CMIP6 Design and Plans

&

WGCM Infrastructure Panel (WIP) Requirements

Karl E. Taylor

(on behalf of the CMIP Panel and the WIP)

Presented at the
4th Annual ESGF & UV-CDAT Meeting

Livermore, CA
3 December 2013

Outline

- New framework for CMIP6

- Infrastructure implications for ESGF

- What is the role of the WGCM Infrastructure Panel (WIP)?

- What specifically is needed in the next few years?

K. E. Taylor
PCMDI

The WGCM and modeling groups are planning a more flexible structure for coordinated modeling activities

- CMIP5

  ➡ Monolithic

  ➡ Resource-intensive

- Future coordinated model activities (CMIP & CMIP6)

  ➡ Basic, routinely performed limited set of experiments (CMIP)

  ➡ Specialized additional experiments focusing on specific science questions (CMIP6): Modeling groups pick and choose.

- Fundamental requirement set by WGCM:

  ➡ All activities make use of common infrastructure for archiving and accessing data

  ➡ Expectation that ESGF can evolve to meet all the needs.

WCRP Grand Challenges: (1) Clouds, circulation and climate sensitivity, (2) Changes in cryosphere, (3) Climate extremes, (4) Regional climate information, (5) Regional sea-level rise, and (6) Water availability, plus an additional theme on "Biogeochemical forcings and feedbacks"
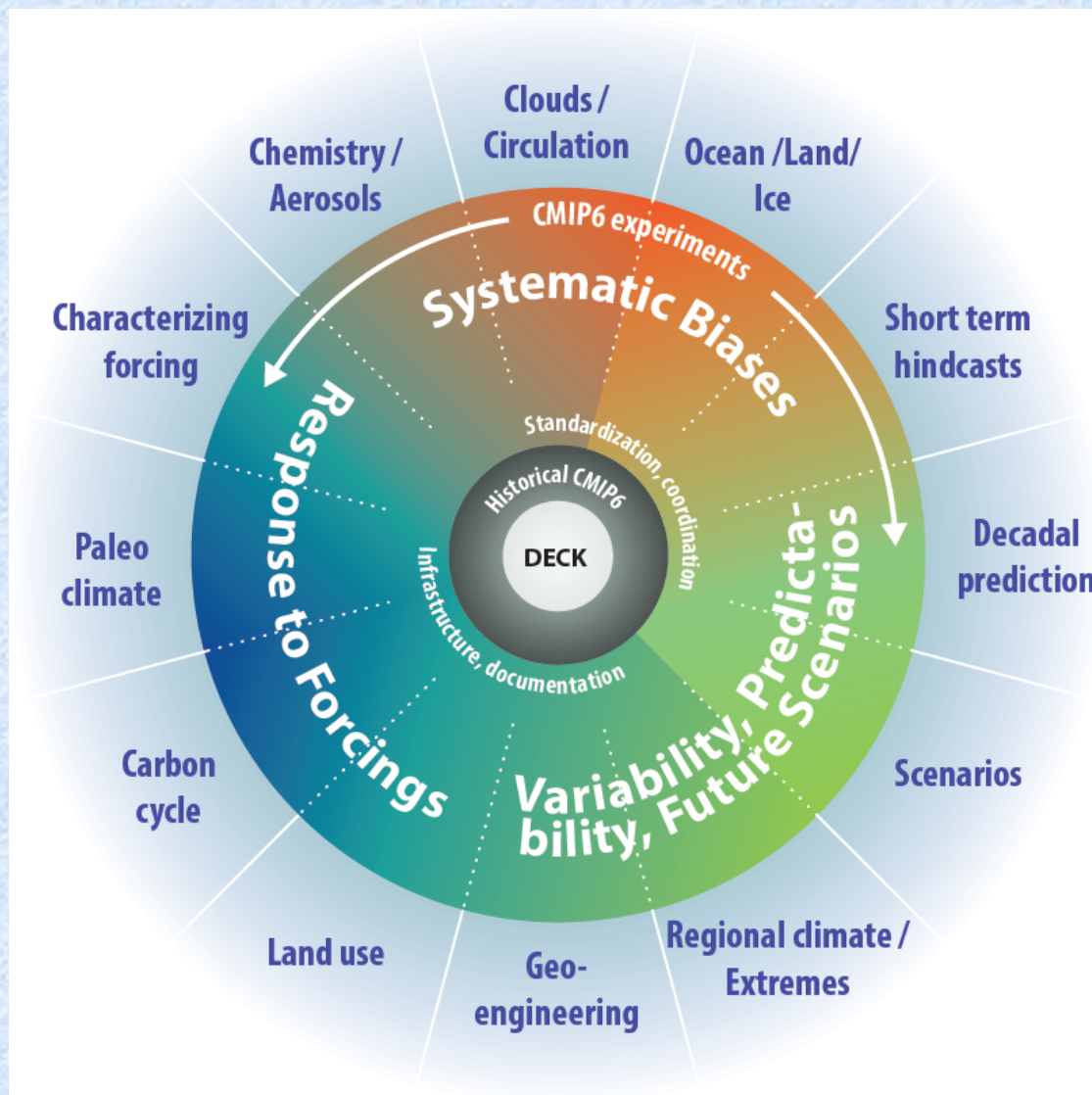

