CMIP's Requirements of the ESGF

A Report on Behalf of the WGCM Infrastructure Panel (WIP)

Karl E. Taylor & V. Balaji

(WIP co-chairs)

Presented at the
Earth System Grid Federation (ESGF)
Face to Face Conference

Monterey, CA
8 December 2015

WIP members: 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

Francesca Guglielmo, LSCE

Eric Guilyardi: IPSL

Martin Juckes: BADC

Slava Kharin: CCCma

Michael Lautenschlager: DKRZ

Bryan Lawrence : NCAS, BADC

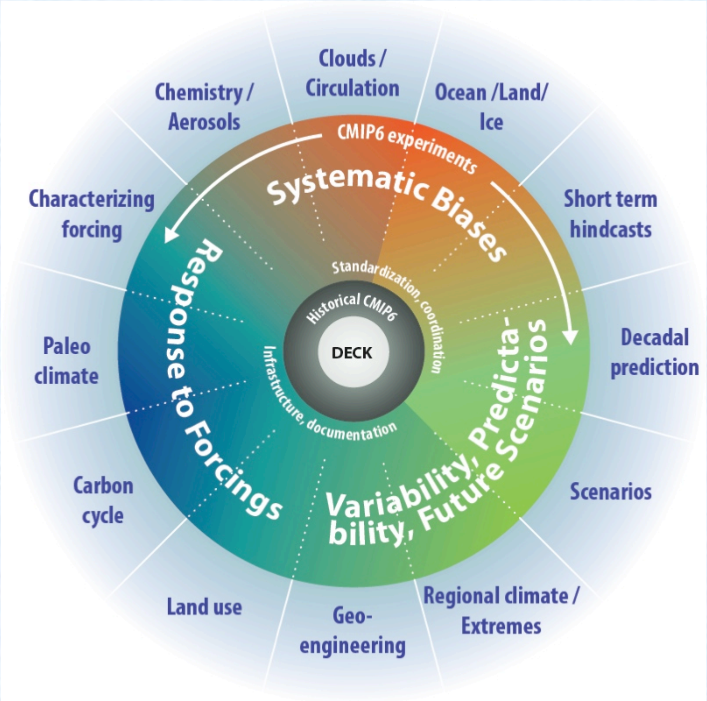
Dean Williams: PCMDI

Outline

- CMIP project overview.
- What are the connections between CMIP's organizational structure and ESGF?
- What are CMIP6's requirements of ESGF?

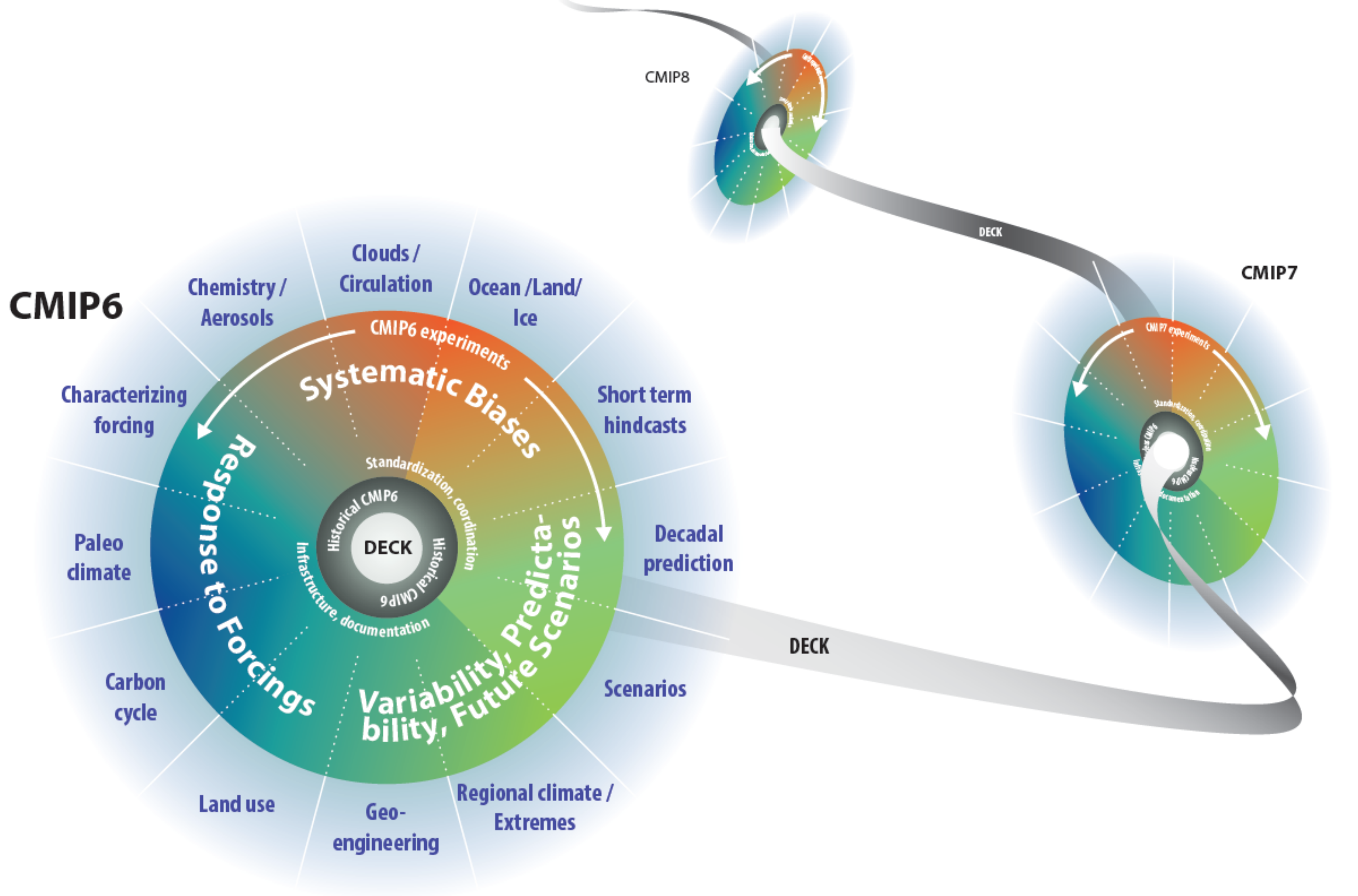
CMIP6 builds on an ongoing core of CMIP experiments

