A24D-08 Rapid Evaluation Framework for the Assessment Fast Track

Forrest M. Hoffman
Oak Ridge National Laboratory (ORNL)

On behalf of the REF Delivery Team and the CMIP7 Model Benchmarking Task Team

AGU 2025
16 December 2025

With thanks to everyone for making the REF possible:

- Acknowledgement of funders

esa

U.S. DEPARTMENT of ENERGY

Being delivered by:

Climate Resource

OAK RIDGE National Laboratory

DLR

netherlands eScience center

Lawrence Livermore National Laboratory

ESMValTool
Earth System Model Evaluation Tool

ILAMB

PMP

CMEC

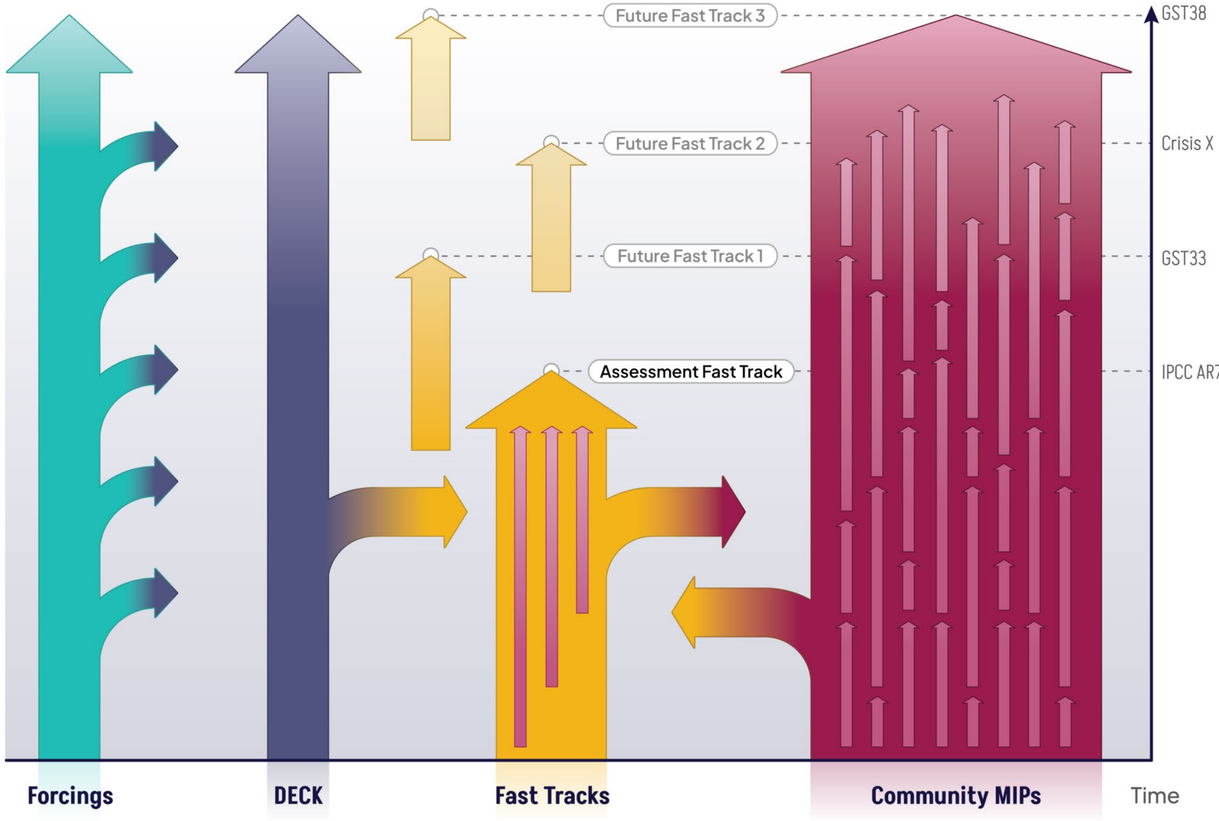
- Acknowledgement of the WCRP community who have responded to our surveys, attended dedicated events and provided constructive feedback.

CMIP7 Model Benchmarking Task Team

- Prerequisite for reliable information from Earth system models: understand model capabilities and limitations.
- Essential to evaluate the models systematically and comprehensively with the best available observations and reanalysis data.
- Full integration of routine benchmarking and evaluation of the models into the CMIP publication workflow.
  - Technical challenges: high resolution, memory limits, unstructured grids.
  - Science challenges: innovative diagnostics, novel observations, ML based analysis

CMIP Phase 7

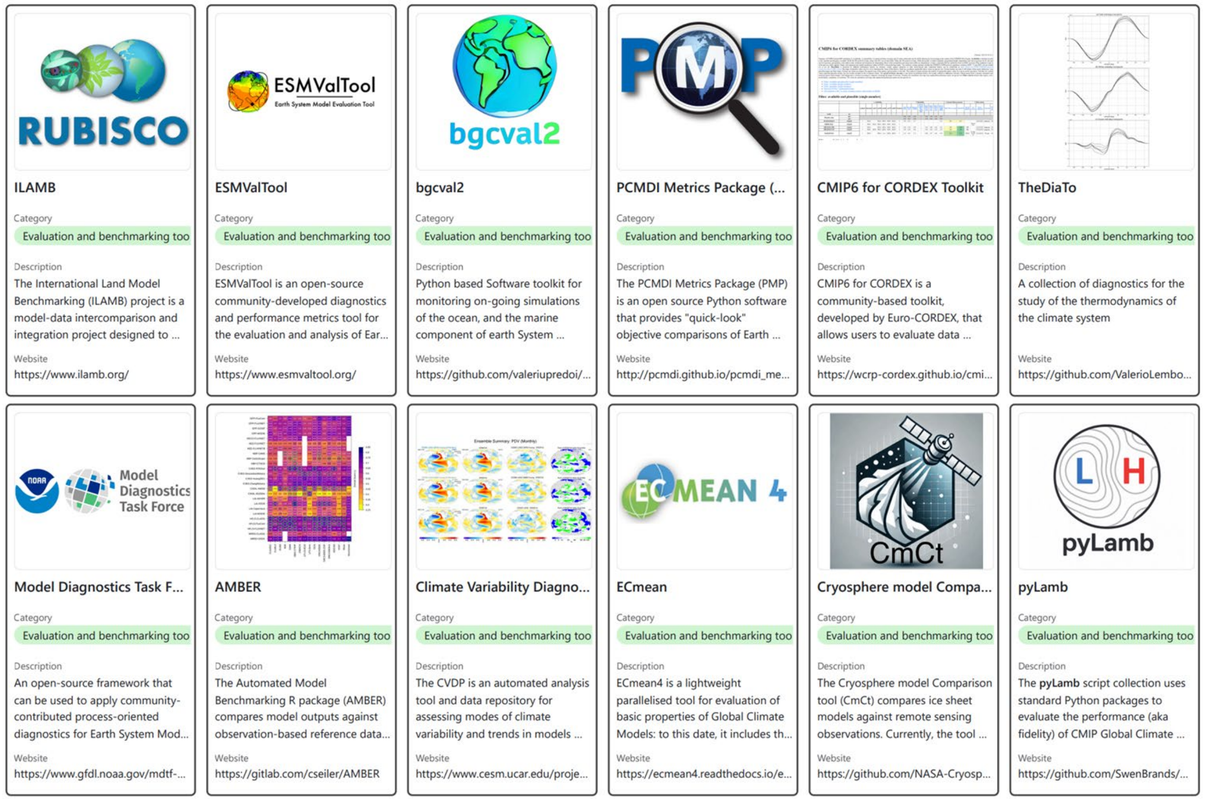
- CMIP is evolving into a more continuous approach.
- Now includes small targeted “fast track” experiment sets, in addition to the DECK and the growing number of Community MIPs.
- The first fast track, “Assessment Fast Track” will respond to the needs of national and international assessments.



