

The WGCM Infrastructure Panel (WIP) and Priorities for CMIP6

Karl E. Taylor and V. Balaji

(on behalf of the CMIP Panel and the WIP)

Presented by Eric Guilyardi at the
2015 Aspen Decadal Workshop

June 2015

Outline

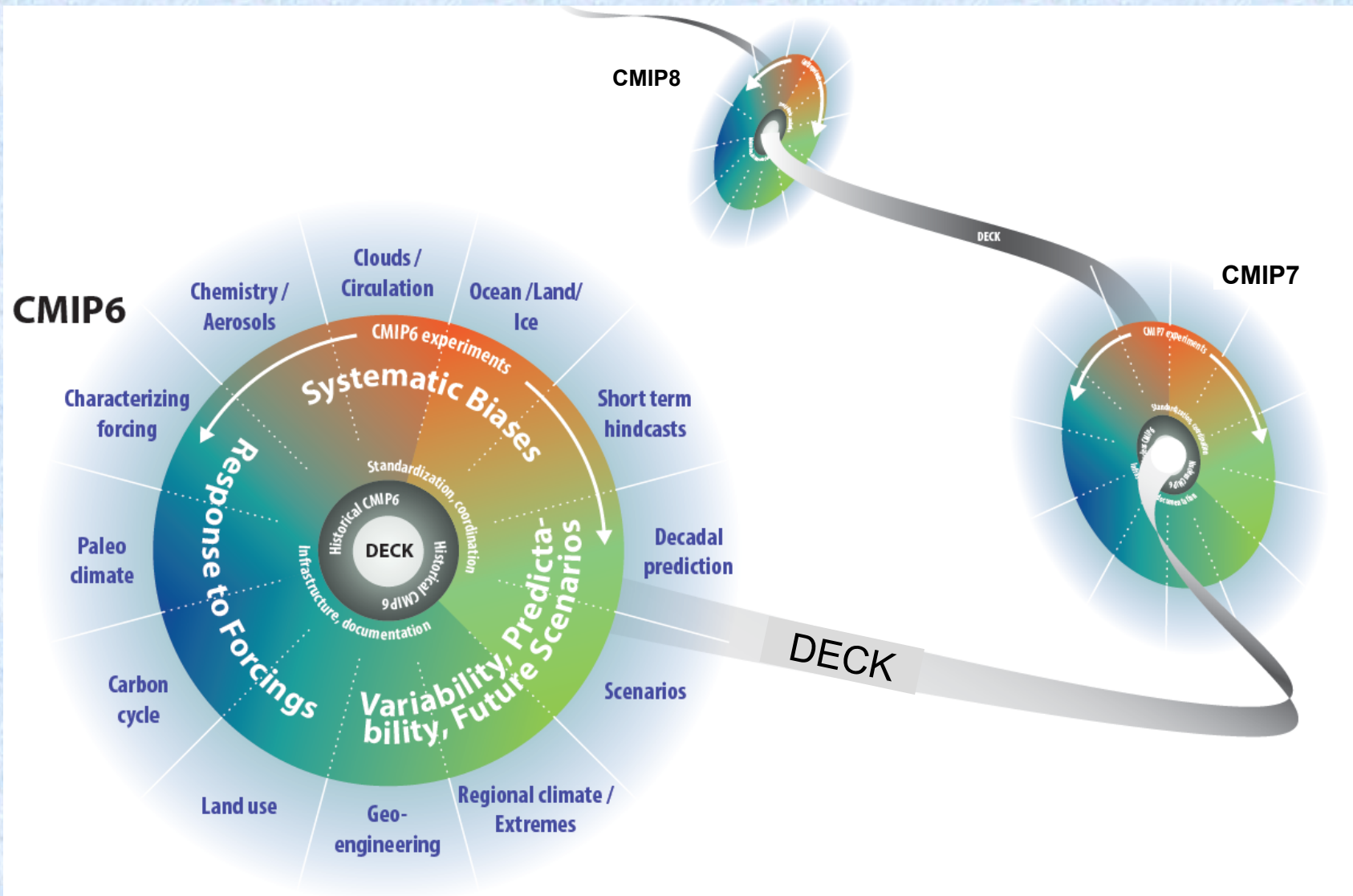
- Infrastructure implications for climate modeling infrastructure
- What is the role of the WGCM Infrastructure Panel (WIP)?
- What specifically is needed in the next few years?