- Diagnosis, Evaluation and Characterization of Klima (DECK)
  - Include:
    - AMIP (~1979-2014)
    - Pre-industrial control
    - 1%/yr CO2 increase
    - Abrupt change to 4xCO2
  - Performed whenever a new model is developed (no deadlines)

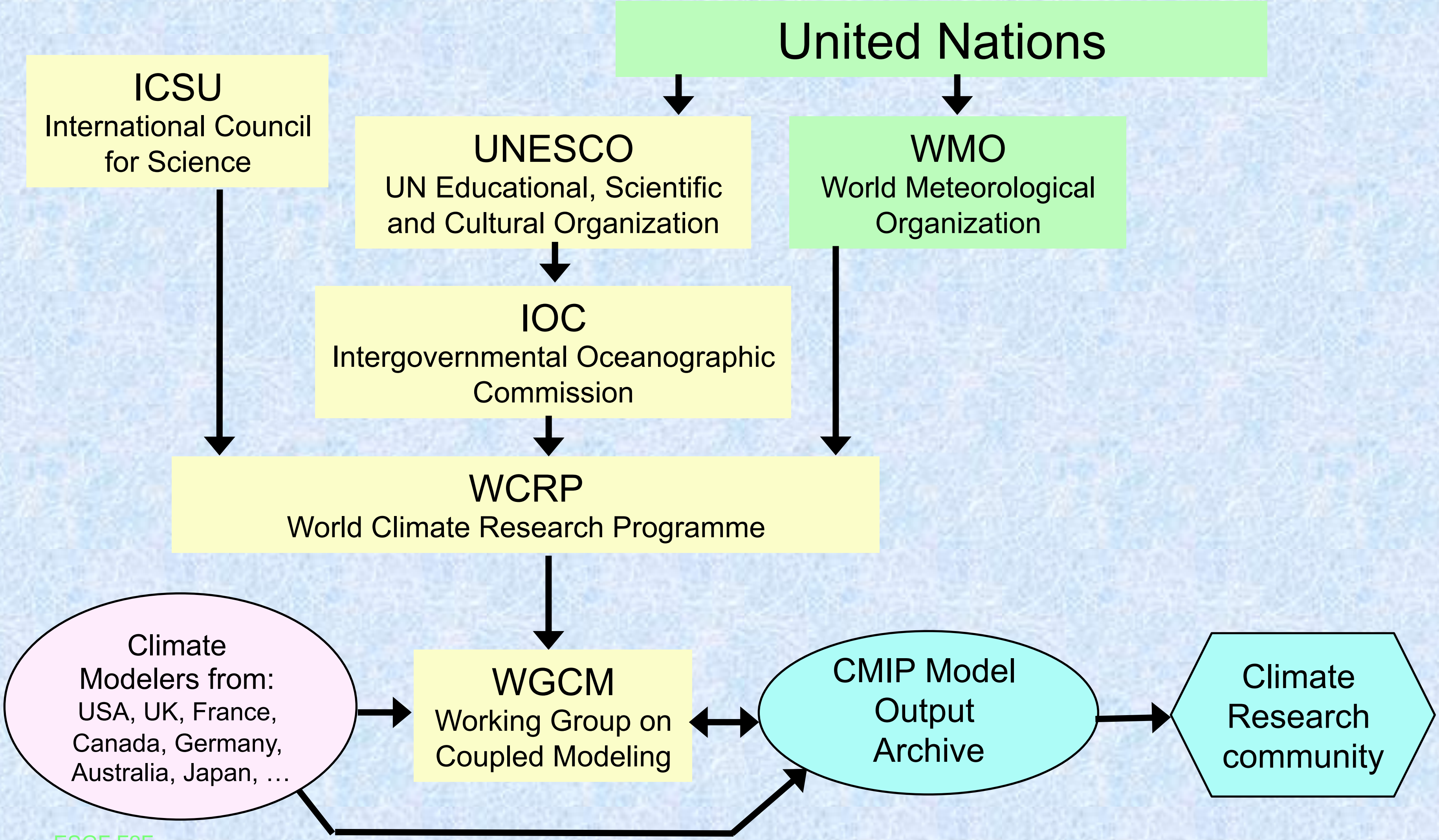


- Historical run
  - Historical forcing updated for each CMIP phase
  - Required for CMIP6 participants
- CMIP6-endorsed MIPs
  - Modeling groups will choose to participate in a subset, depending on scientific interest and resources.

New CMIP design is adapted to address new science questions as they emerge.



