CMIP6:

WIP Status Update

A summary of the CMIP6 project to-date from the WGCM Infrastructure Panel (WIP)

Paul J. Durack, Matthew Mizielinski and Karl E. Taylor

WGCM23
9th December 2020 - Virtual

LLNL-PRES-817548
This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

Lawrence Livermore National Laboratory

Outline

- WIP background and membership transitions
- Timelines - calibrate where we've come from and where we are
- CMIP6
    - Infrastructure components
    - Successes to date
    - Components and their contribution
        - CVs, Data Request, CMOR, ES-DOC, ESGF Citation service
    - Lessons learned
- Conclusions

What is the WGCM Infrastructure Panel* (WIP)?

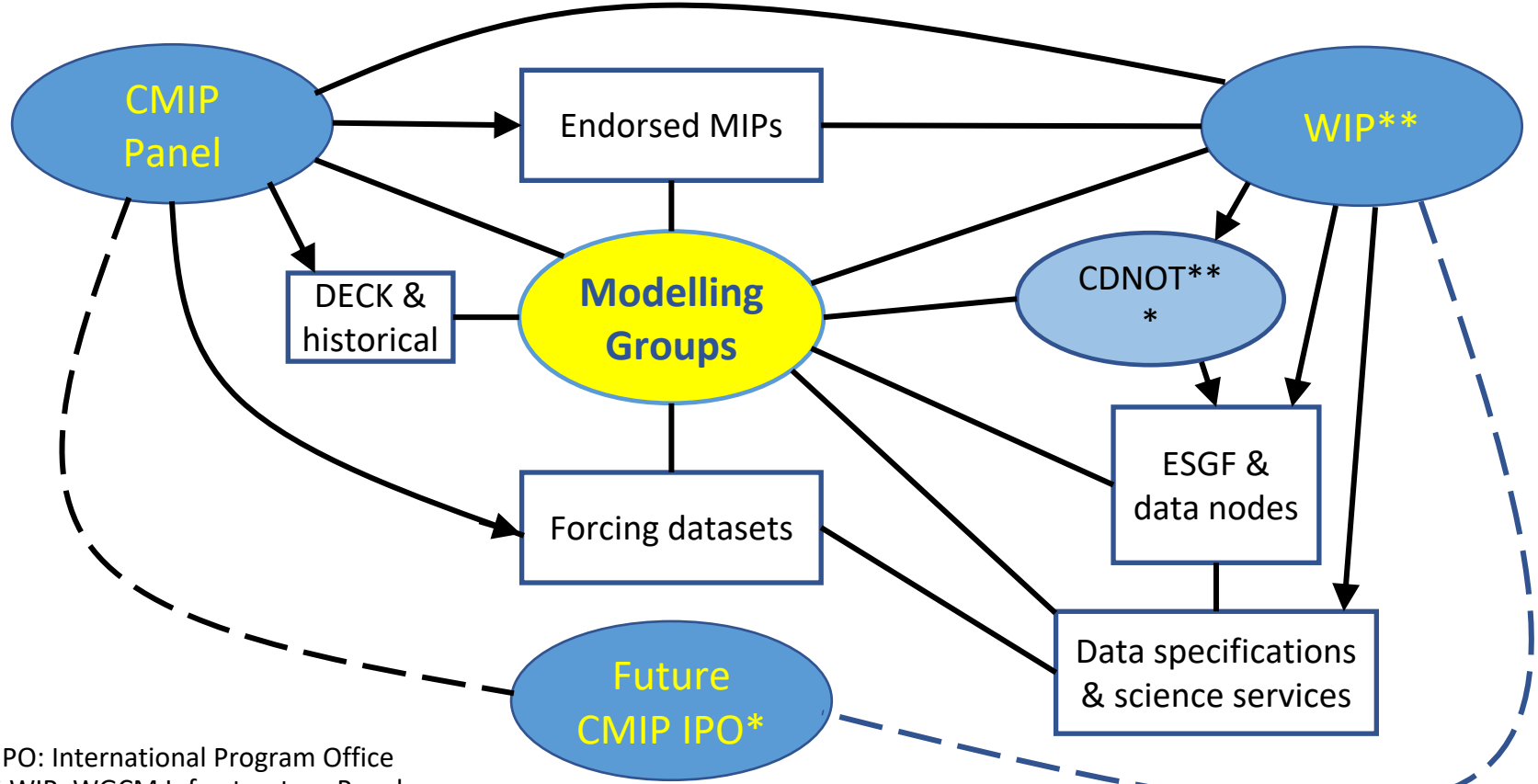
Established by the WGCM in June 2014 to:

- Set requirements ensuring the CMIP infrastructure will serve its purpose
- Write documents defining specifications for the infrastructure and data it hosts.
- Help coordinate development work done by funded infrastructure projects so that the elements work well together
- Communicate and coordinate with data managers at modeling groups via a “CMIP Data Node Operations Team” (CDNOT)

* G. Abdulla, S. Ames, Y. Bai,, P. J. Durack, D. Hassell M. Juckes, S. Kharin, M. Lautenschlager, M. Mizielinski, R. Petrie, M. Stockhause, K. E. Taylor

https://www.wcrp-climate.org/wgcm-cmip/wip

WCRP/WGCM Organizational Structure for CMIP



* IPO: International Program Office
** WIP: WGCM Infrastructure Panel
*** CDNOT: CMIP Data Node Operations Team

Arrows indicate direct responsibility or oversight role

WIP membership transitions

- The WIP lost some key members about a year ago:
  - The WIP co-chair and the 2 P.I.s for ES-DOCs resigned
    - Multiple reasons were cited, but fundamentally it reflects the stress placed on the inadequately funded projects that support CMIP infrastructure
    - There was also a breakdown in communication across the CMIP management structure
    - It is a warning sign that all CMIP participants need to include infrastructure planning and support in all phases
  - The CDNOT chair accepted a position at ECMWF
- Ruth Petrie (CEDA) is the new CDNOT chair
- Matthew Mizielinski (Hadley Centre) is the new WIP co-chair
- Next month Paul Durack (PCMDI) will replace Karl Taylor as the other WIP co-chair

Outline

- WIP background and membership transitions
- Timelines - calibrate where we've come from and where we are
- CMIP6
    - Infrastructure components
    - Successes to date
    - Components and their contribution
        - CVs, Data Request, CMOR, ES-Doc, ESGF Citation service
    - Lessons learned
- Conclusions

Timeline of the *MIPs

*MIPs temporal evolution and their support infrastructure

Planning begins	1989	1993	1995	1997	2003	2008	2014
Simulations	AMIP1	AMIP2	CMIP1	CMIP2/ CMIP2+	CMIP3	CMIP5	CMIP6
Data volume	1GB	500GB	1GB	500GB	50TB	2PB	>20PB
Host infrastructure	LLNL FTP*	LLNL FTP	LLNL FTP	LLNL FTP	LLNL FTP	ESGF 41# nodes	ESGF 31 nodes
Operations begin	1989	1995	1996	1999	2004	2011	2018

Next slide focuses on this period

*For some groups in addition to data distribution, LLNL computing facilities were used to run contributing simulations

#Just 17 CMIP5 nodes remain to-date, which means a ~50% loss rate from the CMIP5 peak - highlights tier 1 node importance for data preservation