https://wcrp-cmip.org/cmip-phases/cmip7/

Benchmarking and evaluation tools

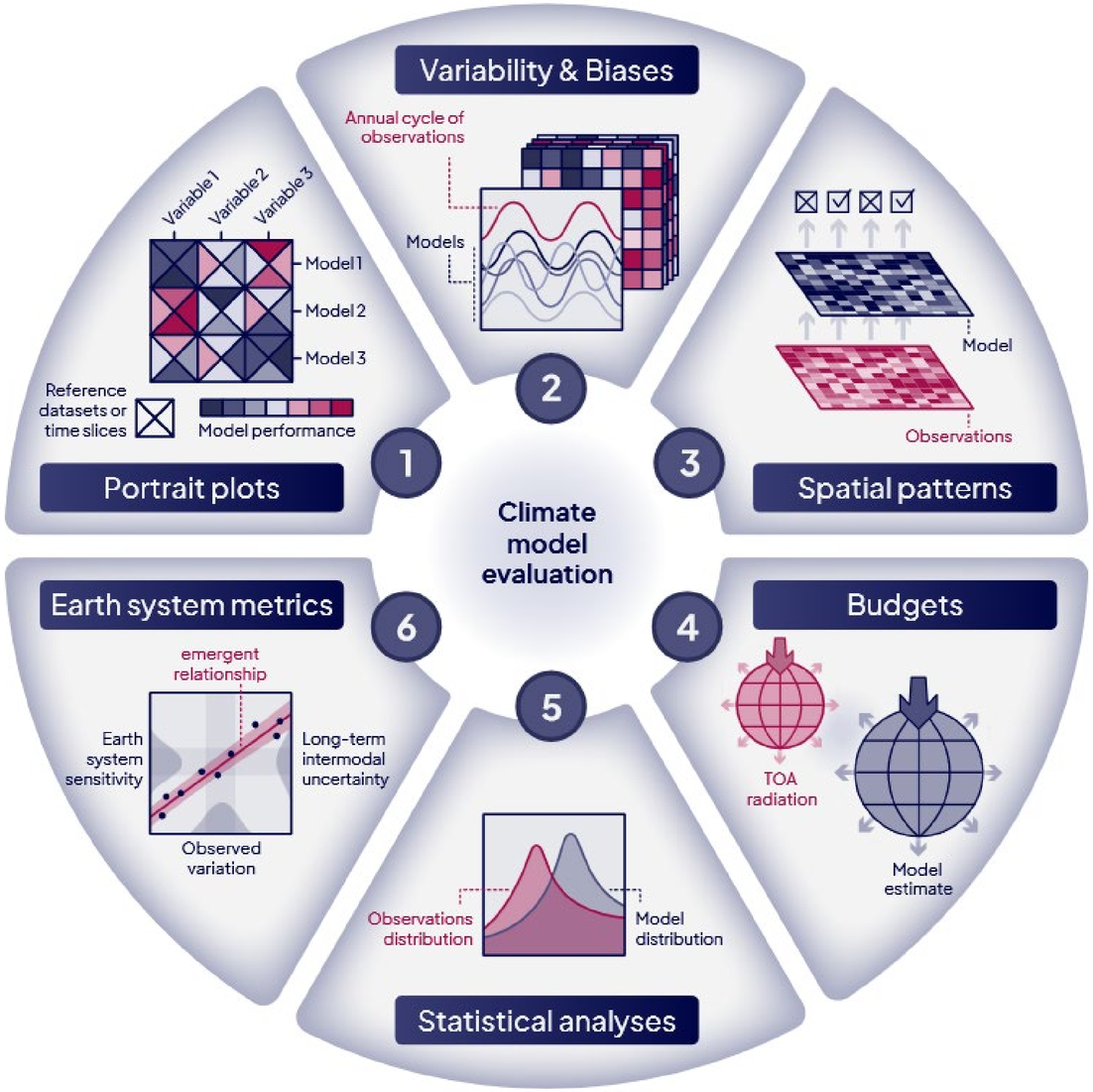
CMIP have created a community database of tools that can be used for analyses with Earth system data.



- Open-source
- Useable for CMIP data analyses
- 12 tools in the database so far
- Add your tool today!

Available at: https://wcrp-cmip.org/tools/model-benchmarking-and-evaluation-tools/

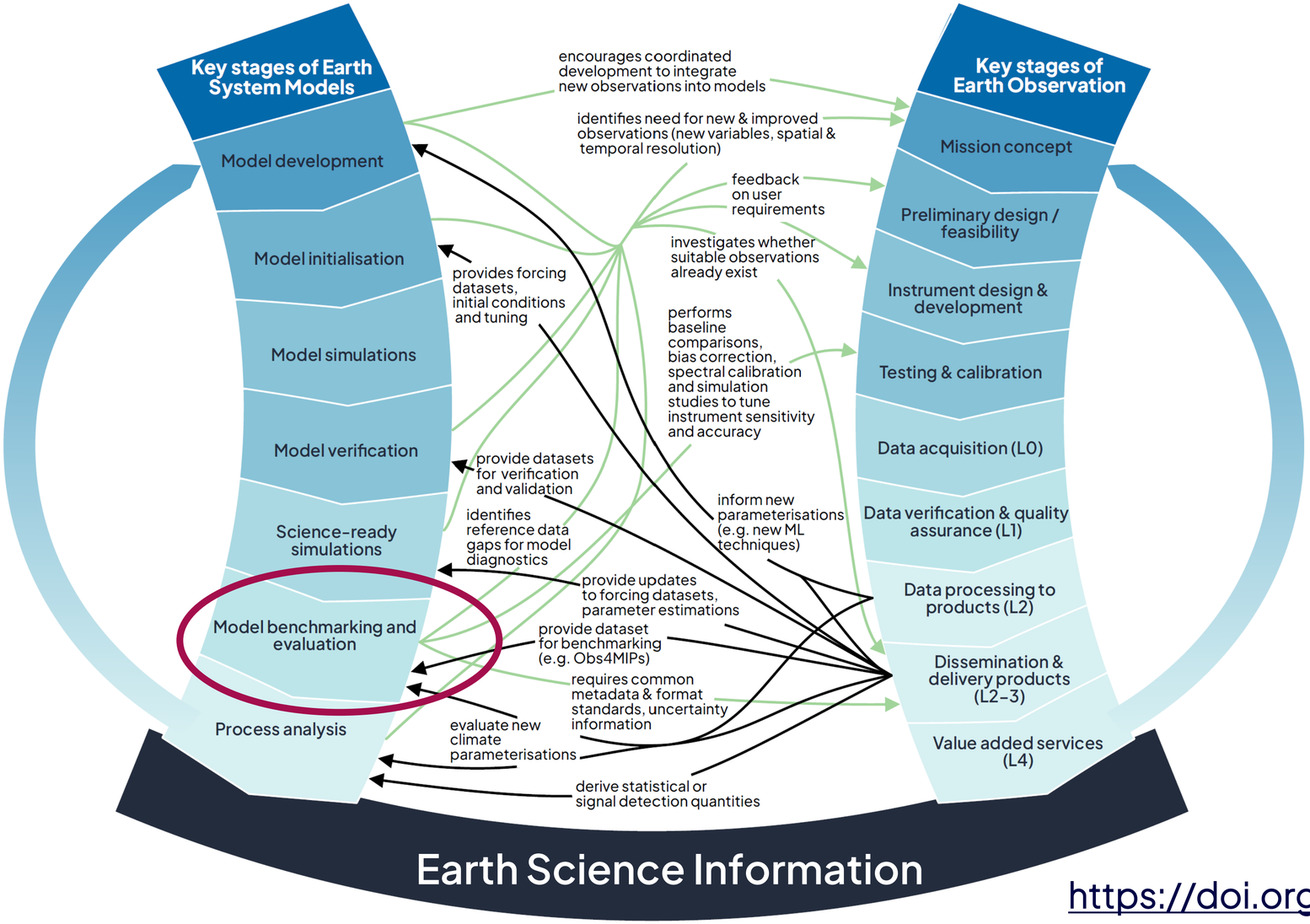
Benchmarking and evaluation approaches



- Six general approaches, grouped according to their underlying evaluation principles
- Portrait plots, Variability & Biases, Spatial patterns, Budgets, Statistical analyses, Earth System metrics
- Most of them can be applied regionally or globally