CMIP: Part of an internationally-coordinated research program



United Nations

ICSU
International Council for Science

UNESCO
UN Educational, Scientific and Cultural Organization

WMO
World Meteorological Organization

IOC
Intergovernmental Oceanographic Commission

WCRP
World Climate Research Programme

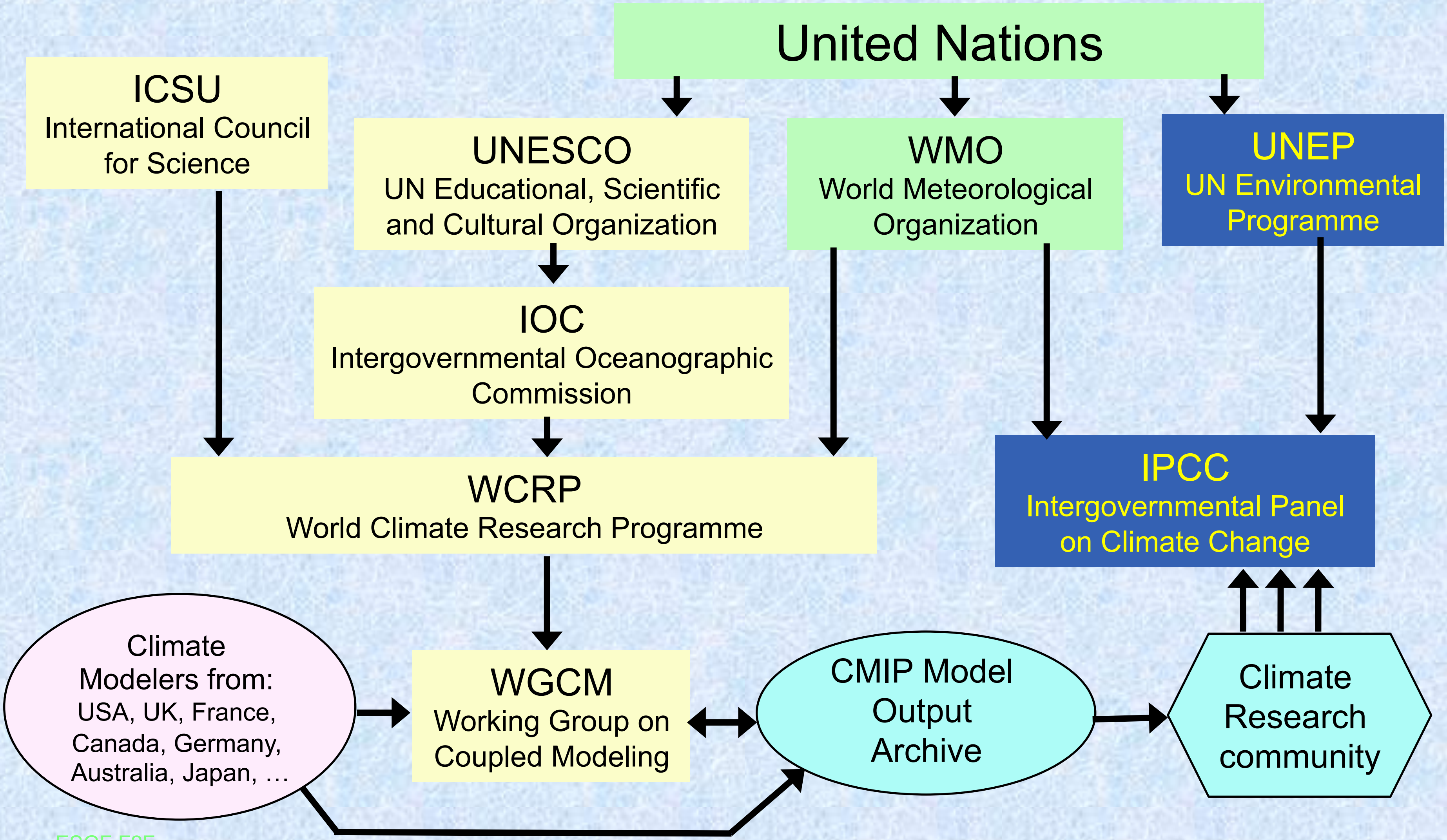
Climate Modelers from:
USA, UK, France, Canada, Germany, Australia, Japan, …

WGCM
Working Group on Coupled Modeling

CMIP Model Output Archive

Climate Research community

IPCC assessments are separate from the international climate research programs



United Nations

ICSU
International Council for Science

UNESCO
UN Educational, Scientific and Cultural Organization

WMO
World Meteorological Organization

UNEP
UN Environmental Programme

IOC
Intergovernmental Oceanographic Commission

WCRP
World Climate Research Programme

IPCC
Intergovernmental Panel on Climate Change

Climate Modelers from:
USA, UK, France, Canada, Germany, Australia, Japan, …

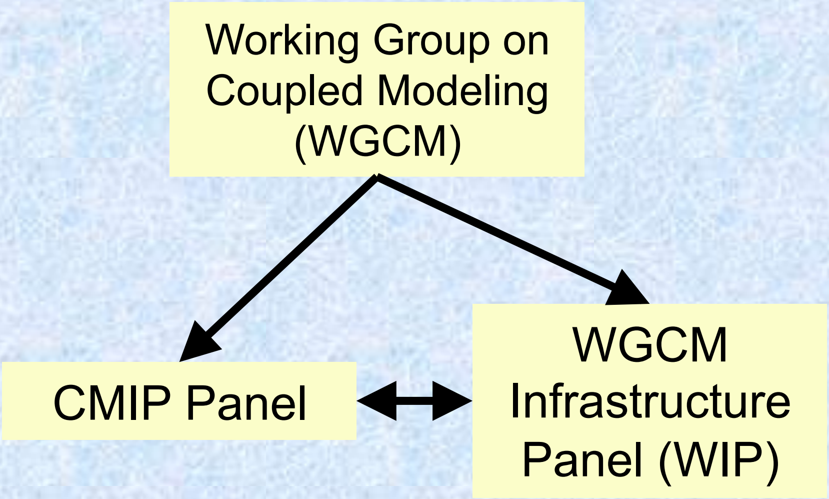
WGCM
Working Group on Coupled Modeling

CMIP Model Output Archive

Climate Research community

CMIP coordination

- The WGCM (WCRP's Working Group on Coupled Modelling)
  - Represents the modeling groups and wider climate research community
  - Founded CMIP and provides oversight
- The CMIP Panel manages and coordinates scientific aspects (experiment design, output request, forcing data sets, etc.)
- The WGCM Infrastructure Panel (WIP) manages and coordinates infrastructure development, implementation, and operations.


Working Group on Coupled Modeling (WGCM)
CMIP Panel
WGCM Infrastructure Panel (WIP)

Why a WCRP Infrastructure Panel (WIP)?