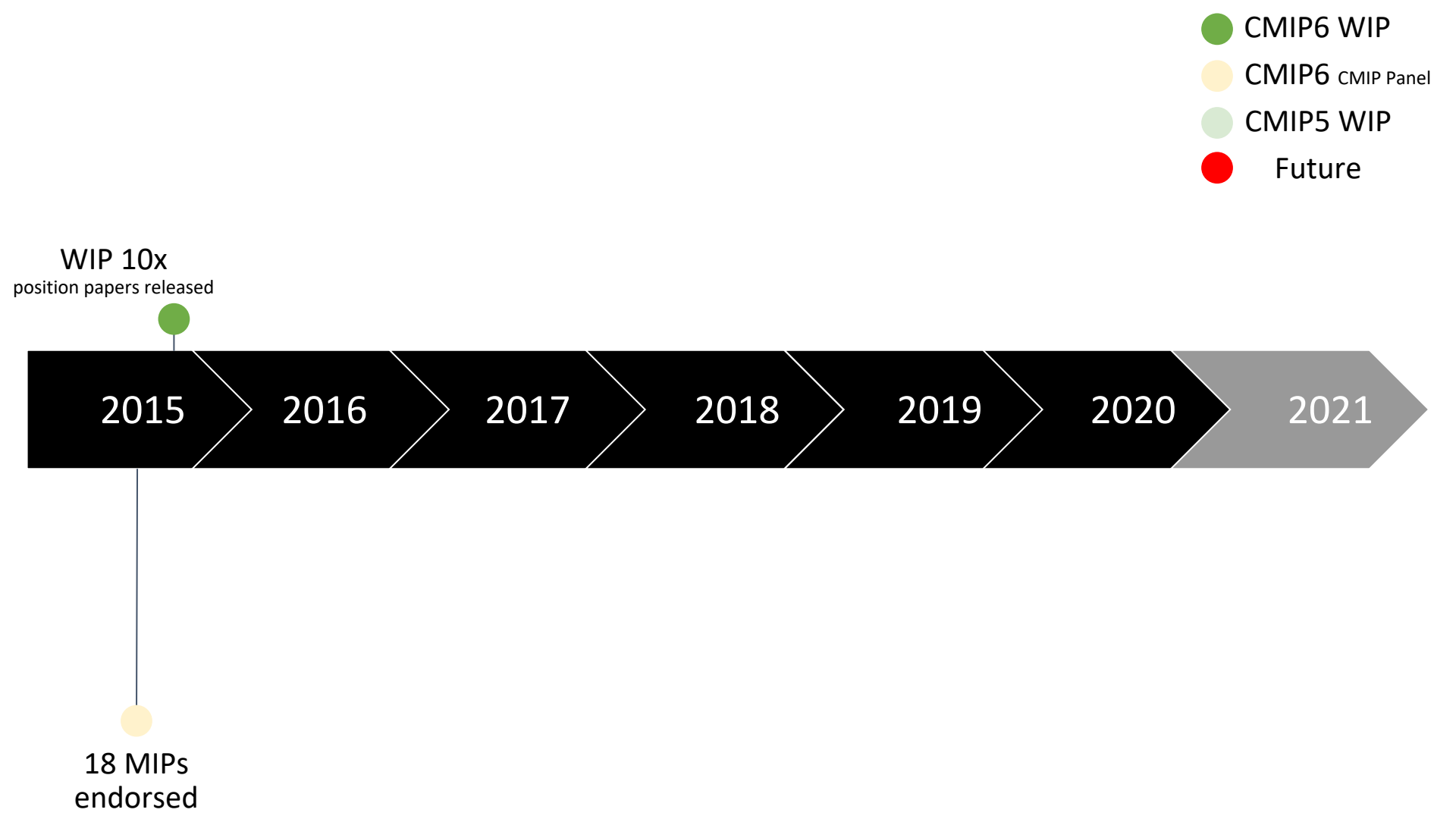
Timeline of CMIP6

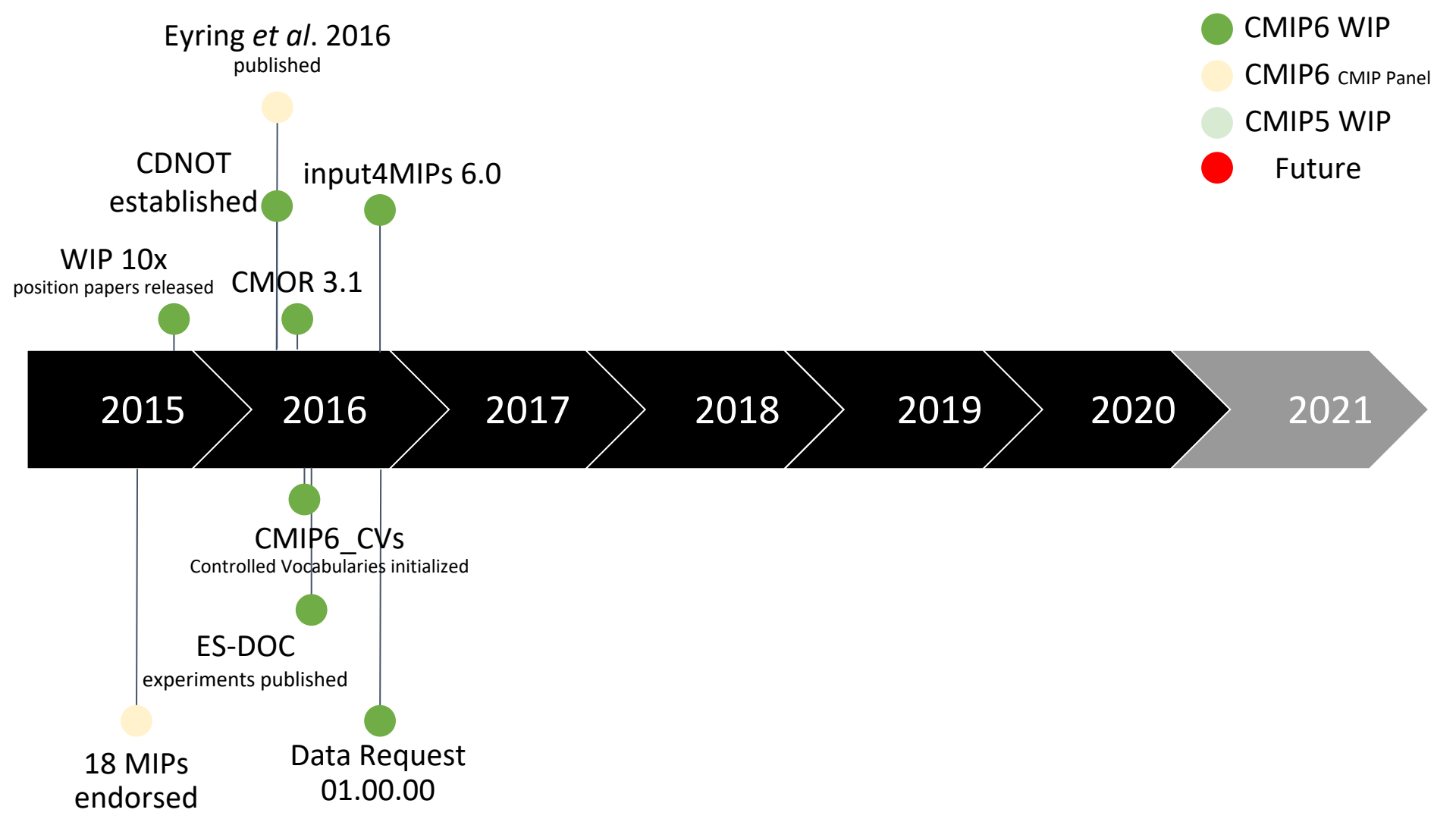
CMIP6 WIP

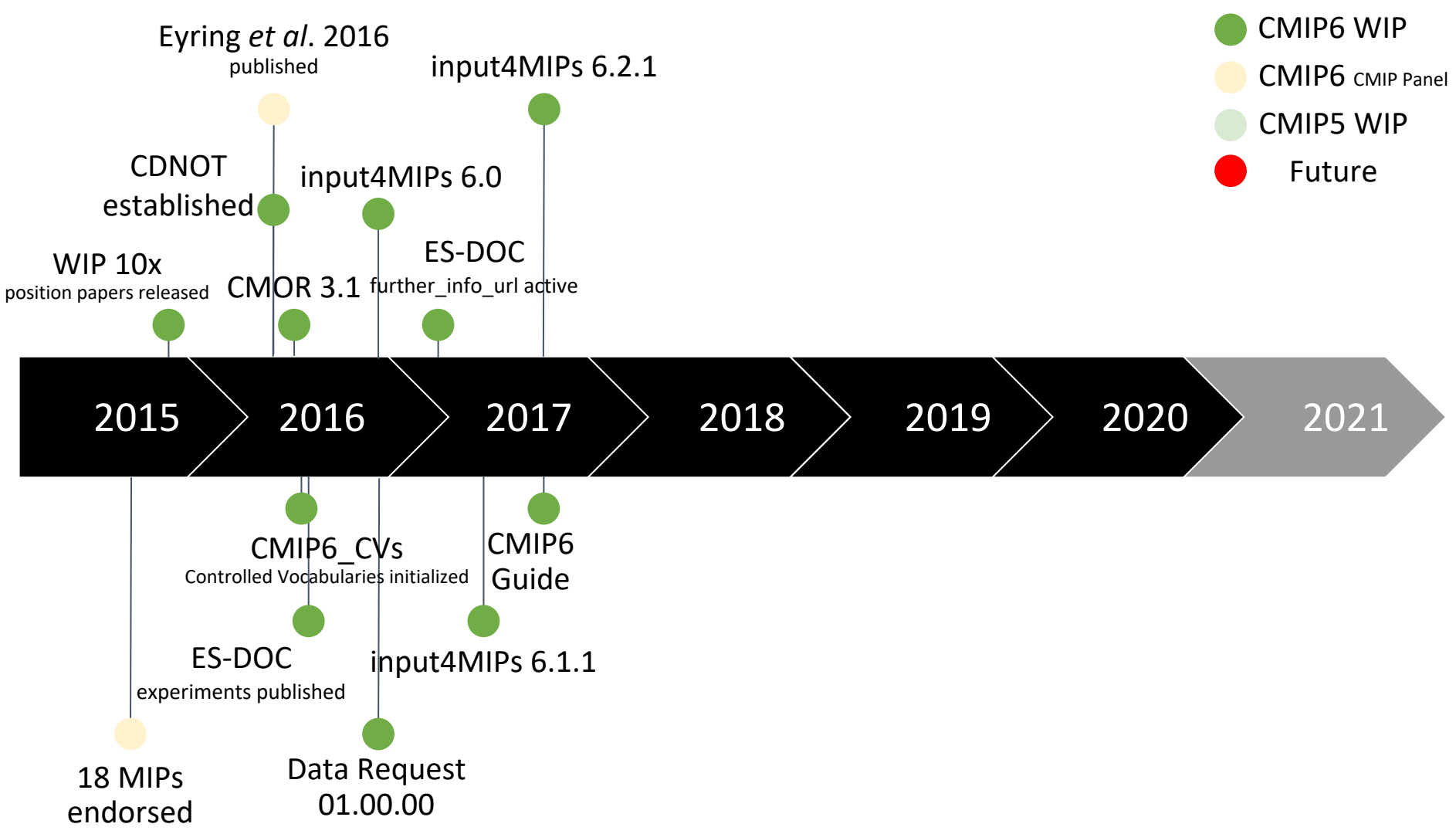
CMIP6 CMIP Panel

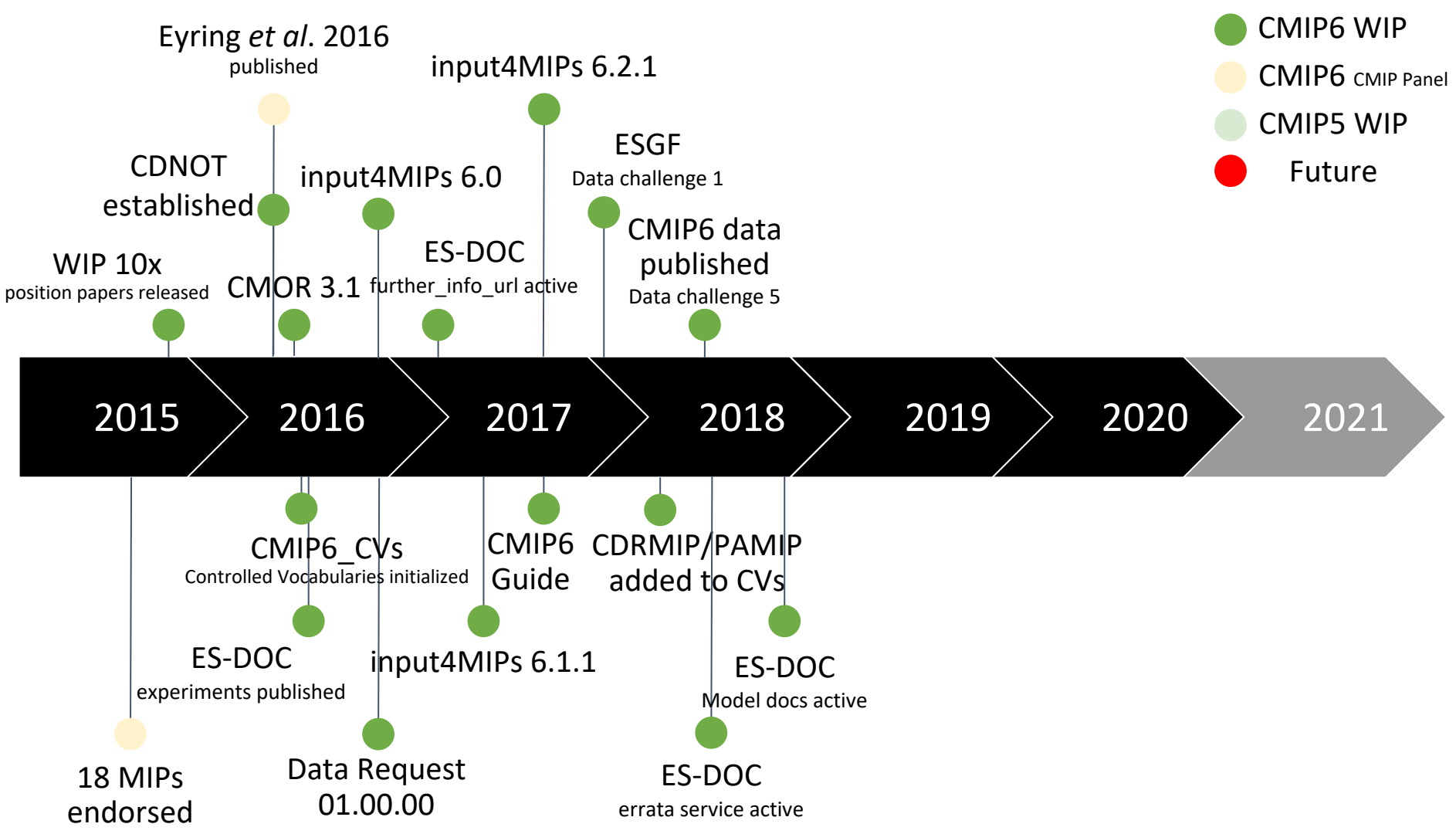
CMIP5 WIP

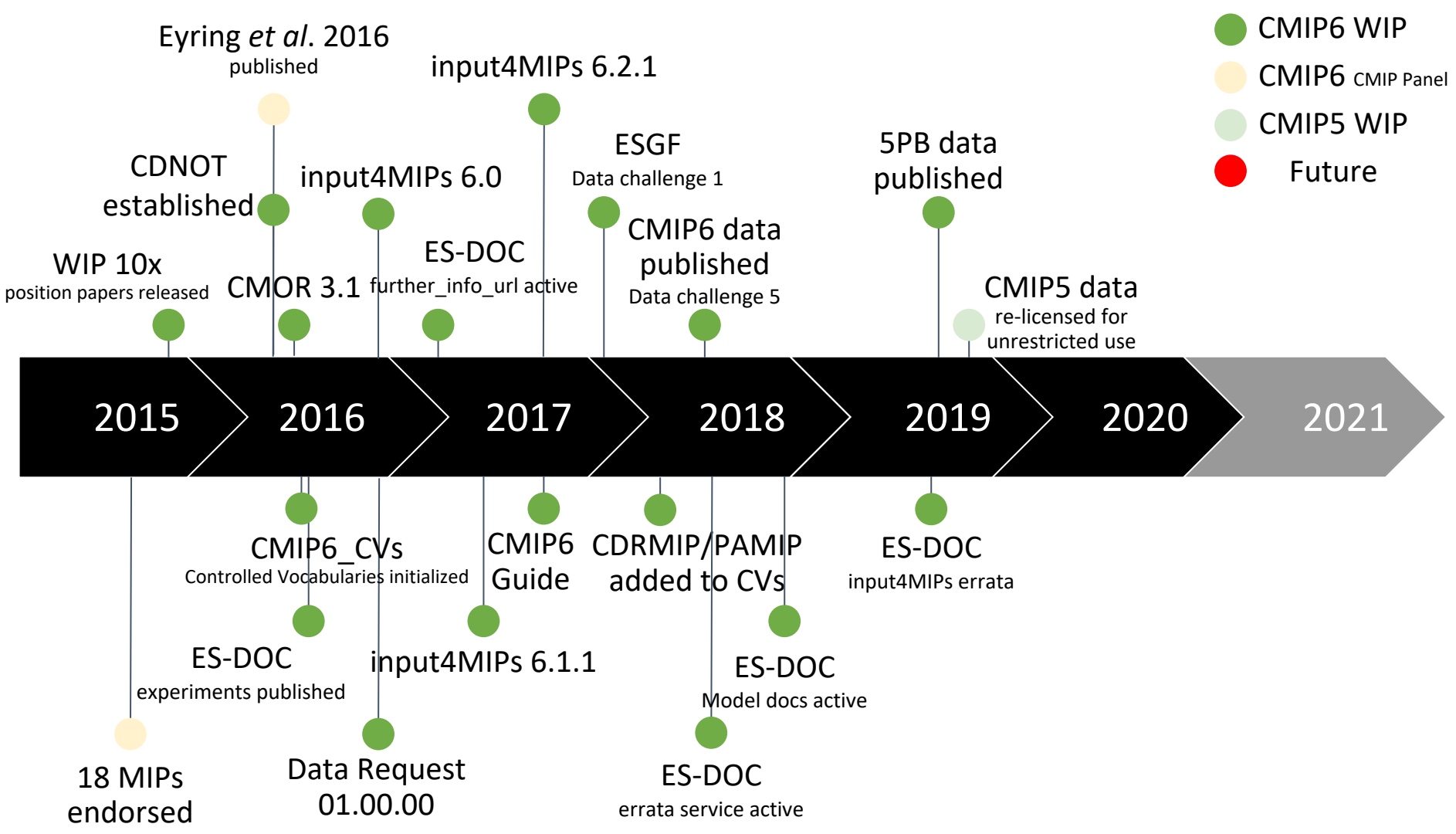