DECK (entry card for CMIP)
i.   AMIP simulation (~1979-2014)
ii. Pre-industrial control simulation
iii. 1%/yr CO2 increase
iv. Abrupt 4xCO2 run

Historical CMIP6 (entry card for CMIP6)
i.   Historical simulation (1850-2014)

(DECK & Historical CMIP6 to be run for each model configuration used in the subsequent CMIP6-Endorsed MIPs)

The DECK is expected to evolve only slowly across phases of CMIP.
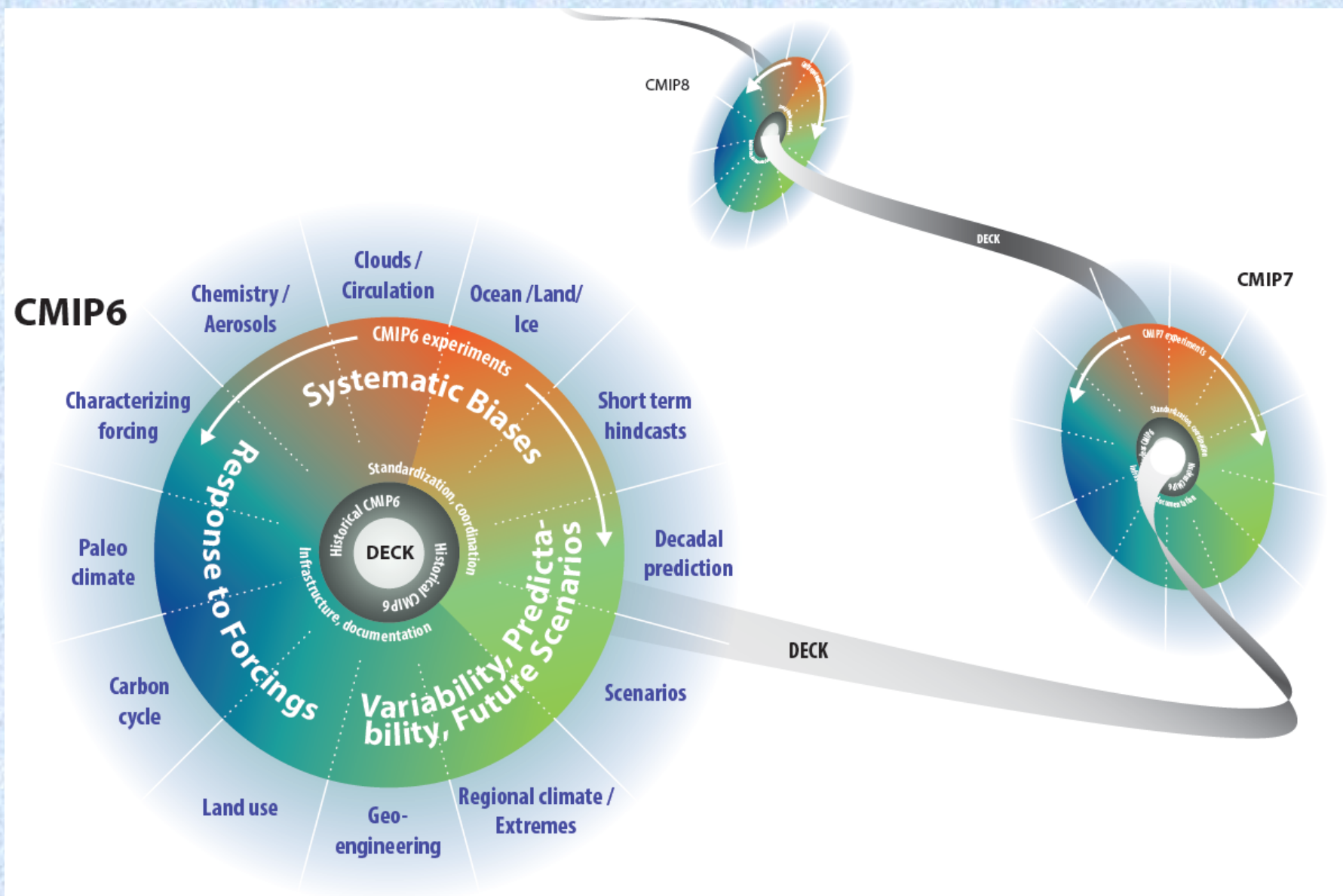
Timeline CMIP6 (~2015-2020)

- May 2015: Endorsed MIPs established and data request published

- January 2016: preindustrial forcing data sets ready.