DOI: 10.5281/zenodo.13934006

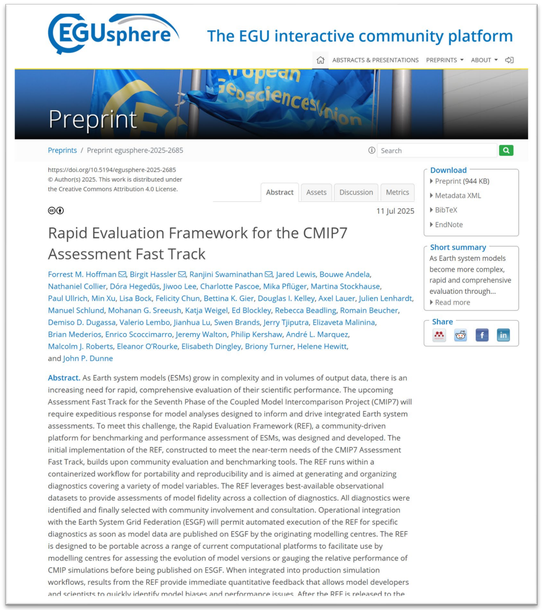
Observations for model evaluation



https://doi.org/10.5281/zenodo.14886515

Overview of the Rapid Evaluation Framework

- **Vision:** A community owned evaluation framework, built upon, and compatible with, existing community evaluation packages that incorporates an application programming interface (API) for executing metrics generation from those community evaluation packages, across the globe.
- **Goal:** The Rapid Evaluation Framework (REF) is a complete end to end system providing a systematic and rapid performance assessment of the expected models participating in the CMIP Assessment Fast Track, supporting the next IPCC Assessment Report 7 (AR7) cycle.
- **Outcome:** The REF provides ability to fully integrate evaluation tools into the CMIP publication workflow, and their diagnostic outputs published alongside the model output on the Earth System Grid Federation (ESGF) through an easily accessible website. The REF serves as a starting point for deeper exploration of CMIP model data.



Hoffman *et al*. (2025) Rapid Evaluation Framework for the CMIP7 Assessment Fast Track, EGUsphere [preprint] DOI: 10.5194/egusphere-2025-2685

https://wcrp-cmip.org/cmip-phases/cmip7/rapid-evaluation-framework/

The Rapid Evaluation Framework (REF):

- Brings together analyses from different evaluation tools in a consistent manner, showcasing open-source software and data diversity.
- REF diagnostics give a broad overview of the spread across models in terms of calculated metrics.
- REF diagnostics can be useful to answer a specific research question, by examining diagnostic criteria related to model performance in key processes or regions relevant to the study.

We surveyed the community in 2024 and they told us they wanted the development of a **single community analysis tool** to streamline and unify the analysis process.

https://wcrp-cmip.org/cmip-phases/cmip7/rapid-evaluation-framework/

Co-creation of the REF -19 months to date!

The brief:

The Model Benchmarking task team was tasked by the CMIP panel to provide a systematic and rapid performance assessment of the expected models participating in CMIP7 with a set of new and informative diagnostics and performance metrics, ideally along with the model output and documentation.

Several rounds of consultation:

- Early scope and structure shared with CMIP panel, ESMO and the WIP for feedback.
- Modelling centres and observation providers asked via surveys and through drop-ins for their views.
- Hackathon enabled early beta testers to directly feedback to the delivery team.
- Beta release issued for open consultation with the wider CMIP and user community.

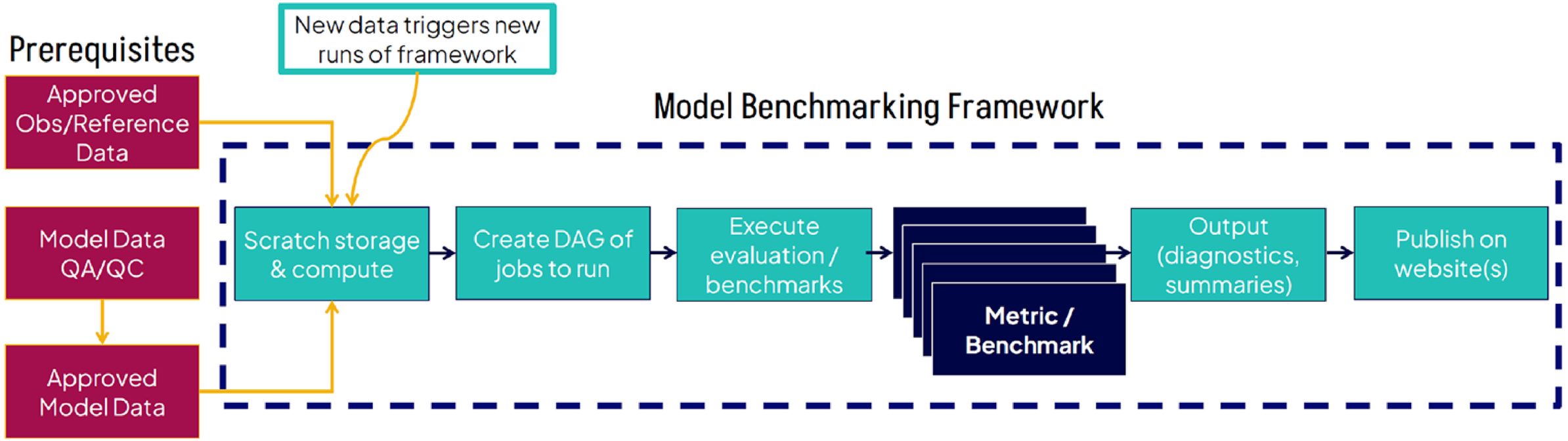
What we've done:

VERSION 1: In October 2025, CMIP6- & 6Plus-ready REF release

VERSION 2: In March 2026 there will be a second release, the CMIP7-ready REF.

https://wcrp-cmip.org/cmip-phases/cmip7/rapid-evaluation-framework/

Overview of the Rapid Evaluation Framework



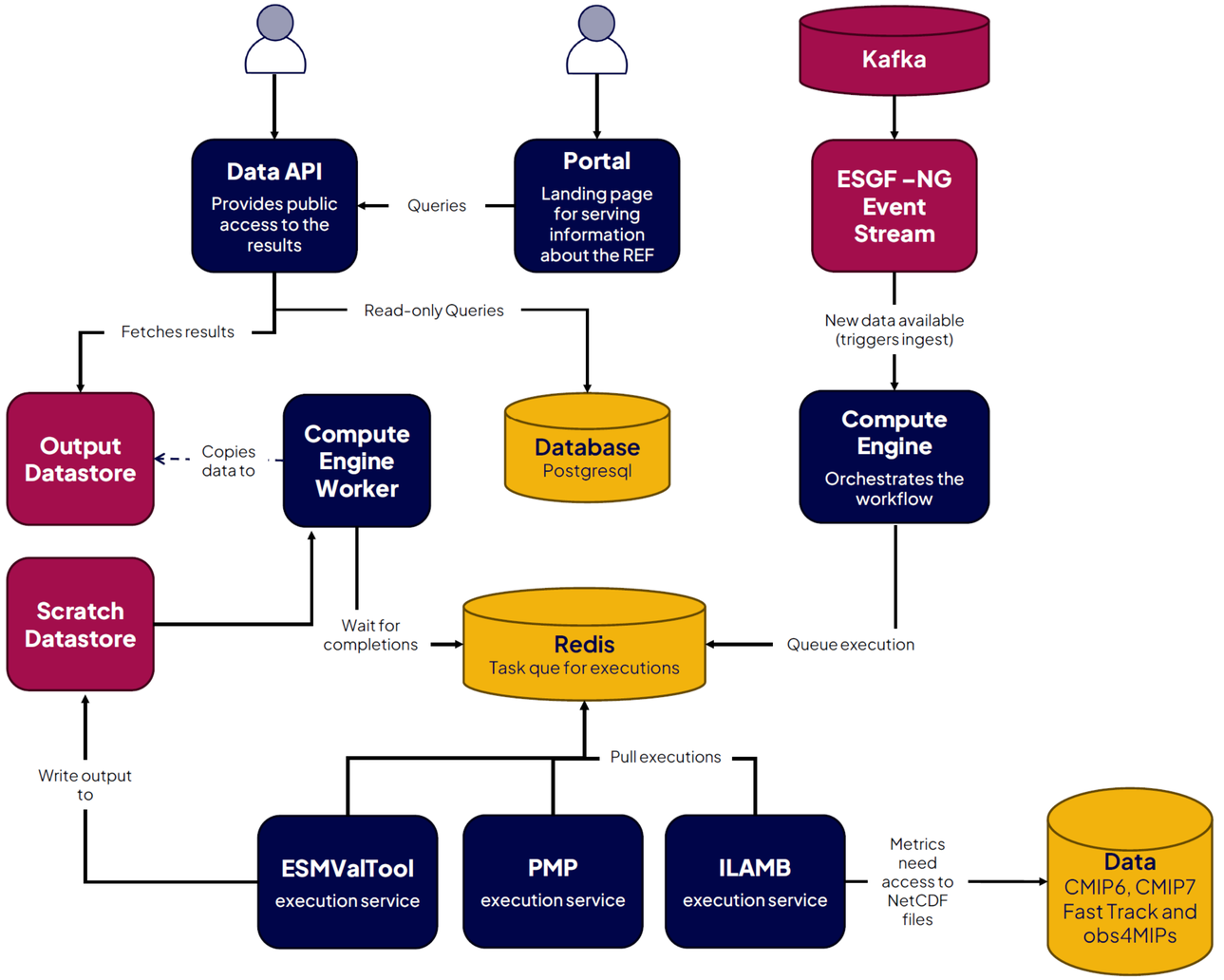
For the Version 2 release, it will only be run for experiments that have satisfied QA checks via the ESGF QA process.

