising with the CF conventions
  - Data reference syntax (DRS)
  - Quality assurance software

Initial strategy: Develop a series of "position papers" on data infrastructure in support of CMIP activities

- Protocol document for the "endorsed MIPs".
- 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.
- Strategies for managing the growth of CMIP data volumes

White paper: Endorsed MIP protocols

This document outlines the data and metadata protocols the MIP managers will be required to define and enforce, so that there is

- Consistency across all MIPs and DECK.
  - The DECK will be a refined version of what was done in CMIP5
- Minimal extensions and additions to the DECK model output request and data requirements except as needed
  - To answer specific scientific questions (e.g., new variables & vocabularies)
  - To accommodate new types of data (e.g., two time coordinates for near-term prediction: forecast time and forecast lead time)

MIP checklist: A list of actions, issues and bottlenecks for MIP coordinators

Scientific issues (CMIP panel):

- Initialization, experiment description, forcing data, justification of variable request

Infrastructure issues: (WIP and service providers/governance bodies)

- ESGF coordinating host, ESGF data node(s), model documentation plan, volume estimate, standard names, ESGF extensions [if required], quality control procedure:

Vocabularies and technical specification (WIP)

- Data reference syntax, institutions and models, other vocabularies

White paper: CMIP licensing and access control

For CMIP6 the WIP proposes a change in the how modeling centers specify terms of use.

- In CMIP5: Users signed a terms of use agreement when they registered and then were given access only to files falling under that agreement
  - The complicated ESGF access control mechanisms impaired smooth and easy downloading.
- For CMIP6 data licenses will be embedded in the data files (netCDF global attribute)
  - There will be choice of two different licenses ("unrestricted" and "non-commercial research") Required registration for updates (in the event of retraction or republication)
  - This will enable direct access to data without sign-in
  - If secondary ("dark") repositories are established, the data will continue to be served under license.
  - Users can register for updates (to learn of retraction or republication)

White paper: Data citation

The WIP proposes to encourage accurate identification of data used in research

- Provide credit and attribution (for data creators and contributors)
  - Enable direct citation in publications
- Uniquely identify data used in research
  - Provide services for recording and retrieving provenance information
  - Provide services for retrieving data
  - Services need to be compatible with other provenance mechanisms
- DOI assigned to the ensemble of runs produced by a single modeling group for a single experiment.