Future

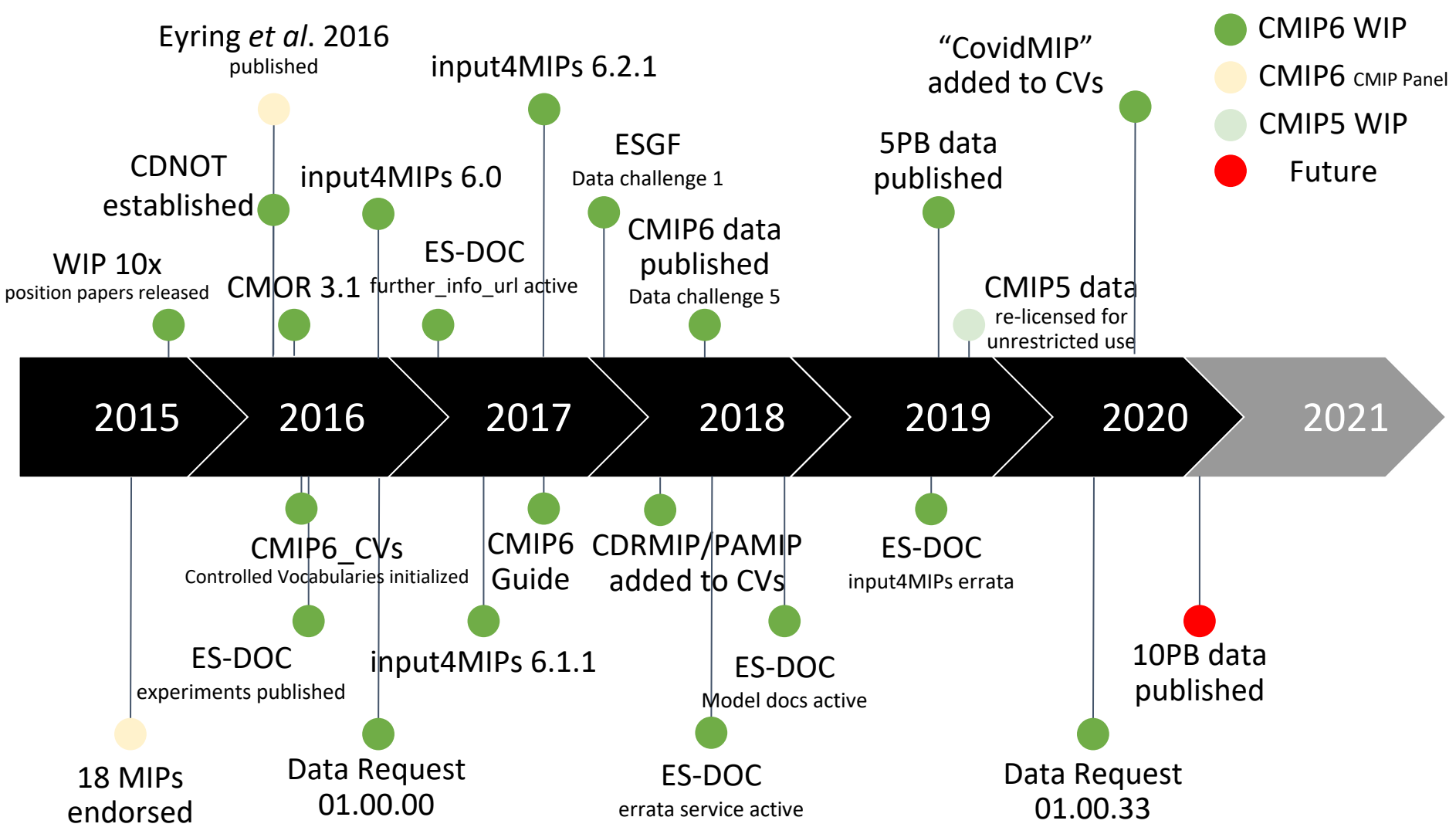

CMIP6 WIP
CMIP6 CMIP Panel
CMIP5 WIP
Future
WIP 10x
position papers released
2015
2016
2017
2018
2019
2020
2021
18 MIPs
endorsed


Eyring *et al*. 2016
published
CDNOT
established
input4MIPs 6.0
WIP 10x
position papers released
CMOR 3.1
2015
2016
2017
2018
2019
2020
2021
CMIP6_CVs
Controlled Vocabularies initialized
ES-DOC
experiments published
18 MIPs
endorsed
Data Request
01.00.00
CMIP6 WIP
CMIP6 CMIP Panel
CMIP5 WIP
Future