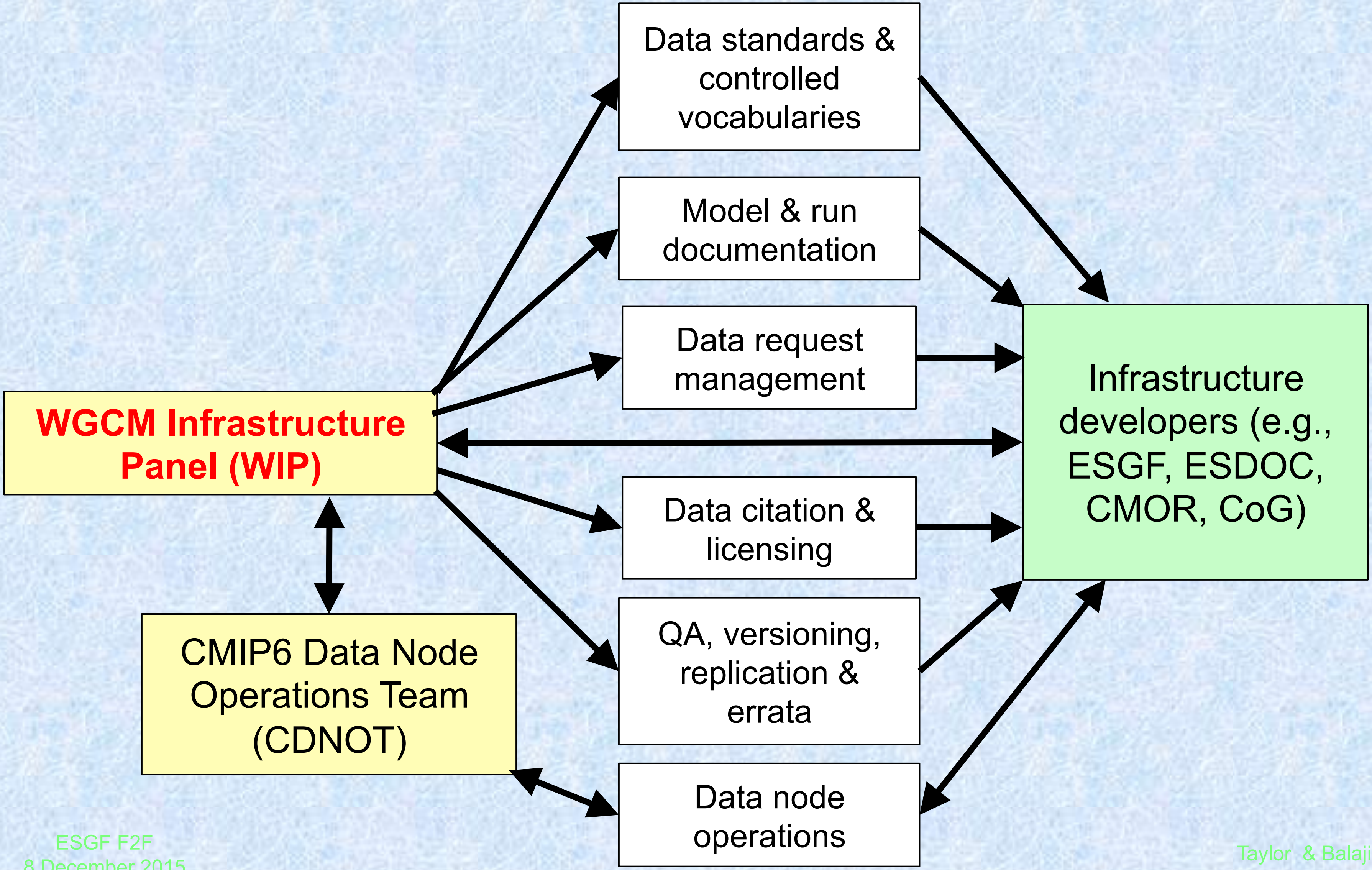
- CMIP and other "MIPs" involve a huge resource commitment by modeling groups

- The modeling groups through the WGCM rely on:
  - The CMIP panel to coordinate MIP design to maximize scientific return on their resource investment
  - The WIP to identify and articulate the infrastructure requirements of CMIP and minimize the technical burdens it places on modeling groups

- The WIP's authority derives from its connection to the WGCM and the modeling groups.

- The WIP is responsible for facilitating access to CMIP data, providing input to ESGF as to priorities, and setting and enforcing CMIP data node requirements.

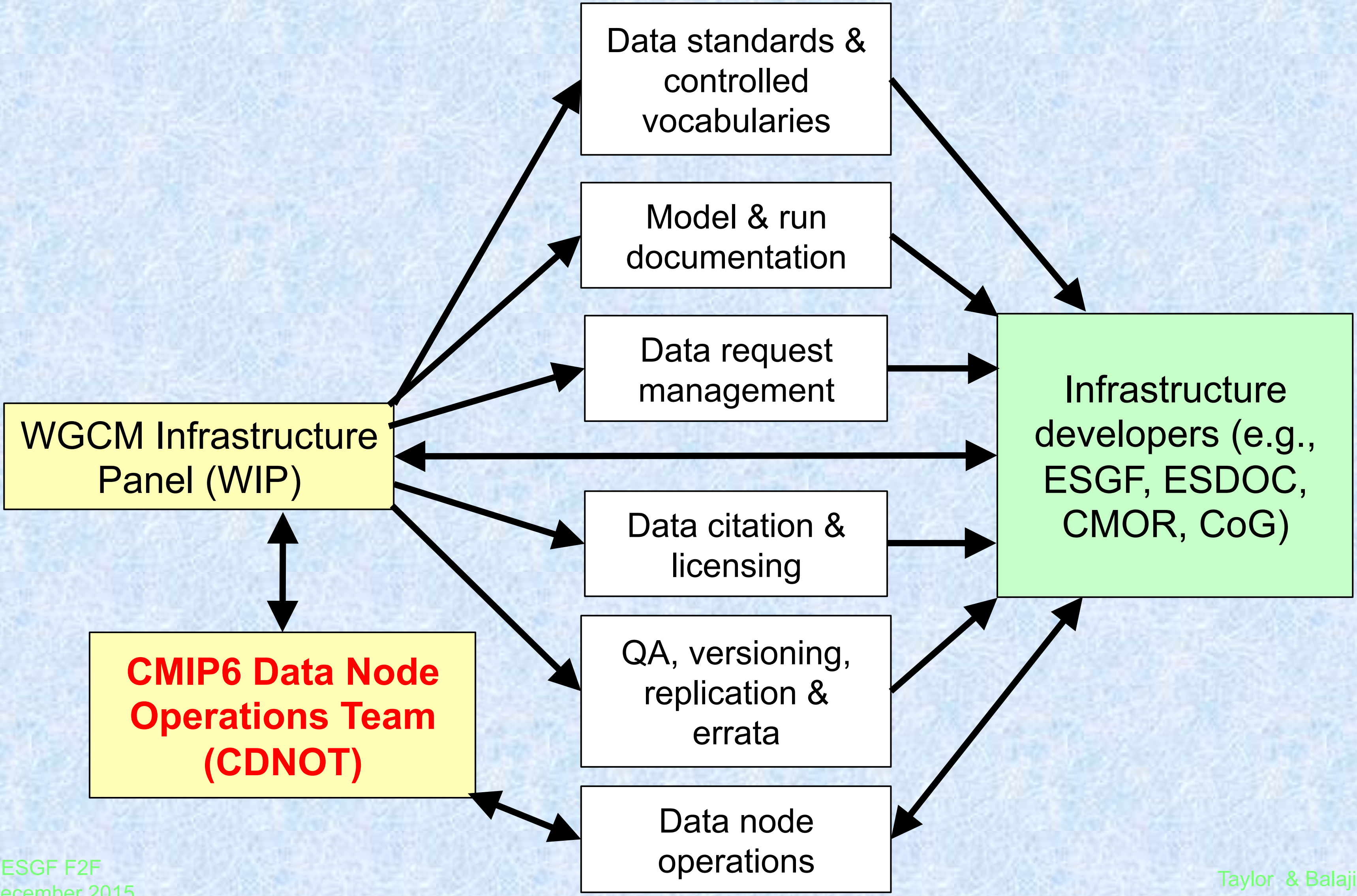
MIP infrastructure coordination, development, and implementation