- January 2016: First model output produced.

- July 2016: historical forcing ready.

- October 2016: future scenario forcing ready.

WGCM / modeling group concerns

- They will devote substantial resources to participate in CMIP and other MIPs

  - ➡ Imperative to minimize their effort
  - ➡ All MIPs should adopt similar data requirements
  - ➡ All MIPs should rely on common software and IT infrastructure

- The infrastructure needs to be robust and take into account modeling group resource commitments as ESGF evolves

  - ➡ The WGCM has expressed strong support for establishing a governance structure for ESGF that fully considers its impact on modeling group resources.

- They favor establishment of a more formal agreement between the data nodes and the ESGF software managers so that all parties clearly understand what resource commitments must be made and what procedures need to be followed.

WGCM / modeling group suggestions:

- Communicate plans and requirements at all stages

- Better document all operational procedures and formally establish a release schedule.

- Implement a procedure for testing and mandating installation of new releases of ESGF node software that takes into account resource impact on modeling groups

A number of activities must be coordinated in the development of modeling infrastructure

- 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

The WGCM established the WIP "to promote a robust and sustainable global data infrastructure in support of the WGCM's scientific mission"

- 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.

K. E. Taylor
PCMDI

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

WIP 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

- The WIP is also responsible for all the technical specifications for the CMIP data request.

K. E. Taylor
PCMDI

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

- CMIP6: similar

K. E. Taylor
PCMDI

Projected data volumes for CMIP6

- Model resolution likely to increase in CMIP6

  ➡ Some output might be reported at a coarser resolution

- Model output request not expected to increase by much

- Modeling groups say they will be able to simulate about the same number of years in CMIP6 as in CMIP5

- CMIP6 data volume preliminary estimate: less than 10 Pbytes

Needs: Reduce data volume transferred from archive to users

- Subset and concatenation capability (republishing all datasets with OPeNDAP should satisfy this, I think).

  - Single pressure level or subset of layers of multi-layer variables
  - "rectangular" (lat-lon) portion of a global field
  - Segment of or selected times from a time-series
  - Concatenate so data returned spans time samples contained in multiple files

- Data compression options?

- Simple server-side calculations (CDAT and LAS should satisfy this).

  - Collapse one axis
    - mean or sum
    - Variance , max, min
  - Form climatological annual cycle (from multiple years of data produce mean Jan., mean Feb., ... mean Dec.)

K. E. Taylor
PCMDI

Needs: Replication and versioning

- An automated dataset "replication" method is needed

- Establish a more uniform federation-wide method of identifying different versions of datasets

- Make it easy to trace reasons for withdrawal/replacement of datasets

  ⇒ Flawed metadata?

  ⇒ Flawed data?

  ⇒ Additional variables?

  ⇒ ???

Needs: Metrics, credit, provenance, etc.

1. Modeling groups want credit for the data they produce

    ➡ Cite models (documentation publication for each model?)

    ➡ Generate federation-wide download statistics

2. Researchers need to document what data were used in published research

    ➡ DOI's (or some tracking i.d.)

Problem: lots of models and lots of tracking i.d.'s per publication

K. E. Taylor
PCMDI

Needs: QC & Errata

- Can we move to a community-based approach?

  ➡ Web-based reporting of errors and responses to these reports

- Notification service

- Web-based service for user enquiries about whether files have been withdrawn and updated files are available

Additional needs

- The WCRP advocates free access to data, so consider developing a "relaxed-security" version of ESGF to

  ➡ Simplify software and make it operationally more robust?

  ➡ Make it easier for users?

- Increased capability/flexibility in searching and automating download procedures:

  ➡ Implement additional search options ("and" "or" constructions)

  ➡ Simplify scripted downloads

The WIP and the CMIP panel will continue to communicate evolving needs.

- CMIP6:

  ➡ http://www.wcrp-climate.org/wgcm-cmip/wgcm-cmip6

- WIP:

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

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.

Years simulated by each modeling group for CMIP5