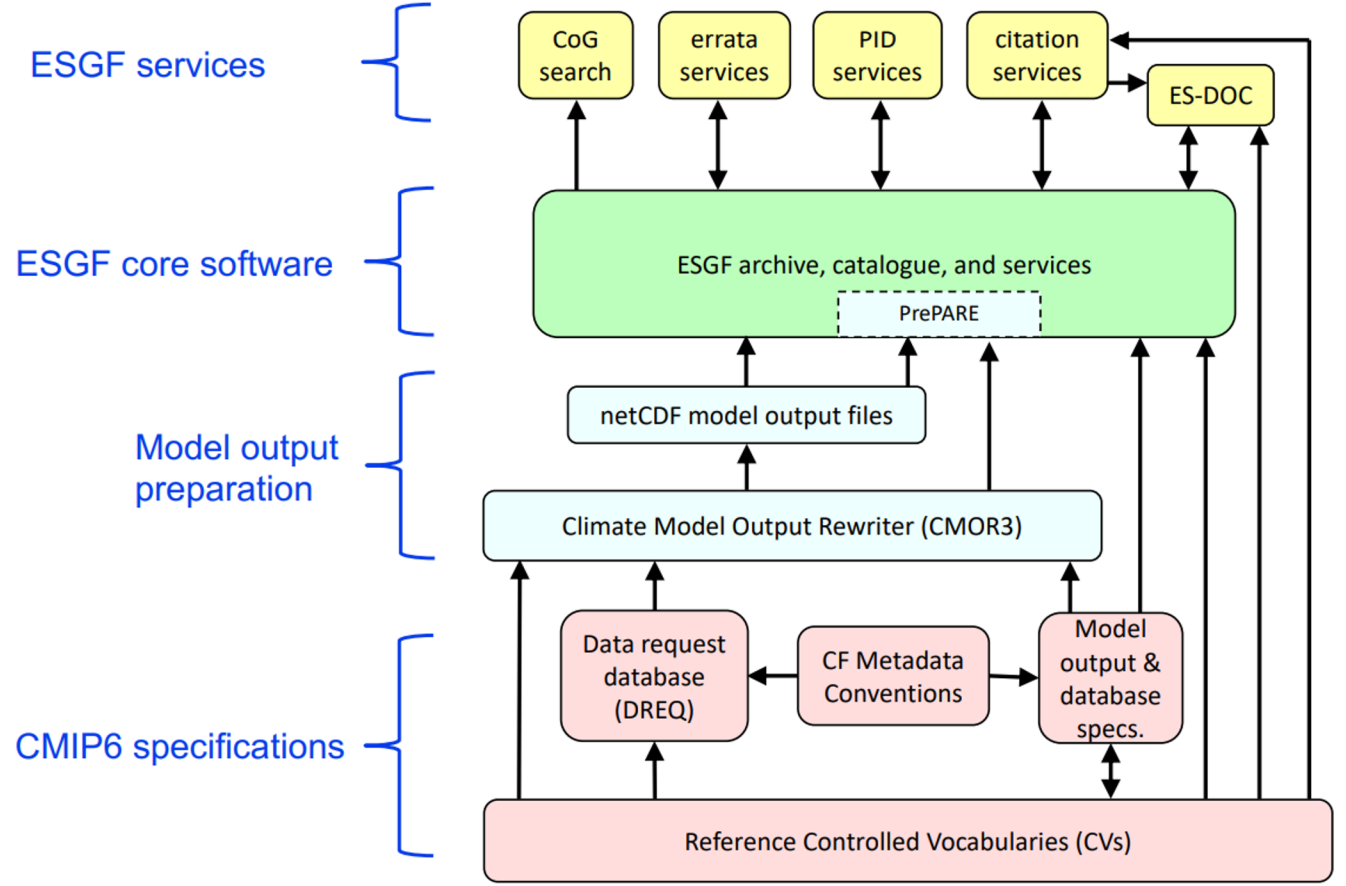

CMIP6 WIP
CMIP6 CMIP Panel
CMIP5 WIP
Future
Eyring *et al*. 2016
published
input4MIPs 6.2.1
CDNOT
established
input4MIPs 6.0
WIP 10x
position papers released
CMOR 3.1
ES-DOC
further_info_url active
2015
2016
2017
2018
2019
2020
2021
CMIP6_CVs
Controlled Vocabularies initialized
CMIP6
Guide
ES-DOC
experiments published
input4MIPs 6.1.1
18 MIPs
endorsed
Data Request
01.00.00


CMIP6 WIP
CMIP6 CMIP Panel
CMIP5 WIP
Future
Eyring *et al*. 2016
published
input4MIPs 6.2.1
CDNOT
established
input4MIPs 6.0
ESGF
Data challenge 1
WIP 10x
position papers released
CMOR 3.1
ES-DOC
further_info_url active
CMIP6 data
published
Data challenge 5
2015
2016
2017
2018
2019
2020
2021
CMIP6_CVs
Controlled Vocabularies initialized
CMIP6
Guide
CDRMIP/PAMIP
added to CVs
ES-DOC
experiments published
input4MIPs 6.1.1
ES-DOC
Model docs active
18 MIPs
endorsed
Data Request
01.00.00
ES-DOC
errata service active


CMIP6 WIP
CMIP6 CMIP Panel
CMIP5 WIP
Future
Eyring *et al*. 2016
published
input4MIPs 6.2.1
CDNOT
established
input4MIPs 6.0
ESGF
Data challenge 1
5PB data
published
WIP 10x
position papers released
CMOR 3.1
ES-DOC
further_info_url active
CMIP6 data
published
Data challenge 5
CMIP5 data
re-licensed for
unrestricted use
2015
2016
2017
2018
2019
2020
2021
CMIP6_CVs
Controlled Vocabularies initialized
CMIP6
Guide
CDRMIP/PAMIP
added to CVs
ES-DOC
input4MIPs errata
ES-DOC
experiments published
input4MIPs 6.1.1
ES-DOC
Model docs active
18 MIPs
endorsed
Data Request
01.00.00
ES-DOC
errata service active


CMIP6 WIP
CMIP6 CMIP Panel
CMIP5 WIP
Future
2015
2016
2017
2018
2019
2020
2021
Eyring *et al*. 2016
published
CDNOT
established
WIP 10x
position papers released
CMOR 3.1
input4MIPs 6.0
ES-DOC
further_info_url active
input4MIPs 6.2.1
ESGF
Data challenge 1
CMIP6 data
published
Data challenge 5
5PB data
published
"CovidMIP"
added to CVs
CMIP5 data
re-licensed for
unrestricted use
CMIP6_CVs
Controlled Vocabularies initialized
CMIP6
Guide
CDRMIP/PAMIP
added to CVs
ES-DOC
input4MIPs errata
ES-DOC
experiments published
input4MIPs 6.1.1
ES-DOC
Model docs active
10PB data
published
18 MIPs
endorsed
Data Request
01.00.00
ES-DOC
errata service active
Data Request
01.00.33

Infrastructure components and dependencies



CMIP infrastructure support

- PCMDI
  - DOE has provided 31-years of *MIP support
- ESGF
  - Originated by U.S. DOE
  - More recent major contributions from numerous others
- IS-ENES
  - European contribution to ESGF & CMIP infrastructure
- Numerous other projects and institutions, including DKRZ, IPSL, CEDA, ES-DOC, NASA, NOAA, …
- 31+ ESGF nodes and 52 Modeling institutions around the world representing 26 countries ...

ESGF Earth System Grid Federation

PCMDI

Lawrence Livermore National Laboratory

DEPARTMENT OF ENERGY · UNITED STATES OF AMERICA

is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

Sciences de l'environnement — Institut Pierre Simon Laplace

es-doc
Earth System Documentation

DKRZ

NOAA
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION · U.S. DEPARTMENT OF COMMERCE

NASA

ESGF Preparation

- In addition to software/hardware development/configuration
- Data challenges 1-5
  - Tested infrastructure to ensure a smooth/seamless roll out of CMIP6 for users
- Plans began at ESGF Face-to-Face (F2F) December 2017
- Began January 2018
- Finished (challenge 5) with first CMIP6 data published June 2018

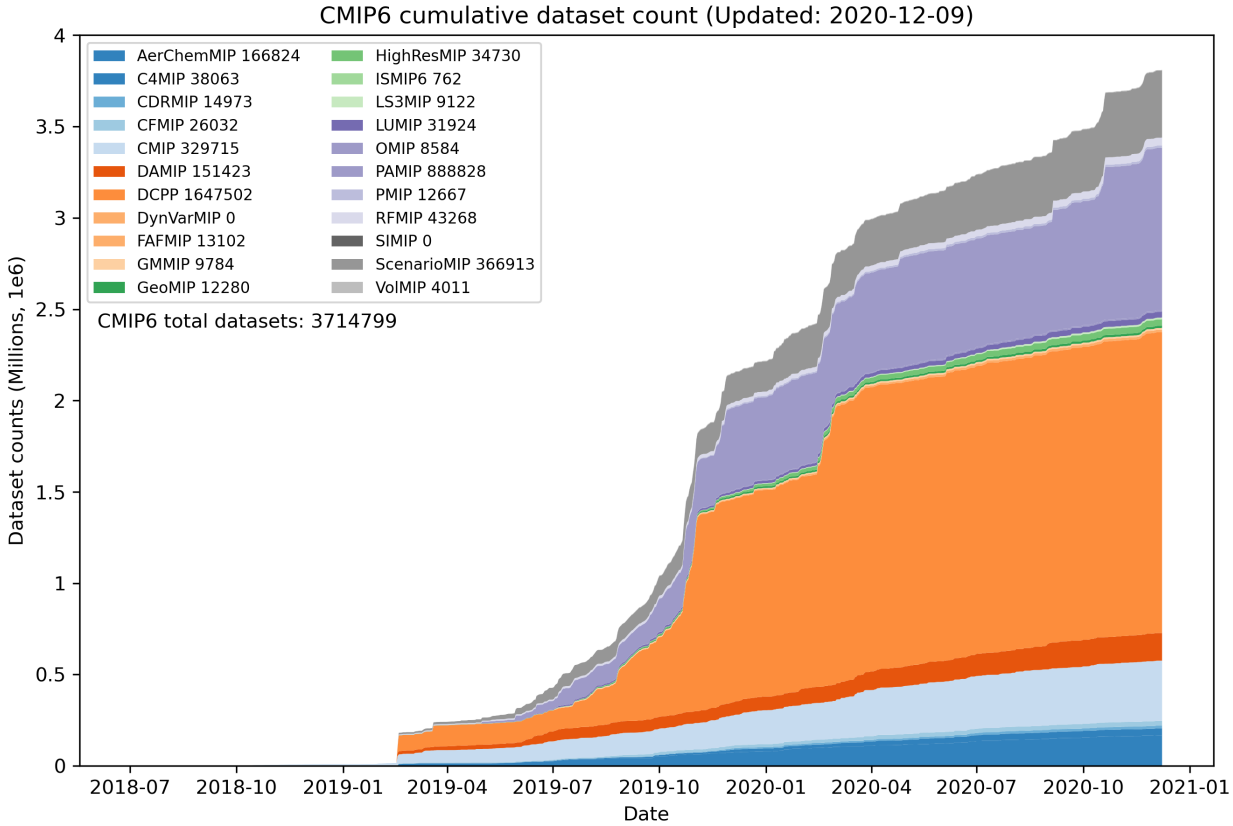
Tasks 1-16 vs Challenges 1-5	1	2	3	4	5
1. Install (or update) the ESGF software stack	✔	✔	✔	✔	✔
2. Run quality control on primary data	✔	✔	✔	✔	✔
3. Publish primary data	✔	✔	✔	✔	✔
4. Publish replica data	✔	✔	✔	✔	✔
5. Verify search and download is functional	✔	✔	✔	✔	✔
6. Register data with PID assignment service		✔	✔	✔	✔
7. Verify Citation Service registers DOIs for published data		✔	✔	✔	✔
8. Populate "further_info_url" through ES-DOC scanning		✔	✔	✔	✔
9. Replicate published data			✔	✔	✔
10. Apply the "test suite"			✔	✔	✔
11. Verify the metrics collection for the dashboard			✔	✔	✔
12. Register an errata with the Errata Service			✔	✔	✔
13. Retract a version of the data				✔	✔
14. Publish a new version of the data				✔	✔
15. Ensure homogeneity across ESGF CoG sites				✔	✔
16. Move testing to production environment				✔	✔