Basics of CMIP6 coordination

- WCRP's Working Group on Coupled Modeling (**WGCM**) is responsible CMIP (independent of IPCC)
- WGCM oversees two panels:
 - **CMIP Panel** responsible for scientific aspects including experiment design and list of requested model output.
 - WGCM Infrastructure Panel (**WIP**) responsible for technical details, data standards, software infrastructure, etc.

(More about this later.)

CMIP provides continuity through DECK and an evolving suite of additional experiments addressing specific science questions.



Timeline CMIP6 (~2015-2020)

- May 2015: Endorsed MIPs established and data request compiled
- January 2016: preindustrial forcing data sets ready.
- January 2016: CMIP6 runs can begin.
- July 2016: historical forcing ready.
- October 2016: future scenario forcing ready.

WGCM / modeling group concerns and suggestions

- 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 WGCM encouraged evolution of model documentation approaches
 - Particularly critical to correctly document forcing datasets used
- WGCM / modeling group suggestions:
 - Communicate plans and requirements/expectations at all stages
 - Better document all operational procedures and formally establish a release schedule for ESGF.
 - 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

Purpose of CMIP "infrastructure"

- Ensure all model output can be easily ingested and analyzed by scientists
- Facilitate access to model output
- Make available information needed to interpret model output
 - Experiment details
 - QC and errata
- Provide access to documentation of models
- Record usage statistics

Without CMIP each center would likely follow a different approach impairing multi-model studies.

Summary of key design and infrastructure requirements

- Reduce demands placed on CMIP panel and PCMDI
- Communicate clearly scientists' needs to those developing and maintaining modeling infrastructure
- Establish better communication lines between modeling centers, MIP leaders, and infrastructure developers

For these purposes, the WGCM Infrastructure Panel (WIP) was established.

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.

WIP progress

- Established following the 2013 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 a few 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

Sébastien 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: should we move to more open access which would 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

WIP tasks: mix of detailed short term recommendations and longer-term forward-looking evolution

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: 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 $70000/17 = 4,100$ 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 $330000/26 = 13,000$ years
- **CMIP6**: similar to CMIP5, but somewhat higher resolution models (<10 PB)

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.)

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

Needs: QC & Errata

- Should 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

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

- CMIP6:

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

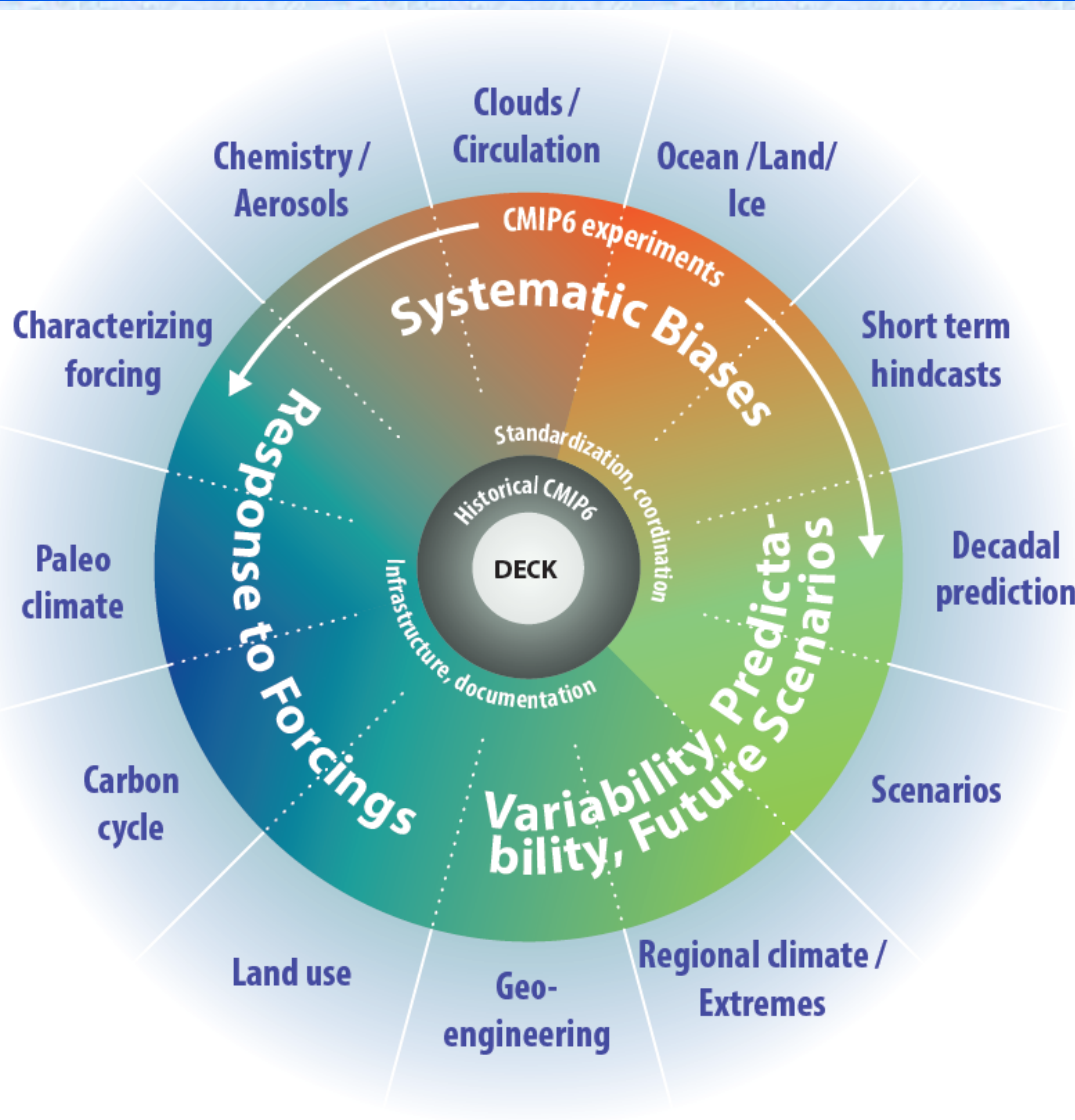
- WIP:

- <https://www.earthsystemcog.org/projects/wip/>

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

- CMIP5
 - Centrally organized
 - Multiple additional MIPs
 - Resource-intensive
- Future coordinated model activities (CMIP & CMIP6)
 - Collection of coordinated independently managed MIPs
 - Basic, routinely performed limited set of experiments (CMIP DECK)
 - 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 and related funded projects will evolve to meet all the needs.

CMIP6 design summary:



DECK

- Small set of benchmark runs
- To evolve only slowly (e.g. OMIP, LMIP)