Data standards & controlled vocabularies
Model & run documentation
Data request management
WGCM Infrastructure Panel (WIP)
Infrastructure developers (e.g., ESGF, ESDOC, CMOR, CoG)
Data citation & licensing
QA, versioning, replication & errata
CMIP6 Data Node Operations Team (CDNOT)
Data node operations

WIP work: Ensure infrastructure is in place to enable scientific research based on CMIP model output.

- Handle technical aspects of the CMIP data request
- Write position papers detailing CMIP requirements for infrastructure
  - 10 papers in the works (half ready for community review)
- Communicate requirements specific to CMIP for configuration and operation of data nodes
- Represent modeling centers in communicating their concerns to ESGF and others developing infrastructure.

MIP infrastructure coordination, development, and implementation



WGCM Infrastructure Panel (WIP)

CMIP6 Data Node Operations Team (CDNOT)

Data standards & controlled vocabularies

Model & run documentation

Data request management

Data citation & licensing

QA, versioning, replication & errata

Data node operations

Infrastructure developers (e.g., ESGF, ESDOC, CMOR, CoG)

CMIP Data Node Operations Team (CDNOT)

- A technical consortium charged with applying and operationalizing ESGF for CMIP6

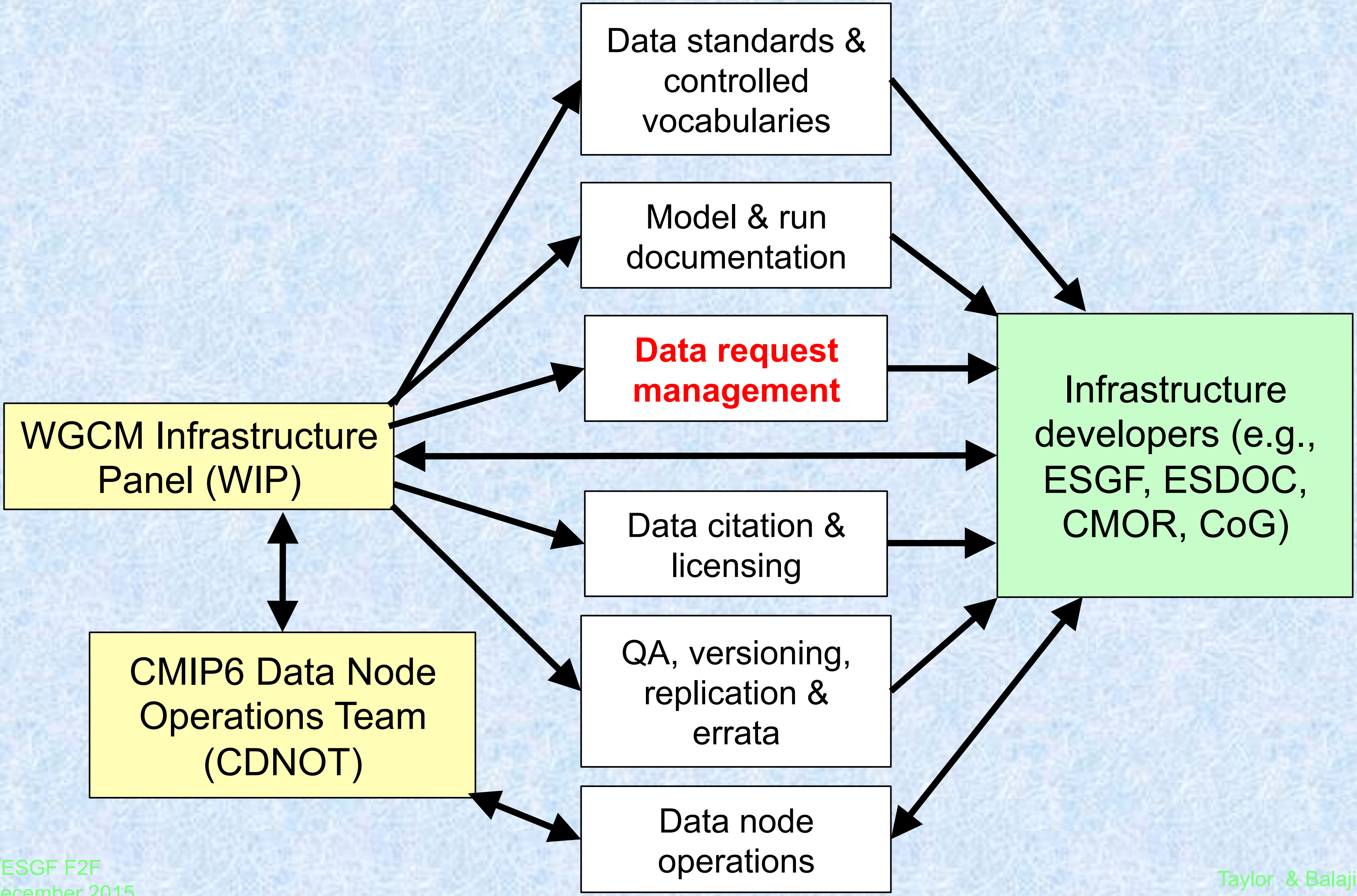
- Sébastien Denvil (IPSL) chairs

- Members representing each site hosting CMIP6 data (i.e., most modeling centers and major data centers)

- Membership overlaps with bodies responsible for requirements (WIP) and software development (ESGF, ESDOC, ...)