https://wcrp-cmip.org/cmip-phases/cmip7/rapid-evaluation-framework/

REF deployment overview

Expansion for:

- CORDEX
- High resolution modelling
- IAMs
- Streamlining preparation of observation datasets for ingest by the REF
- Evaluation/quality control



External existing system

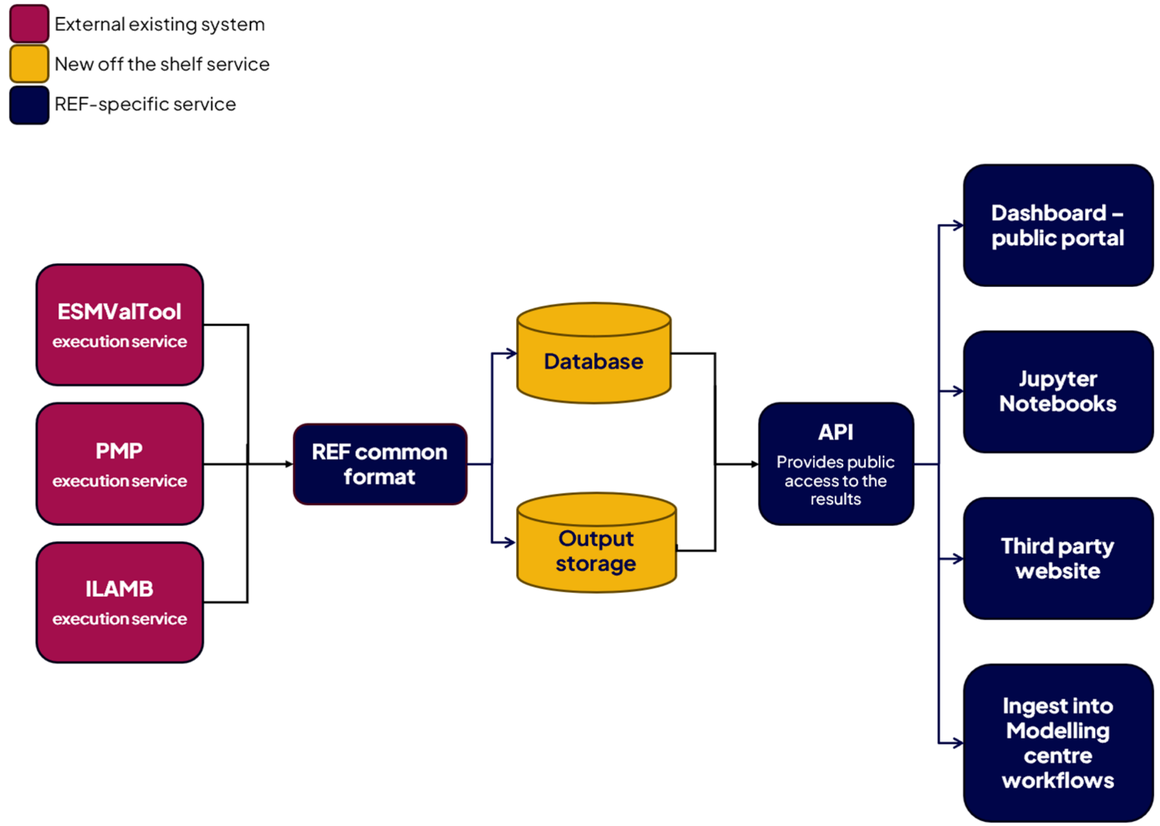
New off the shelf service

REF-specific service

Lewis *et al.* (2025). REF key services and interactions with external services including ESGF. Zenodo. DOI: 10.5281/zenodo.15595006

Components of the REF

- Abstraction for describing a Diagnostic that requires input datasets and produces standardised outputs
- Tooling to run executions in parallel on HPCs, via docker containers or locally
- Database layer for tracking all results for provenance
- API to provide results in a machine-readable format
- Portal to provide some basic information about the result



https://wcrp-cmip.org/cmip-phases/cmip7/rapid-evaluation-framework/

Diagnostic collections including reference datasets

The CMIP Model Benchmarking Task Team, through consultation with the CMIP modelling community, have collected a set of informative diagnostic collections for five different Earth system model themes. for CMIP Assessment Fast Track model evaluation