White paper: Proposed data citation requirements for CMIP6

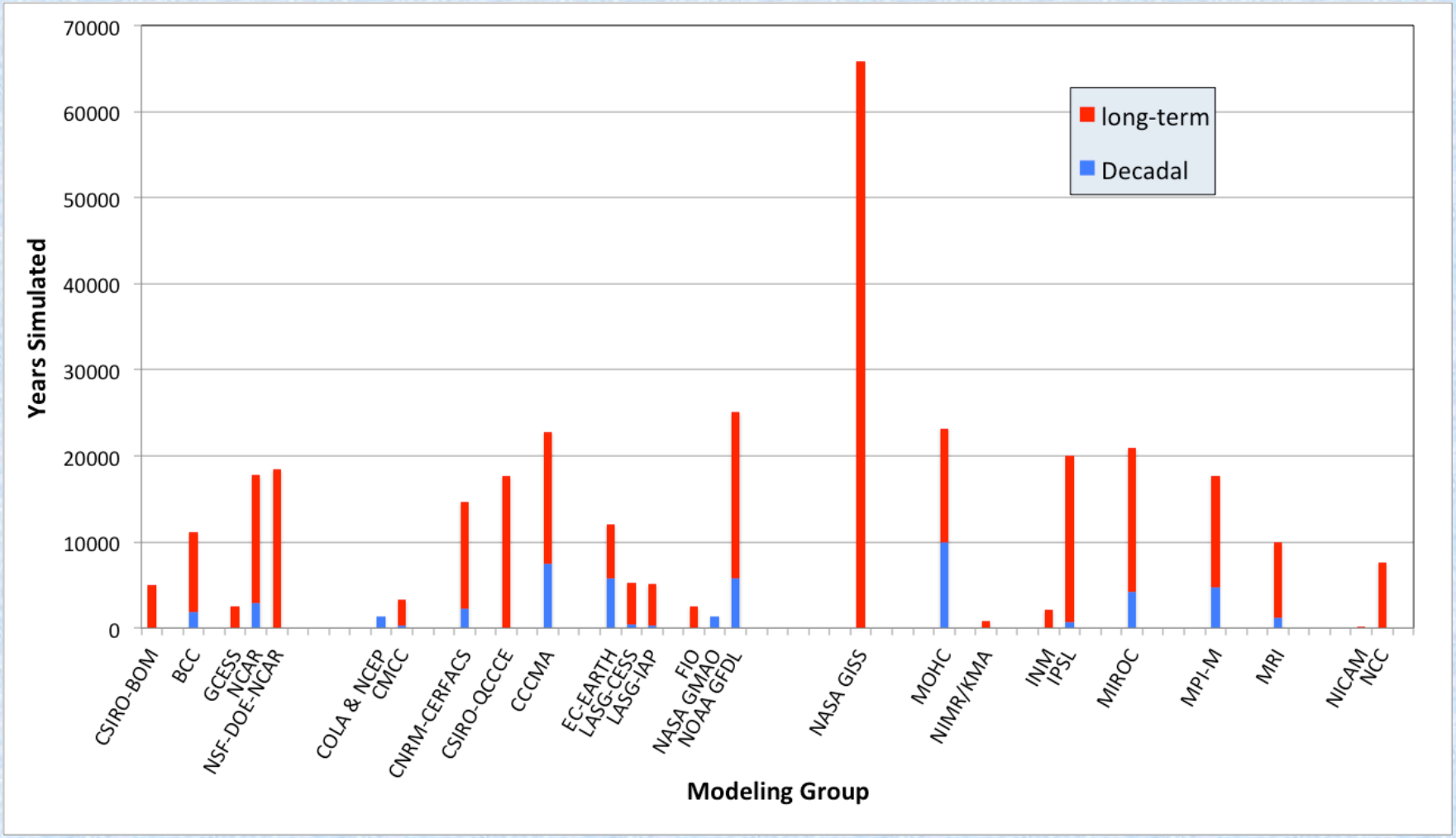
- A WGCM-endorsed policy **requiring** proper citation of datasets in publications
- A recommendations to modeling groups to generate citations in the emerging data science journals
  - e.g., Nature Scientific Data or ESSD
  - Possibly approach one of the journals for a CMIP6 special issue.
- Enhancement of quality control by the modeling groups.
- Demands on the infrastructure:
  - Automated QC mechanisms to ensure adherence to metadata and data quality standards.
  - Automated methods to generate persistent identifiers (PIDs) to collections of files.
  - Commitment to long-term archival by at least some data centers
  - Links connecting datasets to model and experiment documentation (ESDOC/CIM)

White paper: Projected data volumes for CMIP6

Historical data rates:

- CMIP3: 17 institutes(groups) and 25 models (40 TB)
  - total years simulated: 70000
  - individual models simulated 500 to 8400 years with a median of 2200 and a mean of 2800
  - individual groups simulated on average 70000/17 = 4100 years
- CMIP5: 26 institutes (groups) and 60 models (2 PB)
  - numbers estimated on 10/1/2014 (to within about 20%, I guess)
  - total years simulated: 330000
  - individual models simulated on average 330000/60 = 5500 years
  - individual groups simulated on average 330000/26 = 13000 years

Years simulated by each modeling group for CMIP5



Projected data volumes for CMIP6

- The WIP submitted questions to the modeling groups in order to anticipate data volumes for CMIP6:
  - a) What is the expected resolution(s) of your CMIP6 model(s)? (Atmosphere? Ocean?)
  - b) Do you aim to run different configurations (e.g. ESM, physical, etc) of your model(s)?
  - c) Based on your estimate of how much computing you expect to be available to you for CMIP6, how many total model years of these model(s) do you think you can run?
- Model resolution likely to increase in CMIP6
- Modeling groups say they will be able to simulate about the same number of years in CMIP6 as in CMIP5

Grid questions: Should ocean data be regridded?