- Serves to:
  - Communicate WIP discussion to all those of interest
  - Provide input to the WIP of data node/modeling center concerns

MIP infrastructure coordination, development, and implementation



WGCM Infrastructure Panel (WIP)

CMIP6 Data Node Operations Team (CDNOT)

Data standards & controlled vocabularies

Model & run documentation

Data request management

Data citation & licensing

QA, versioning, replication & errata

Data node operations

Infrastructure developers (e.g., ESGF, ESDOC, CMOR, CoG)

4 Position papers: Management of CMIP6 data request

- Led by Martin Juckes (STFC)
- Documents dealing with technical aspects of CMIP6 data:
  - Compilation of list of variables
  - Specification of file name template and file global attributes
  - Specification of “controlled vocabulary” enabling automated management and utilization of CMIP archive
- Status:
  - MIPs submitted list of variables they need
  - CMIP Panel will review for completeness and reasonableness
  - Iterate with MIPs and modeling groups

Relevance of data request to ESGF

- No big changes from CMIP5, but improved database describing variable request suitable for machine interpretation

- A few new global attributes will be harvested for ESG catalogs, e.g.,
  - An extended "tracking_id" (hdl:21.14100/<uuid>)
  - Additional search facets (e.g., a grid indicator)

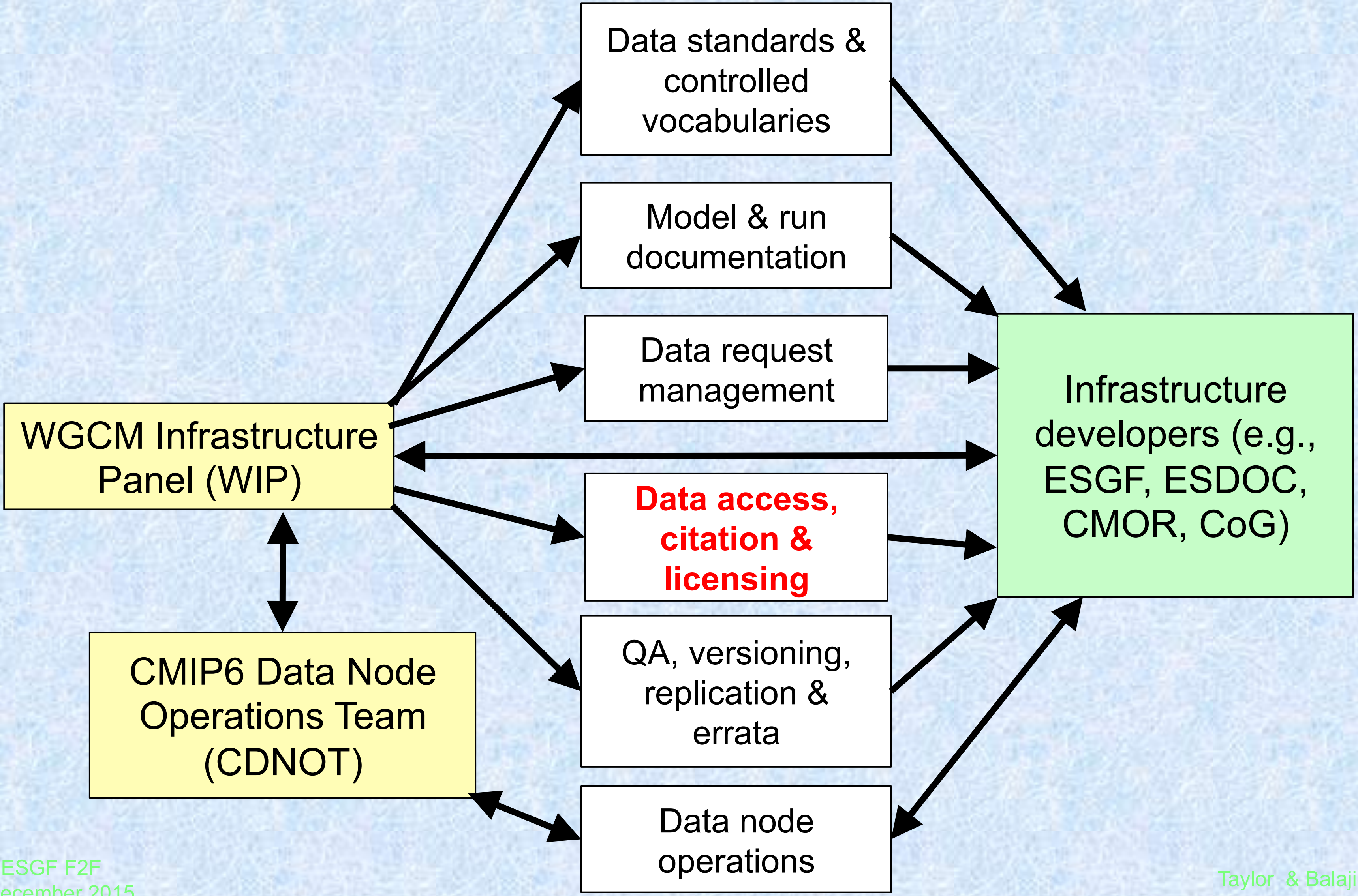
- Filename modified (segment included in CMIP5)

```
file name = <variable_id>_<table_id>_<experiment_id>
        [-sub1_expt_id[-sub2_expt_id]]_
        <source_id>_<run_variant_id>_<grid_id>
        [_<driving_source_id>_<driving_variant_id>]
        [_{time_range}].nc
```

Example:

pr_day_decadal-1960_GFDL-CM2-1_r2i1p1f3_gn_196001-199912.nc

MIP infrastructure coordination, development, and implementation



WGCM Infrastructure Panel (WIP)

CMIP6 Data Node Operations Team (CDNOT)

Data standards & controlled vocabularies

Model & run documentation

Data request management

Data access, citation & licensing

QA, versioning, replication & errata

Data node operations

Infrastructure developers (e.g., ESGF, ESDOC, CMOR, CoG)