Oceans & Sea ice - 6 diagnostic collections, 6 reference datasets

Atmosphere - 6 diagnostic collections, 6 reference datasets

Land & Land ice - 7 diagnostic collections, 7 reference datasets

Earth system - 5 diagnostic collections, 1 reference dataset

Impacts & adaptation - 2 diagnostic collections, 1 reference dataset

For the full list see: DOI 10.5281/zenodo.14284374

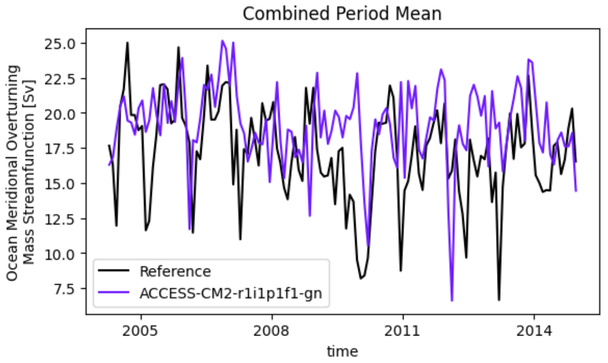
Dashboard

REF Dashboard

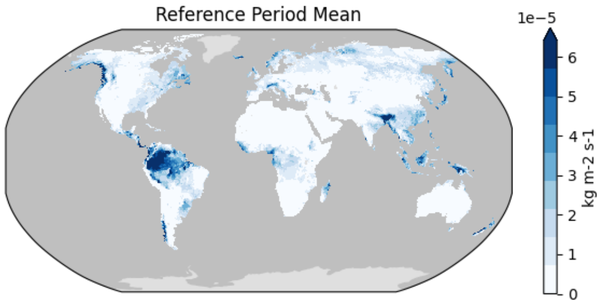
REF Documentation

REF GitHub repository

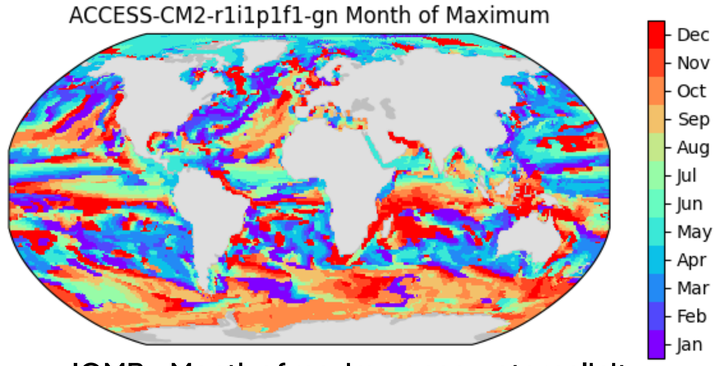
ILAMB and IOMB Diagnostics



IOMB – AMOC

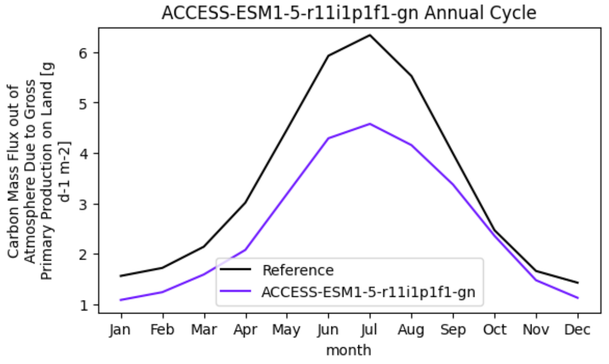


ILAMB – Runoff

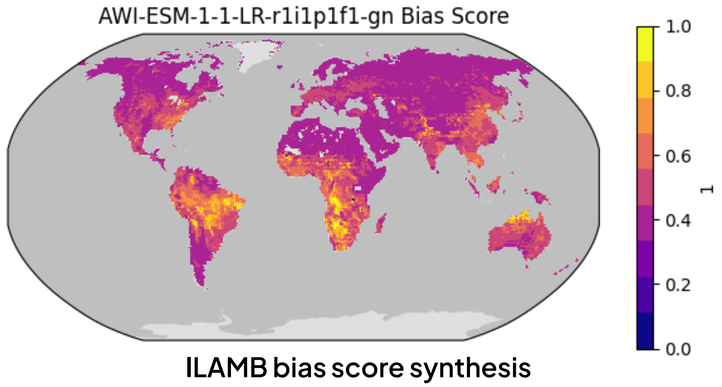


IOMB – Month of maximum sea water salinity

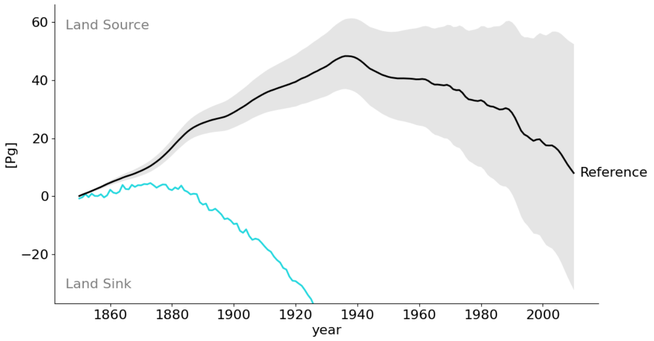
ILAMB contributes key diagnostics covering important variables and processes – some examples include sea water salinity, run off, net biome productivity, gross primary productivity, ocean meridional overturning.



ILAMB – gross primary productivity

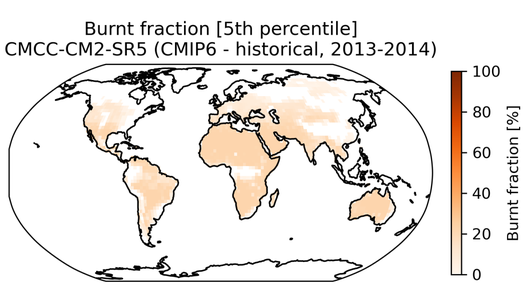


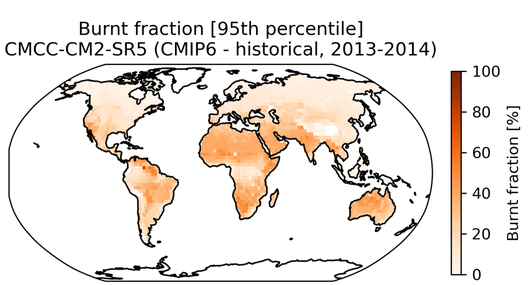
ILAMB bias score synthesis

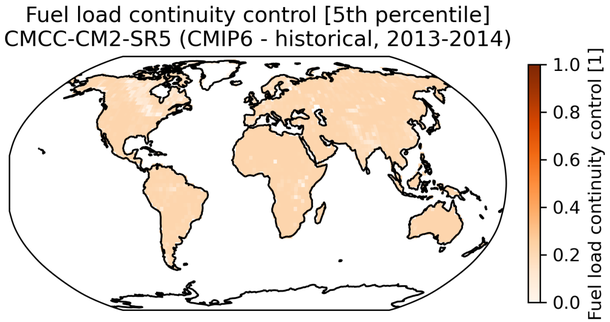


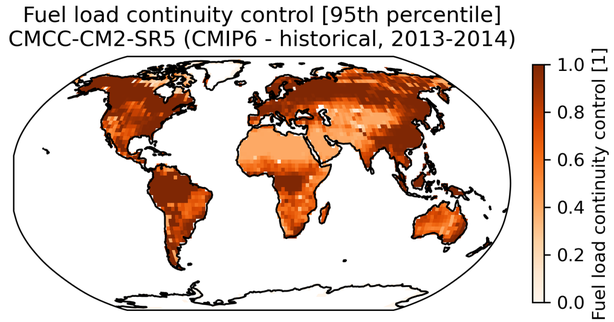
ILAMB – accumulated Net Biome Productivity

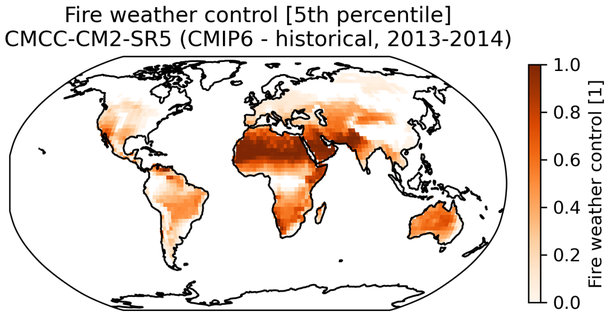
Case study: Fire diagnostics

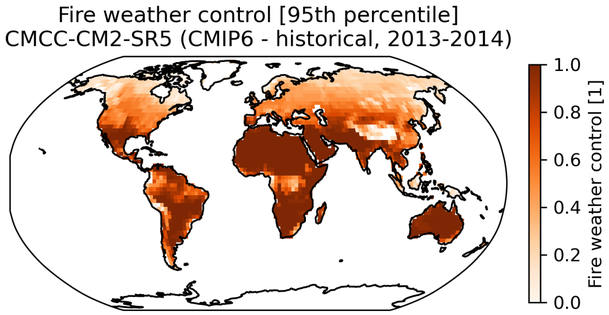






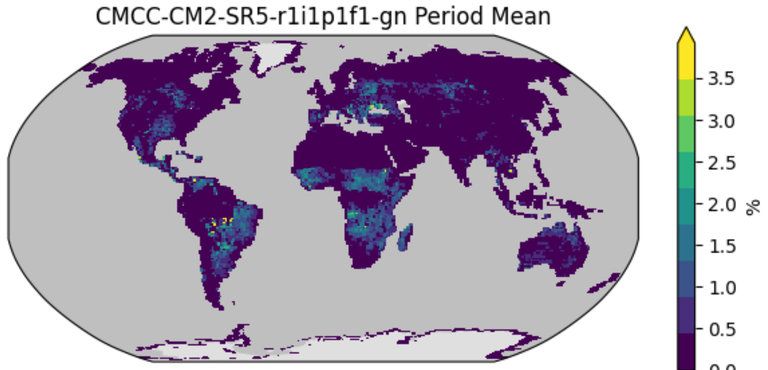




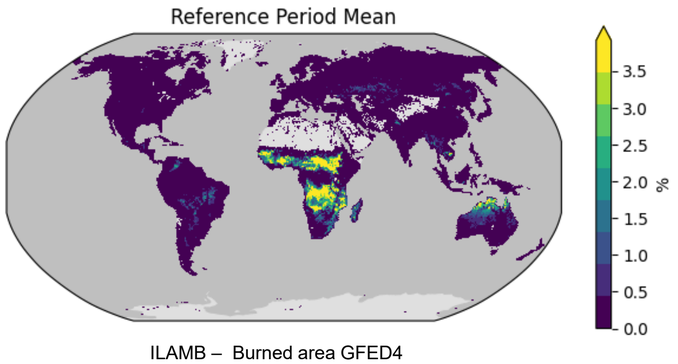


ESMValTool Fire diagnostic produces annual burnt area , fire weather control and fuel continuity at the 5th and 95th percentiles. Calculations based on the ConFire model (Jones et al, 2024).