Table 1. Table of the data challenge tasks and the challenge at which they were implemented. Note the grey tick of Task 11 in Challenge 5 indicates that the task was optional.

Petrie *et al.*, 2020, GMD https://doi.org/10.5194/gmd-2020-153

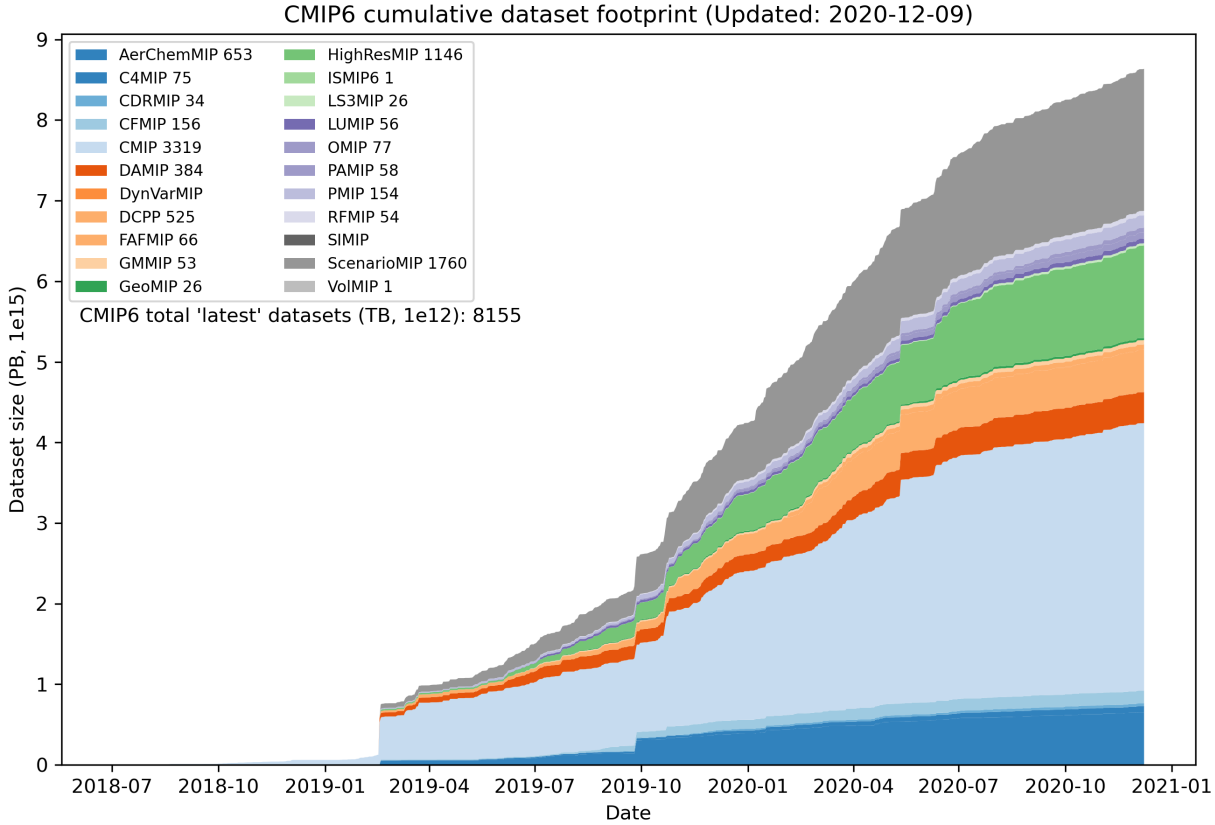
ESGF Published data

- Over 3.7 million datasets on ESGF across all CMIP6 activities/MIPs
- Delivery has been seamless – thanks to data challenges and ESGF stability testing
- Datasets - unique variable collections per experiment RIPF



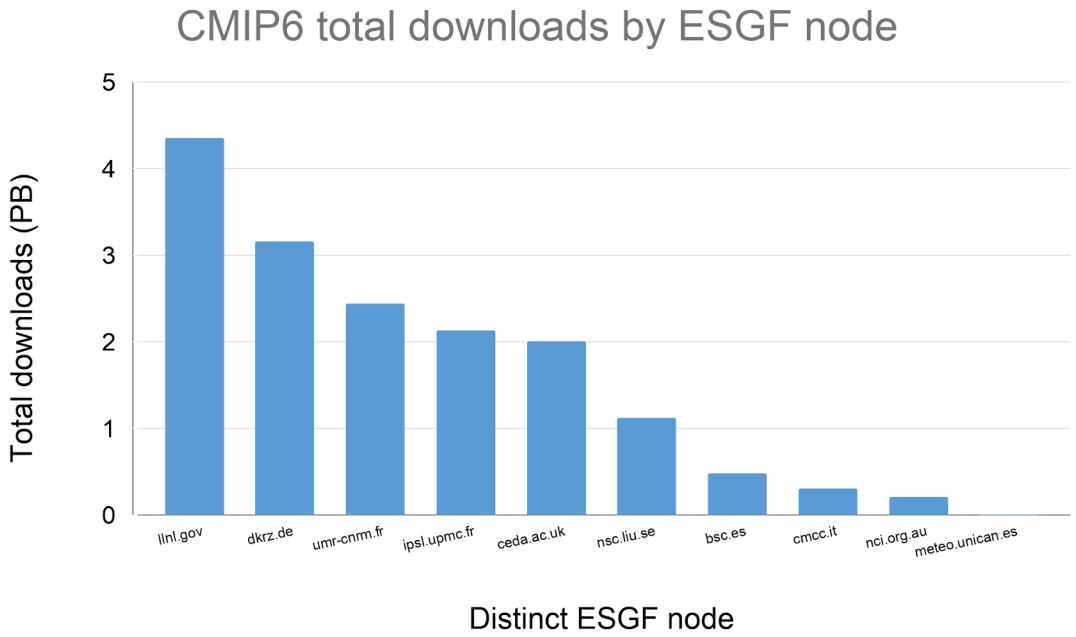
ESGF Published data

- Over 3.7 million datasets on ESGF across all CMIP6 activities/MIPs
- Delivery has been seamless – thanks to data challenges and ESGF stability testing
- Datasets - unique variable collections per experiment RIPF
- Footprint – storage units in PBs