- Many users and common software packages are unable to analyze data on some native grids (e.g., rotated pole grids).
- Regridding by end-users is problematic (e.g conservation, treatment of curvilinear coordinates)
- Should we define a single common grid (e.g., 1x1), and provide data on this grid?
  - Possibly limit regridding to most-commonly analyzed variables (e.g surface temperature and precipitation)
- Other possible technical solutions (all risky!):
  - Can ESGF provide server-side regridding services?
  - Can modeling centres provide regridding software?
  - Can modeling centres provide the interpolation weight tables between their native grid and a common set of targeted grids?

Grids: ocean modelers' perspective

- survey of users, this was principally due to the difficulties of analyzing data on the model's native horizontal grid (e.g tripolar).
- Grid considerations:
- Some operations on regridded data yield incorrect results (e.g products)
- Naive regridding methods are non-conservative (problem compounded on ocean and land surfaces by the presence of coastlines)
- Ocean modeling groups participating in CMIP6 have written a detailed document outlining requirements for ocean data (Griffies et al 2014). Key point re grids:
- A subset of data variables, most widely of interest, are being requested on a spherical longitude-latitude grid and predefined model levels. Modeling centres must take responsibility for proper conservative regridding of these fields

Calendar question: Should we recommend all models adopt a common calendar?

- CMIP5 data has numerous errors related to time axis information
- Calendar matching a common nuisance for analysts.
- Suggest, e.g., JULIAN for near-term experiments?
- Suggest, e.g., NOLEAP for decadal prediction experiments?

ESGF: Status and plans

- Leading international agencies are working toward an agreement on an ESGF governance document.
- ESGF has formed working teams to address technical issues for federation and node architectures and admin:
  - IWT, Installation Working Team
  - PWT, Publication Working Team
  - CWT, Compute Working Team.
  - NWT, Network Working Team
- ESGF major upgrades in progress:
  - A new search and access interface to CMIP archives (COG)
  - Enhanced server-side processing

ES-Doc: Model & simulation documentation

- CMIP5 content review (QC until end 2014)
- Viewer & comparator now operational (see http://es-doc.org)
  - *[show screen shots of viewer and comparator from es-doc.org page]*
- CMIP6 planning:
  - Simplification of contents vs. CMIP5 ?
  - Avoiding redundancy in NetCDF/ES-DOC/DRS content? (WIP to advise)
  - Quality Control Information (of the simulations and of the descriptions)
  - Better tooling?
    - Tools to create metadata from the command line (or from your own information repository, pyesdoc)
    - More than one "CIM questionnaire"?
  - Timeline: define during ES-DOC PIs telco Nov 7th 2014
- Upgrade to CIM itself and tooling

ES-Doc display tool (CMIP5)

es-doc
Earth System Documentation

Documentation Search v0.9.0.3 Support

Doc Type :	Doc Version :	Project :	Institute :	Model :	Experiment :
Model	Latest	CMIP5	*	*	*