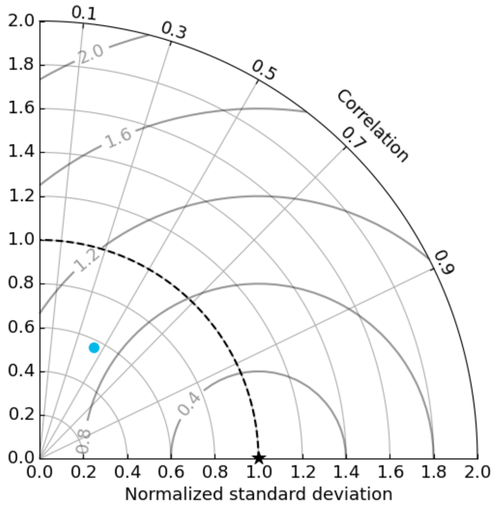
Case study: Fire diagnostics



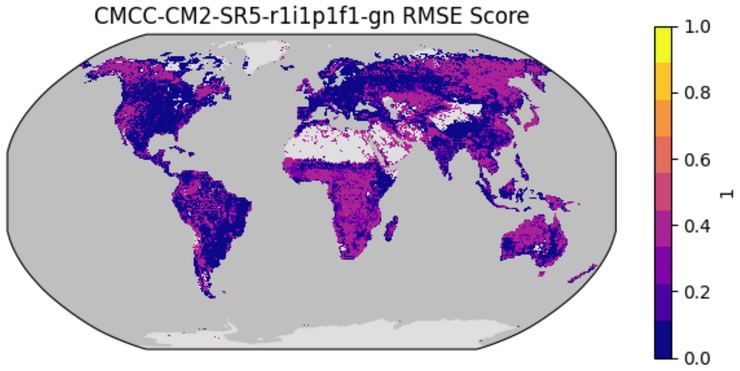
ILAMB – Burned area CMCC-CM2



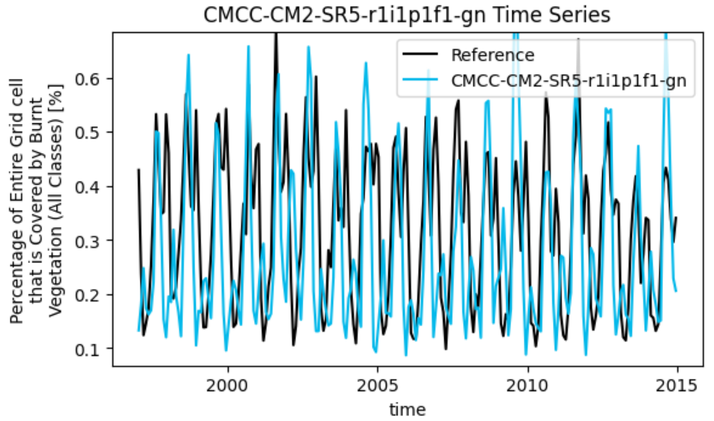
ILAMB – Burned area GFED4



ILAMB – Taylor diagram



ILAMB –CMCC-CM2



ILAMB –CMCC-CM2

ILAMB Fire diagnostics provide different visualizations of burnt area fraction – means, Taylor diagrams and time series, RMSE scores for regional and global values (Collier et al 2018).

CMIP | WCRP

obs4MIPs
Observations for Model Intercomparison Projects

Pathways to reference data publication on ESGF

- Dataset proposal to be submitted to the obs4MIPs Steering Panel
- Once accepted, submit Registered Content (RC) as an issue to obs4MIPs Github repository
  - Contains essential metadata on the dataset and producer
- Dataset preparation:
  - Use CMOR following obs4MIPs instructions to prepare the dataset
  - Create CMOR-like datasets following ODS2.6 (workflows implemented by ESMValTool and ILAMB) – obs4REF
    - For ILAMB community members, ILAMB3 has a workflow for outputting data into ILAMB-ready, CF-compliant, obs4MIPs Data Standards compliant, ESGF publication ready

https://www.wcrp-esmo.org/projects-and-panels/obs4mips

Thank You

@wcrpcmip

wcrp-cmip

cmip-ipo@esa.int

Forrest Hoffman
Model Benchmarking Task Team Co-lead

Birgit Hassler
Model Benchmarking Task Team Co-lead

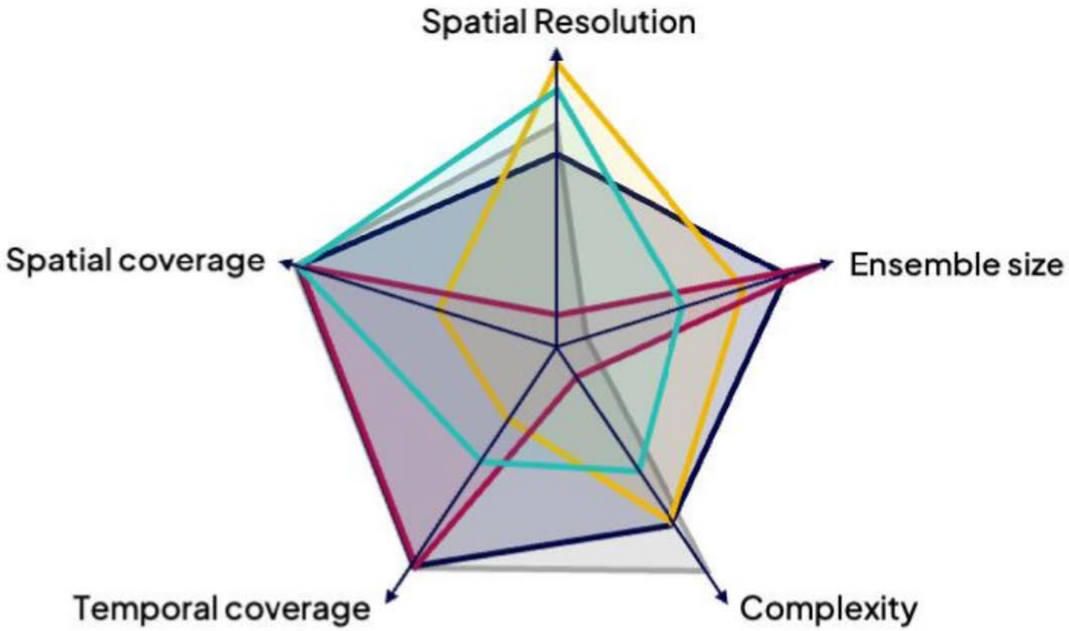
Ranjini Swaminathan
Model Benchmarking Task Team Co-lead

Jared Lewis
REF delivery team Technical Manager

Extra Slides

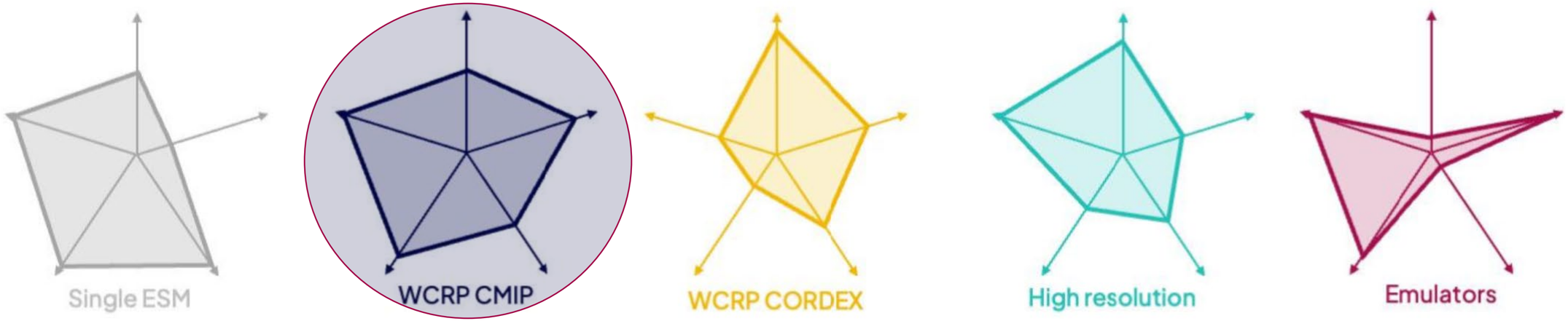
The WCRP Modelling Multiverse

Each type of model or modelling project has a different ability to model over different spatial resolutions, spatial coverage, temporal coverage, model complexities, and ensemble sizes.