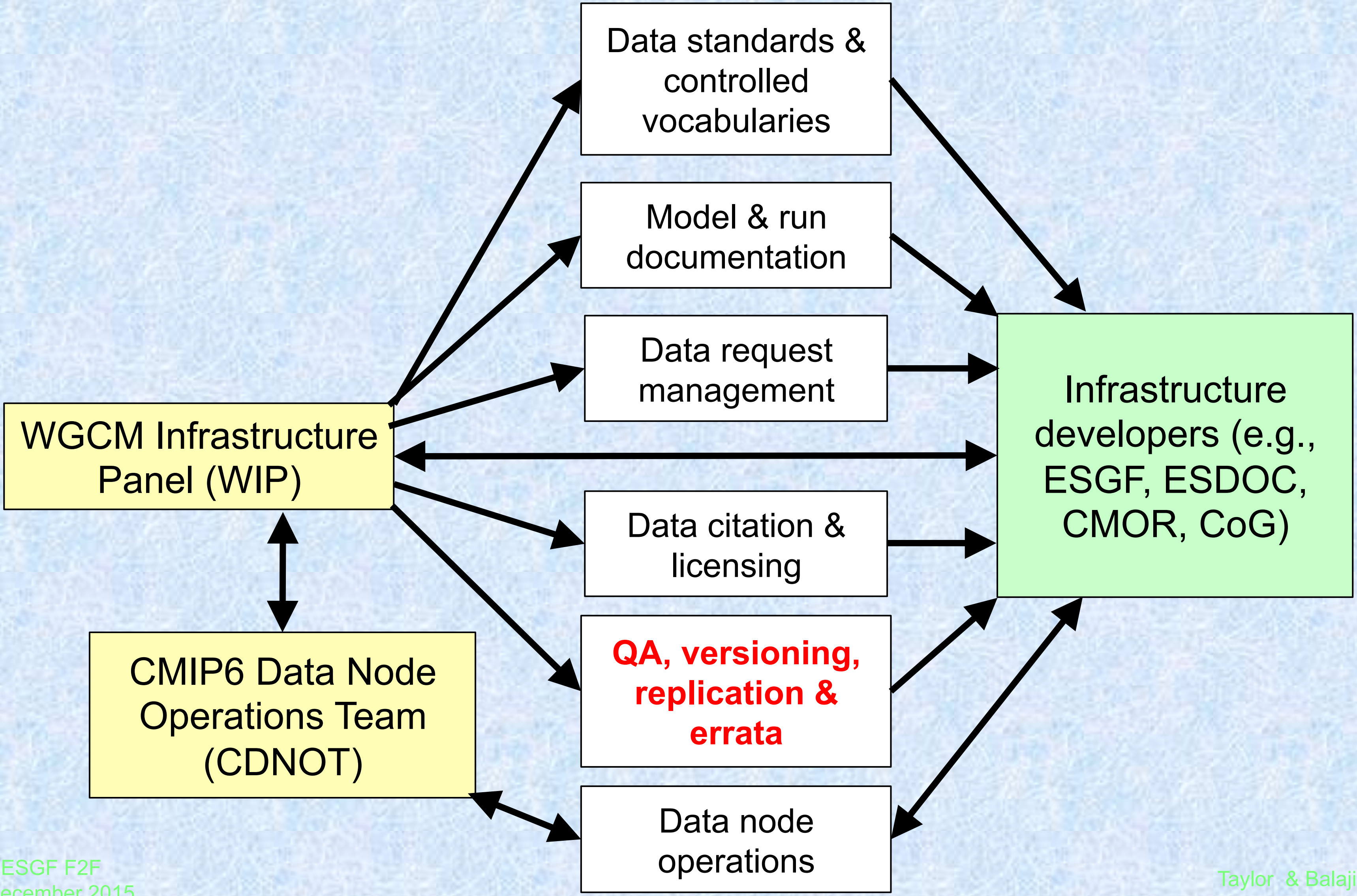
2 Position papers: Data access, citation and licensing

- Main requirements:
  - Ensure proper citation of data acknowledging contributions by modeling groups
  - Reduce ESGF's role in attempting to impose constraints on data usage
- Specifics:
  - A Creative Commons Licensing agreement will be recorded as a global attribute in all files. ("share-alike" or "non-commercial share-alike")
  - Citation requirement will be included in the terms of use
  - Instructions for registration of use of data will also appear in a global attribute
    - Document usage
    - Enable notices of data retraction or updates

Data access, citation and licensing

- Specifics (continued):
  - Enable citation via
    - DOI's assigned at the model/simulation granulation
    - Links connecting datasets to model and experiment documentation (ESDOC/CIM)

MIP infrastructure coordination, development, and implementation



WGCM Infrastructure Panel (WIP)

CMIP6 Data Node Operations Team (CDNOT)

Data standards & controlled vocabularies

Model & run documentation

Data request management

Data citation & licensing

QA, versioning, replication & errata

Data node operations

Infrastructure developers (e.g., ESGF, ESDOC, CMOR, CoG)

4 Position papers: Quality assurance (QA), replication, versioning, and errata

- User requirements:
  - Have data I've downloaded been modified or withdrawn?
  - What errors were detected in modified/withdrawn files?
  - Have relevant data from additional models become available since I last checked the archive?
  - What level of QC has been passed by each dataset?
  - What DOI's should be used (in citations) to indicate which models and experiments were relied on in my scientific study?
- Scientific requirements:
  - For reproducibility, PID's are needed to record as supplementary citation information which files were analyzed.
  - Metadata for at least some withdrawn or replaced dataset versions should be preserved.

Quality assurance (QA), replication, versioning, and errata (cont.)