ESGF publication and replication

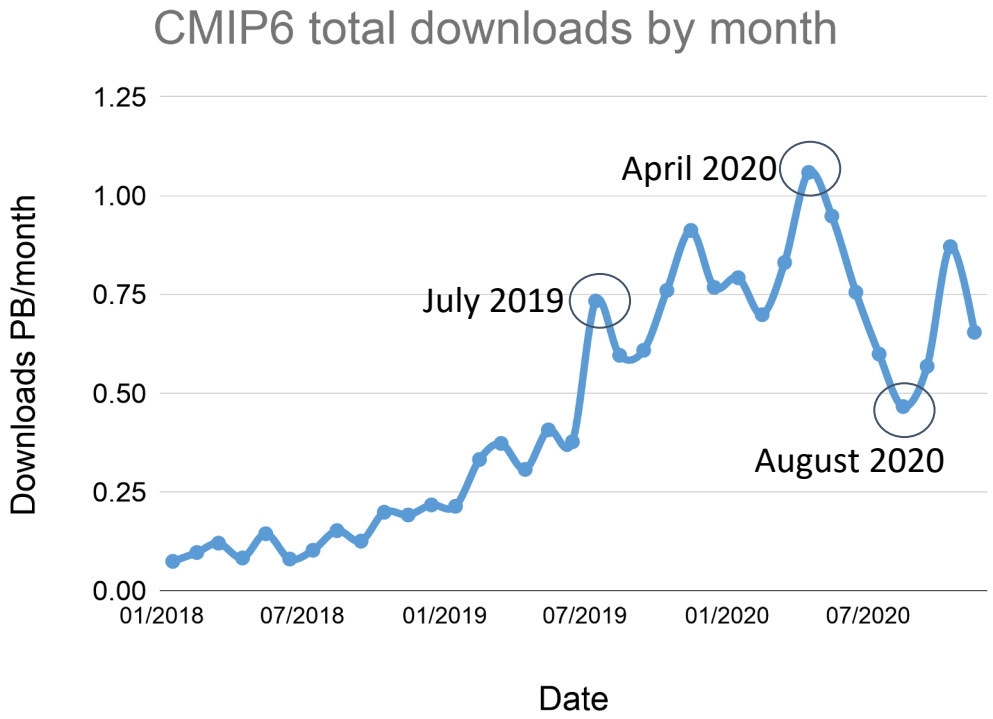
- 17.7 PB CMIP6 data available including ~10 PB of unique and ~8 PB of replicated data
- 16.2 PB of total CMIP6 downloads (to November 2020)
- LLNL delivered 27% of downloads to date
- DKRZ delivered 19%
- CNRM delivered 15%
- IPSL delivered 13%
- CEDA delivered 12%



http://esgf-ui.cmcc.it/esgf-dashboard-ui/federated-view.html

Downloaded data

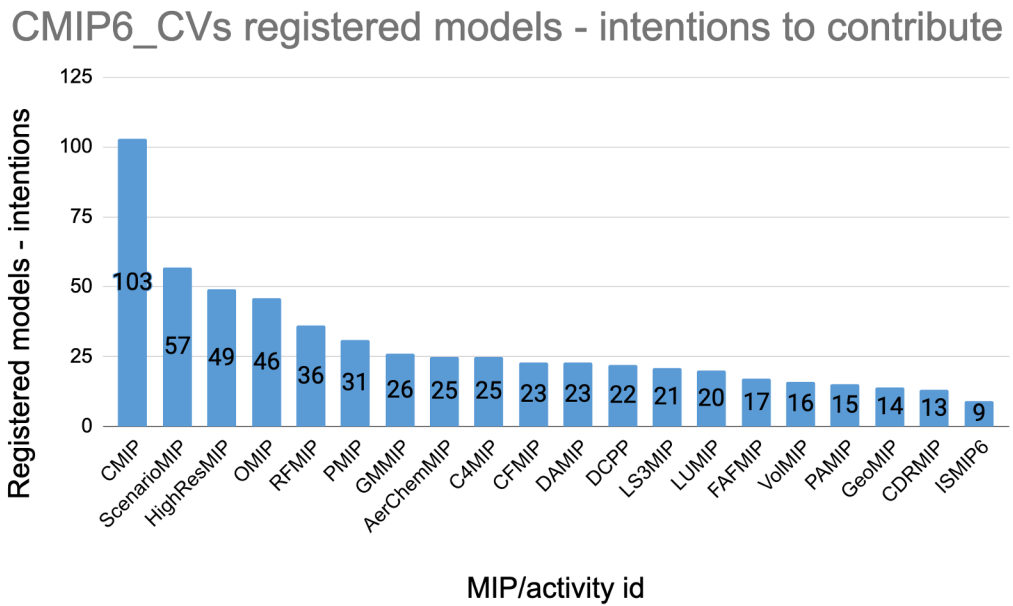
- Over 16 PB of recorded downloads (to November 2020)
- Use of secondary or local “dark” archives (e.g. Pangeo) likely means amount of data globally accessed is significantly larger
- Interesting to see COVID activity impacts



http://esgf-ui.cmcc.it/esgf-dashboard-ui/federated-view.html

CMIP6 controlled vocabulary

- 137 models registered with CMIP6 CVs
- Each model involved in 6* activities on average
- Experiments grown from ~280 to 322 including six recent “CovidMIP” experiments added to DAMIP
- Added CDRMIP and PAMIP in March 2018



https://wcrp-cmip.github.io/CMIP6_CVs/docs/CMIP6_experiment_id.html
https://wcrp-cmip.github.io/CMIP6_CVs/docs/CMIP6_institution_id.html
https://wcrp-cmip.github.io/CMIP6_CVs/docs/CMIP6_source_id.html

https://github.com/WCRP-CMIP/CMIP6_CVs

Data Request & CMOR

Data Request

- Documented in GMD
- Many releases, responsive to MIP tweaks
- Latest version 01.00.33 published to include 6 “CovidMIP” experiments sponsored by DAMIP cloned from hist-aer experiment entries

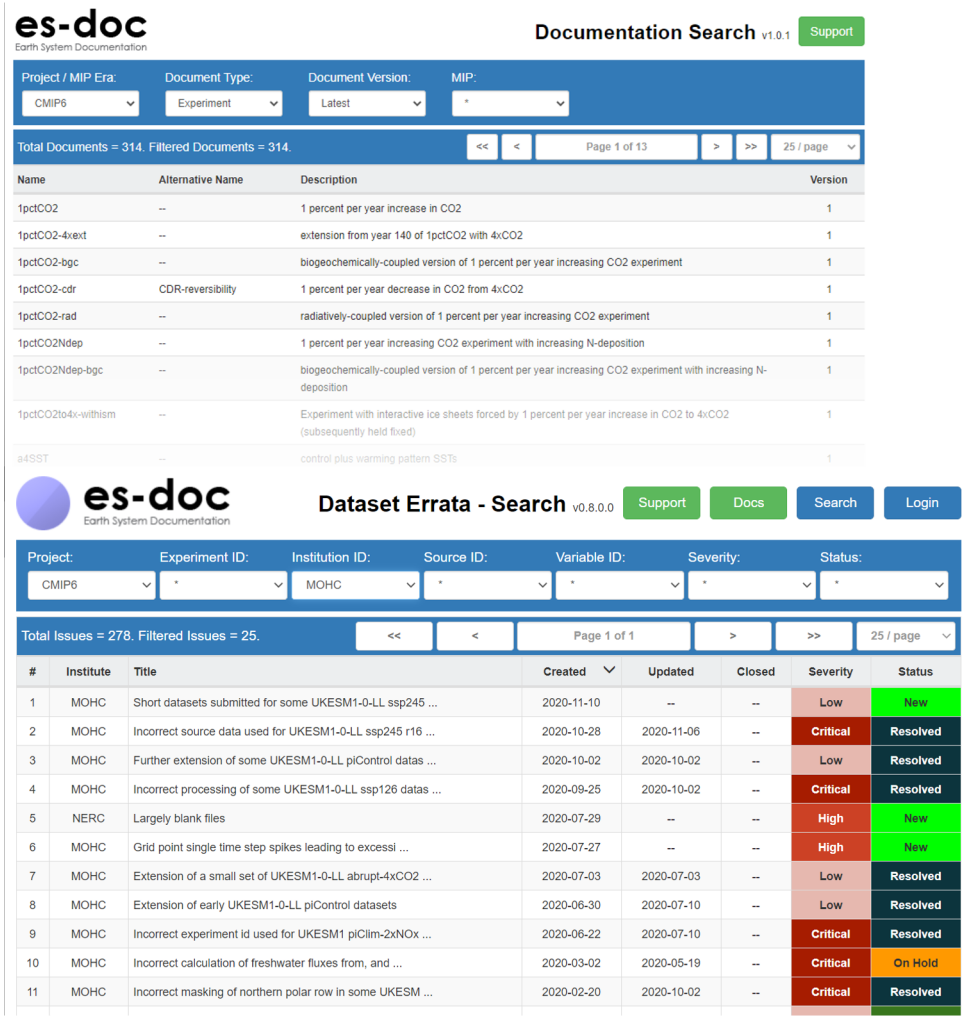
CMOR3

- Updates for python 3 compatibility
- MIP tables generated from data request and CVs when released
- Composite tables require lower investment for modeling group configuration and use

http://clipc-services.ceda.ac.uk/dreq/
https://github.com/PCMDI/cmor

ES-DOC

- Comprehensive documentation that expands upon CMIP6_CVs
- Allows modeling groups to comprehensively document forcing and model configuration per RIPF
- Errata system invaluable for recording dataset issues and extensions



https://search.es-doc.org/
https://errata.es-doc.org

Data Citation

Data Citation Service:

- Provide data DOIs on data collection granularities
- Provide information on data usage in papers
- Disseminate data citation information

Data Citation Status:

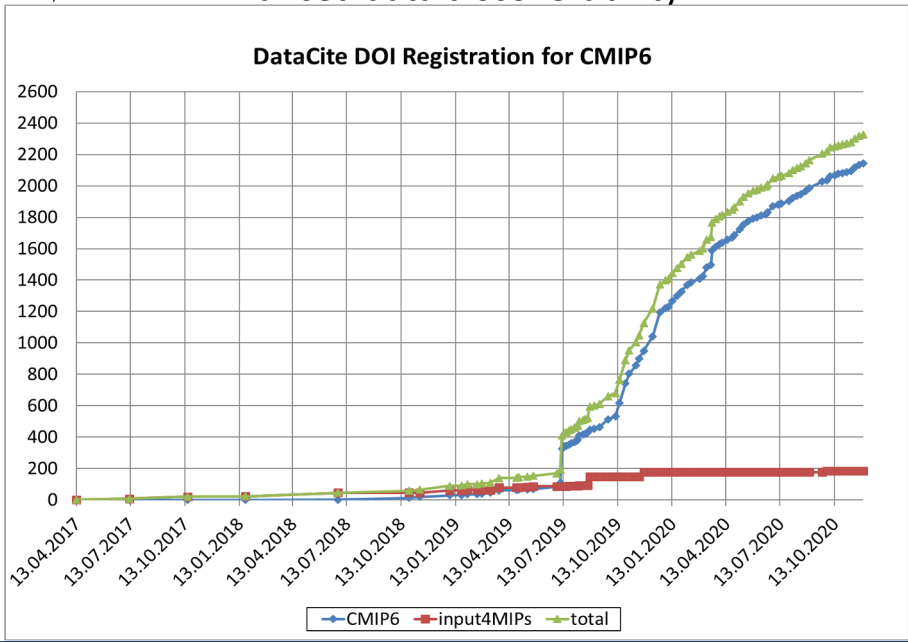
- 2153 CMIP6 DOIs have been registered, which corresponds to a coverage of 99%.
- 1-3 DOIs are registered per day on average.

Data Citation Usage Status:

- IPCC WG1 AR6 will include CMIP6 data references.
- 62 papers referencing CMIP6 data have been added.

Benefit:

- Data is citable in scientific publications.
- Receive credit for data creation
- Enhanced data discoverability



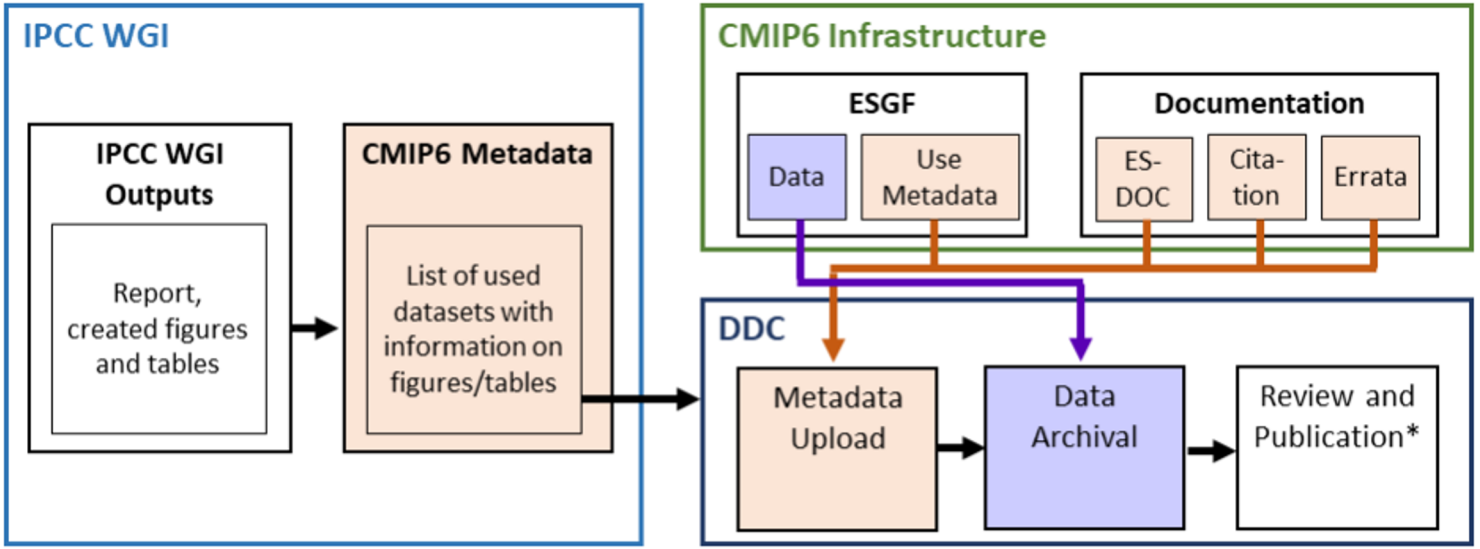
General Information: cmip6cite.wdc-climate.de
Available Data References: bit.ly/CMIP6_Citation_Search
Data Citation Statistics: bit.ly/CMIP6_DOI_Statistics

CC BY

Long-Term Data Archival in the IPCC DDC

IPCC Data Distribution Centre (DDC):

The DDC at DKRZ is the Reference Data Archive for the GCM data underpinning the IPCC Assessment Reports since the SAR. The DDC has the function to steward the data subset on the long-term complying to international data repository standards like CoreTrustSeal. The significance of data has increased in the AR6.



* DataCite publication and publication on IPCC webpages

CC BY

Lessons learned

- WIP is important as the operational phase of CMIPx occurs
  - Also important in planning as infrastructure needs (storage, compute, user-base) begin to coalesce
- Funding is not operational - currently “research funding”
  - Many single points of failures
  - Dependence on single contributors with no backup
- Licensing issues have been raised by the IPCC/AR6
  - Inconsistent licensing across institutions causes problems for downstream use
  - IPCC proposes to publish AR6 and associated data (notably, the ”Atlas") under an attribution license *without the ShareAlike restriction*
- Dialogue between panels (CMIP and WIP) essential to ensure clean delivery across complex contributor network

Lessons learned (cnt'd)

- As complexity of CMIP experiments has expanded, maintaining consistency is an ongoing problem
  - E.g. historical experiments can be spawned from piControl AND past1000 experiments, but their lineage (and difference) may not be clear to a user (CMIP6_CVs#957)
  - E.g. past2k and past1000 overlap same experiment duplicated across the temporal extent of the experiment (CMIP6_CVs#979)
- In future phases an iterative loop of propose, review, and revise is required – likely requiring numerous iterations
- This becomes more important as the interlinkages and dependencies between MIPs/experiments continue

Paul J. Durack
durack1@llnl.gov

Matthew Mizielinski
matthew.mizielinski@metoffice.gov.uk

Karl E. Taylor
taylor13@llnl.gov

Work completed by the PCMDI project is funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Regional and Global Model Analysis Program

Lawrence Livermore National Laboratory

Disclaimer
This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.