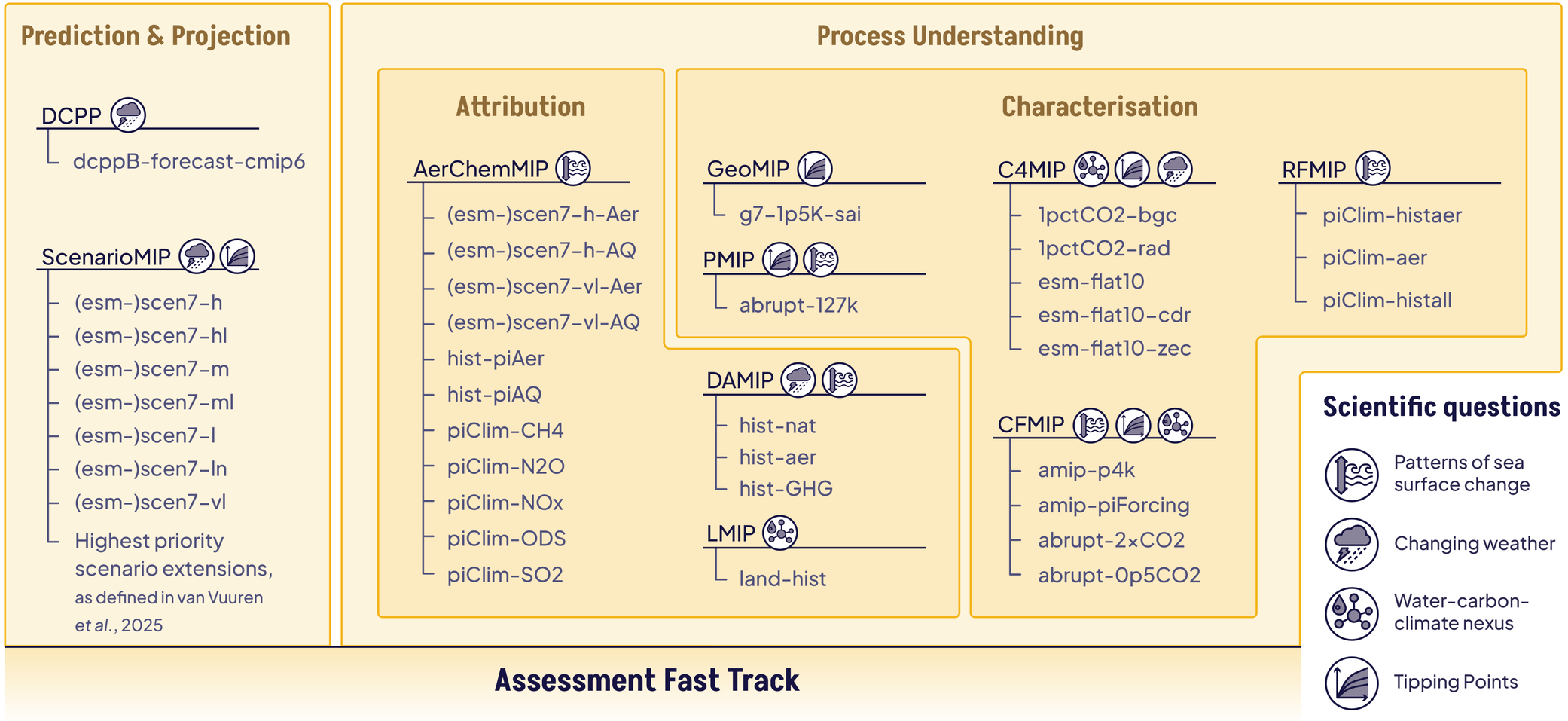
The Rapid Evaluation Framework (**REF**) has been developed for use with **CMIP data.**

It is designed to be a starting point for the community to develop and build upon, with applications across the WCRP Modelling Multiverse.