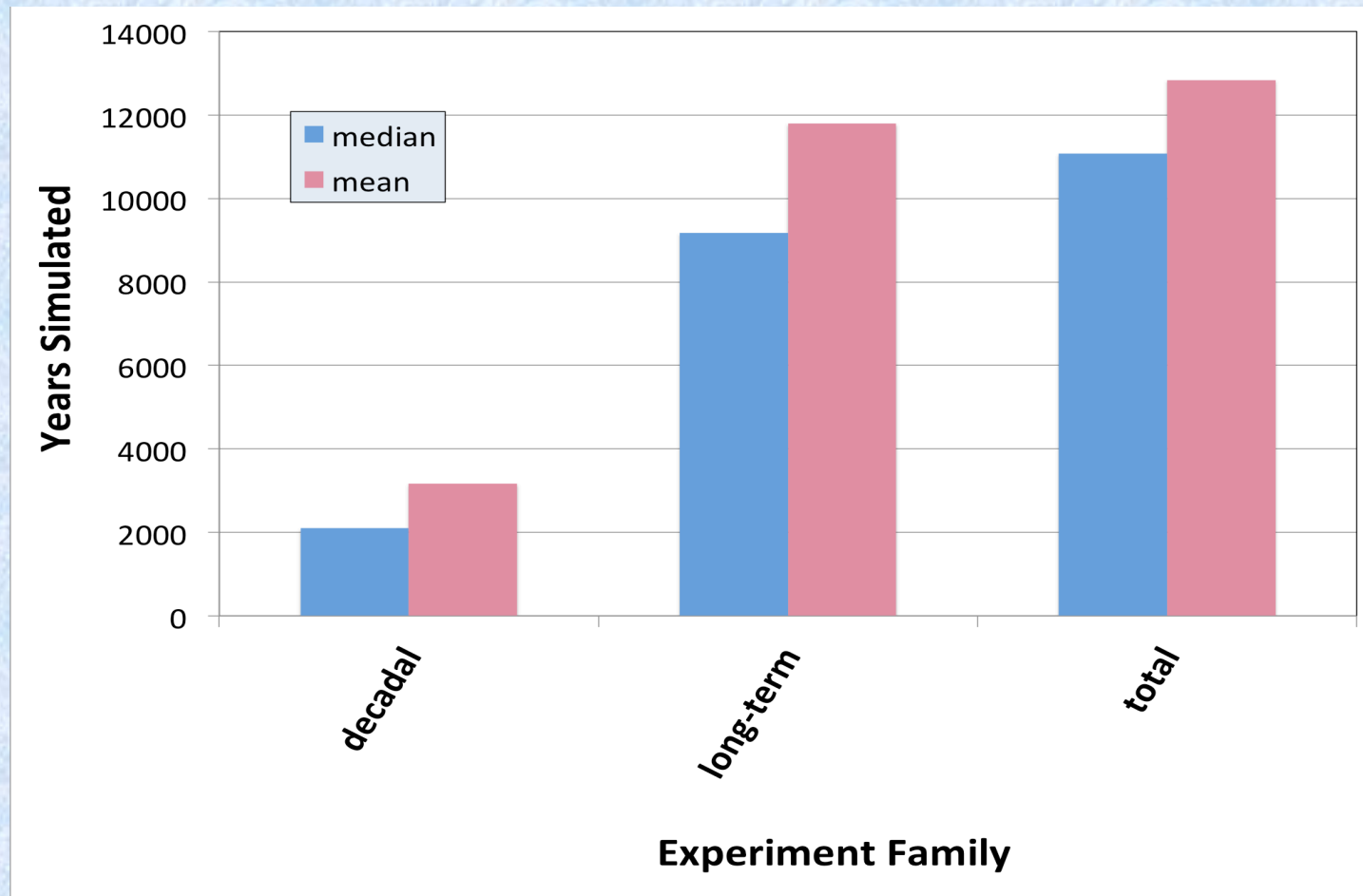
Historical CMIPX

- Forcing to be updated for each new phase

CMIP6-endorsed MIPs

- An evolving collection to address specific scientific issues

CMIP5: Mean and median number of years simulated per modeling group participating in expt. family.



The community-wide adoption of the DECK and common data standards has additional benefits

- Obs4MIPs started
 - NASA, DOE, and now the WCRP are promoting adoption of the same data standards for observational datasets
 - Provide users with datasets tailored to model evaluation.
- Adoption of consistent standards across models and observations facilitates community development of diagnostic and metrics packages
 - Diagnostic and model evaluation software can be shared among modeling centers and the wider community
 - Metrics from all models can be collected and used to highlight and summarize relative performance.

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.