- ESGF operational requirements
  - Automated QC capability to ensure adherence to metadata and data quality standards
  - Automatic rejection of datasets that fail to meet minimal conformance standards (and say why rejected)
  - Automated, monitored data replication between ESGF nodes

Specific requirements for quality assurance (QA), replication, versioning, and errata

- A single "publication unit" or "atomic dataset" should include only data from a single variable from a single model simulation sampled at a single frequency.
  - Makes versioning and replication more practical
  - Compared with CMIP5, this increases the number of datasets by more than an order of magnitude.
- Rely on extended "tracking_id" (hdl:21.14100/<uuid>) as the primary PID.
  - To be used as supplementary citation information in papers.
  - To be harvested by a PID-based query system to see if errata have been reported, or data have been superseded.
  - To aid in automated replication and tracking of versions.
- Provide for community annotation/comment on datasets (QC)

Quality assurance (QA) procedure should:

- Check that all required metadata are include
- Check conformance with controlled vocabularies
- Check filenames conform with required template
- Check that model & experiment documentation exists (on a "landing page" or in ES-DOCs?)
- Enable user access to conformance flag indicating which tests have been passed

Additional CMIP6 requirements (white papers yet to be written)

- Provide search capability to access a single simulation via multiple projects.

- Enable server-side processing, including
  - Subsetting (e.g., single model level; single region)
  - Calculating means (e.g., climatology; zonal mean; ensemble mean)
  - Reformatting/decompressing files (from netCDF4 to netCDF3)
  - Regridding (from native to standard grids)

- Implement procedures and test suites for enhancing ESGF software that minimize impact on CMIP data nodes.

- Seek input from CMIP users and modeling groups (through the WIP, the CMIP panel, and the WGCM) as to priorities for ESGF development.

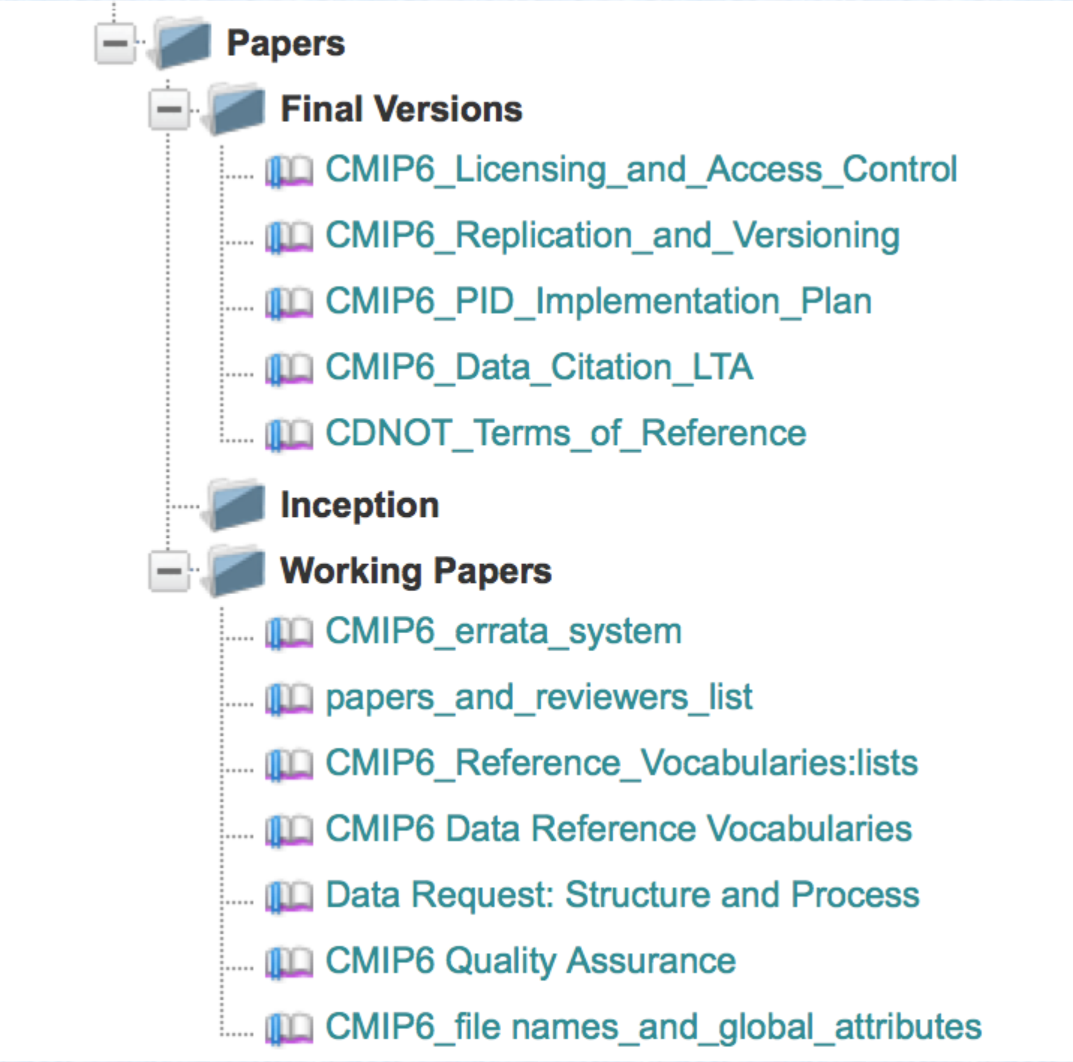
- Link ESGF to ES-DOCs to incorporate model/experiment documentation into QC tests and to provide easy access to users.

Concluding remarks

- CMIP6 presents new challenges for ESGF
- ESGF's CMIP5-era infrastructure provides a solid base on which to build
- The WIP has prepared 10 white papers describing CMIP6 requirements, which have been summarized in a draft document soon to be released.
- Some modeling groups will likely begin preparing model output for publication on ESGF sometime during the first half of next year.
- Total archive storage expected to be order 10's of petabytes
- Your contributions in addressing these requirements are essential to the success of CMIP6.

We invite input.

Further details: https://www.earthsystemcog.org/projects/wip/resources/



- Papers
  - Final Versions
    - CMIP6_Licensing_and_Access_Control
    - CMIP6_Replication_and_Versioning
    - CMIP6_PID_Implementation_Plan
    - CMIP6_Data_Citation_LTA
    - CDNOT_Terms_of_Reference
  - Inception
  - Working Papers
    - CMIP6_errata_system
    - papers_and_reviewers_list
    - CMIP6_Reference_Vocabularies:lists
    - CMIP6 Data Reference Vocabularies
    - Data Request: Structure and Process
    - CMIP6 Quality Assurance
    - CMIP6_file names_and_global_attributes