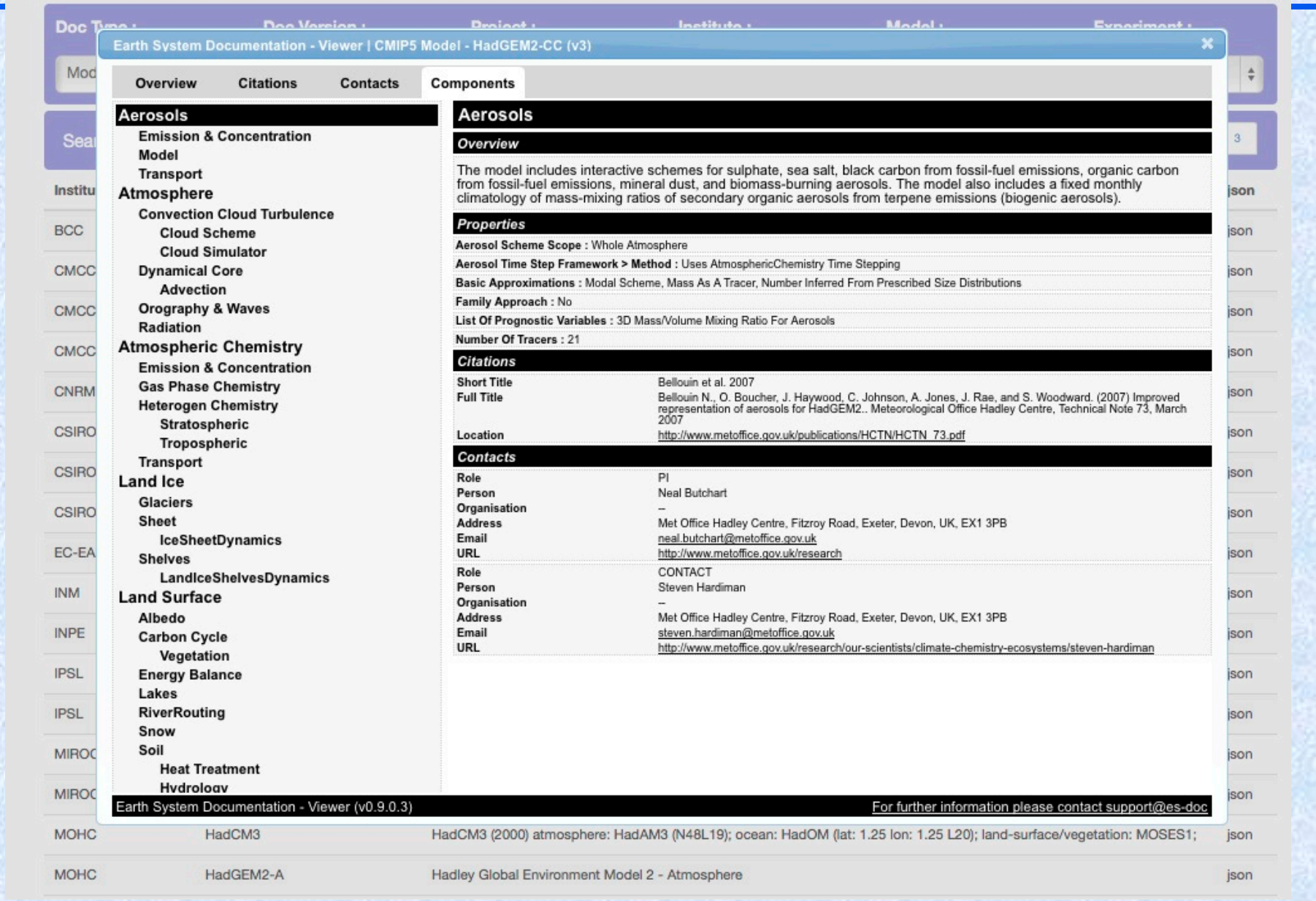
Search returned 42 of 107 records in 0.665s

1 2 3

Institute	Short Name	Long Name	json
BCC	BCC-CSM1.1	Beijing Climate Center Climate System Model version 1.1	json
CMCC	CMCC-CESM	CMCC Carbon Earth System Model	json
CMCC	CMCC-CM	CMCC Climate Model	json
CMCC	CMCC-CMS	CMCC Climate Model with a resolved Stratosphere	json
CNRM-CERFACS	CNRM-CM5	CNRM-CM5	json
CSIRO-BOM	ACCESS1.0	ACCESS1.0	json
CSIRO-BOM	ACCESS1.3	ACCESS1.3	json
CSIRO-QCCCE	CSIRO-Mk3.6.0	CSIRO Mark 3.6.0	json
EC-EARTH	EC-EARTH	EC-EARTH	json
INM	INM-CM4	inmcm4	json

https://www.earthsystemcog.org/projects/es-doc-models/

ES-Doc display sample (aerosols)



ES-Doc comparator tool (CMIP5)