DOI: 10.5281/zenodo.8047805

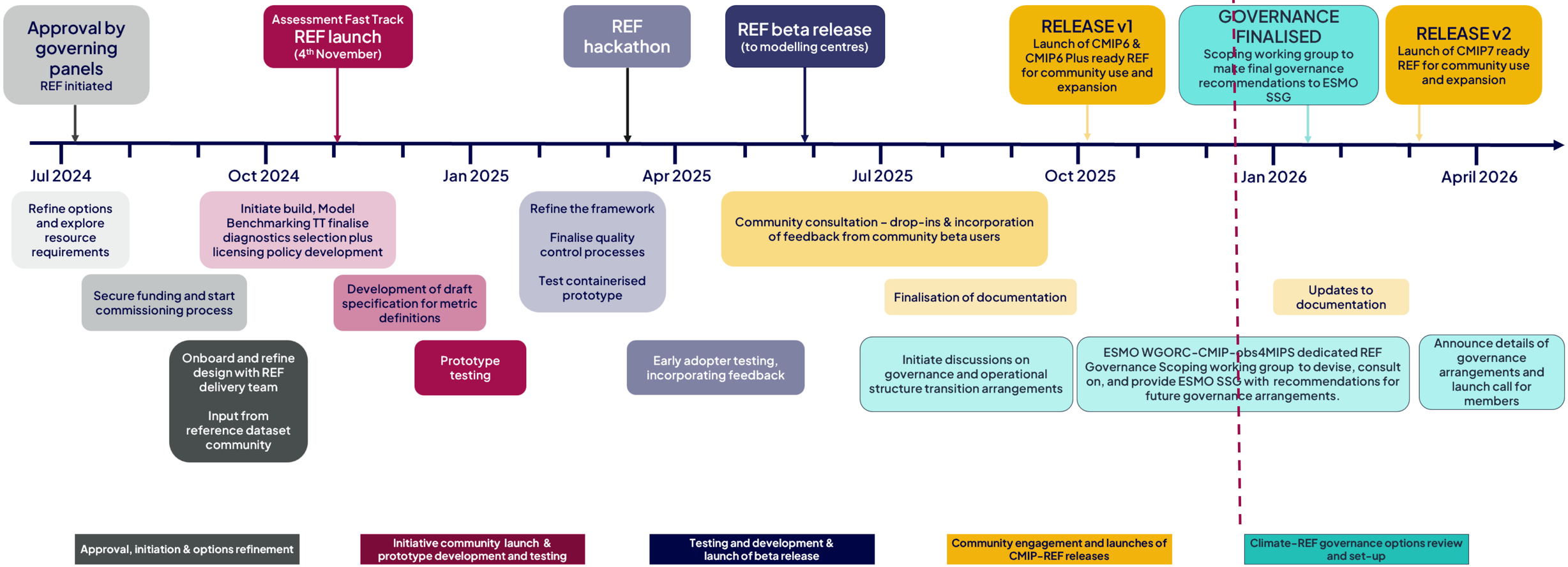
The CMIP Assessment Fast Track



https://wcrp-cmip.org/cmip-phases/cmip7/fast-track/

Rapid Evaluation Framework

REF timeline



Reference datasets

Example uses of the REF

Produced by members of the CMIP Model Benchmarking Task Team

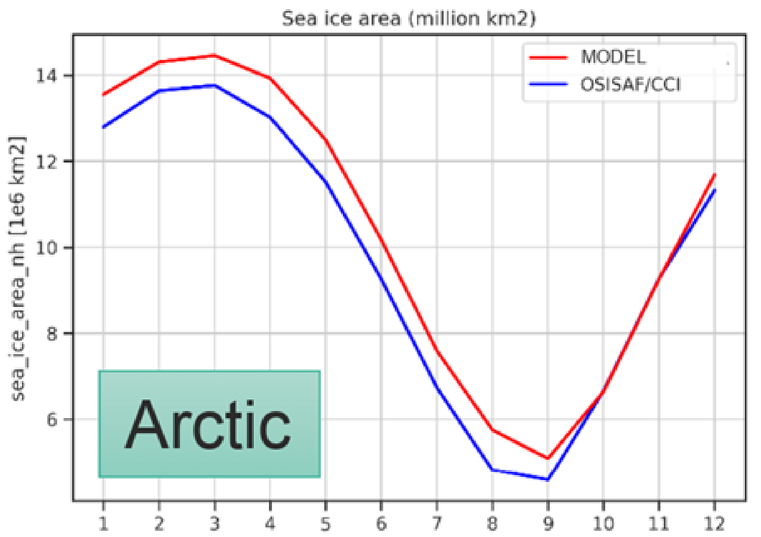
REALM: OCEANS & SEA ICE

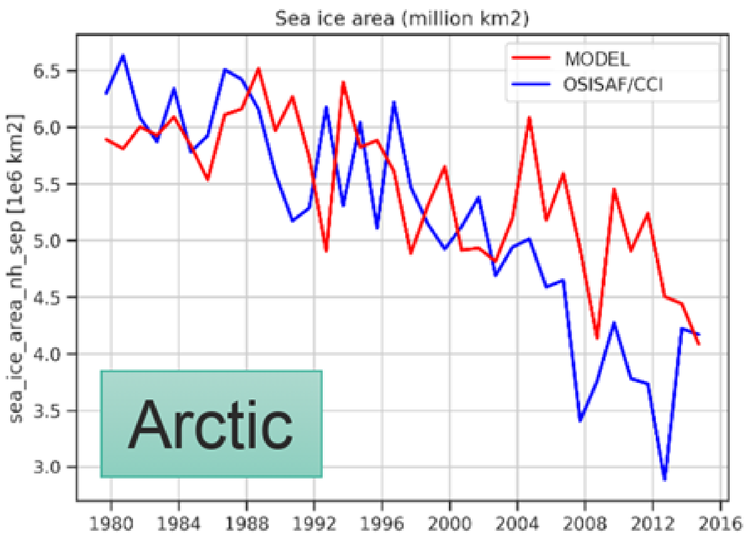
Ed Blockley, Met Office, UK

Met Office

Sea ice area: basic

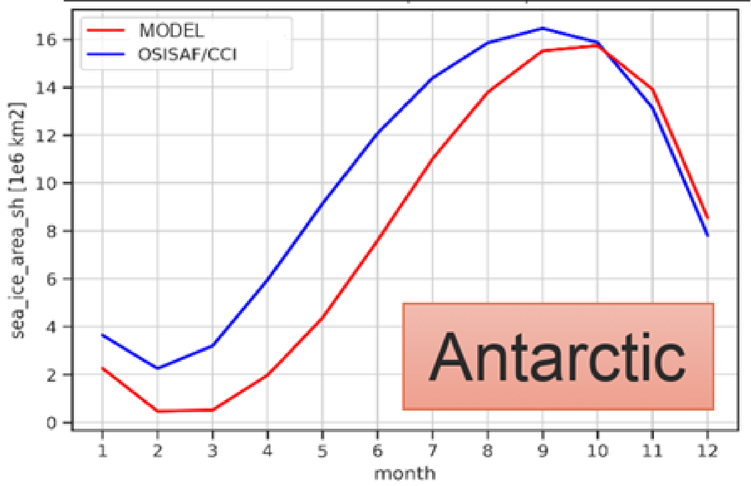
- Hemispheric sea ice area: Arctic & Antarctic
- Mean-seasonal-cycle
- Summer historical evolution:
  - NH (September); SH (February)
- Compared with satellite estimates
- Allows access to quick information on state of sea ice area in each simulation
  - Modelling centres for evaluation
  - Informed data users

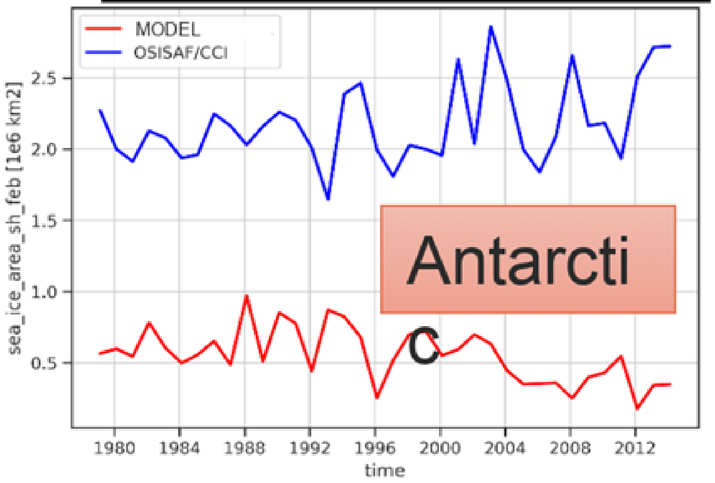




Mean-seasonal-cycle

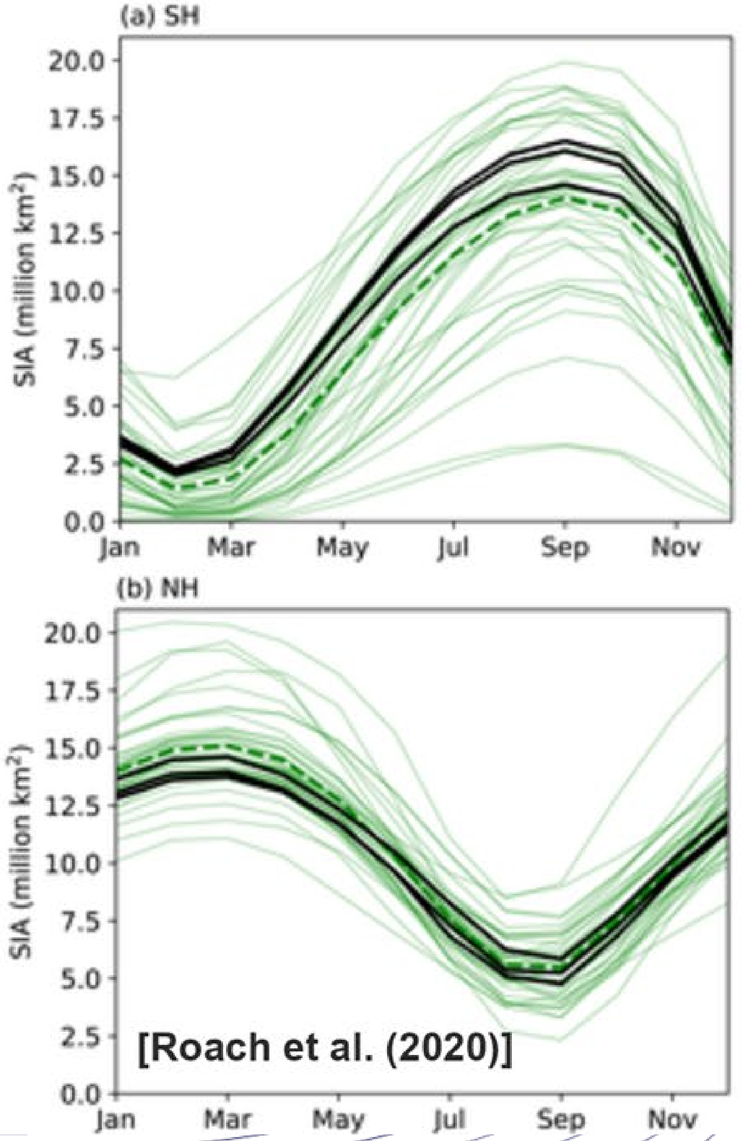
Summer: historical





Sea ice area: comparison

- Individual ensemble members do not give the general picture (high internal variability):
  - Either for a particular model or multi-model consensus
- Plans to include multiple models & ensembles:
- Extend model-obs time-series and mean-seasonal-cycle plots to include ensemble spread
  - Are the observations consistent with the model's ensemble spread?
- Multi-model comparison of hemispheric sea ice area mean-seasonal-cycle
  - Roach et al. (2020) SIMIP benchmarking paper

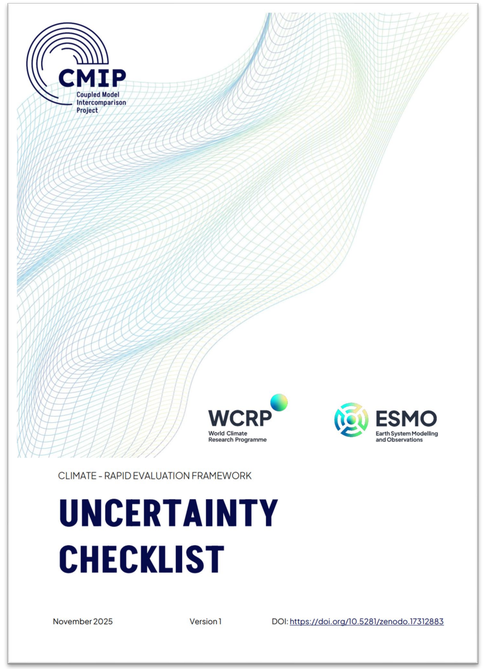


REALM: LAND & LAND ICE

Forrest Hoffman, ORNL

OAK RIDGE National Laboratory

Treatment of uncertainties



Hegedus *et al* (2025)
DOI: 10.5281/zenodo.17312883

- REF requires uniform approach
- Consultations with obs4MIPs, observation community, metric package developers and CMIP7-CVs Task Team
- Additional global attributes describing if and what uncertainty information is included (“has_auxdata”, “aux_variable_id”)
- Different approaches for CMOR and CMOR-like datasets
  - CMOR: each uncertainty information is included in separate files
  - CMOR-like: uncertainty information included as ancillary variables following CF conventions
- Participation of REF as case study in NPL led working group on new Uncertainty Metadata/Nomenclature Convention (UNC)

Transition criteria for REF use of obs4MIPs

- **CMIP7 compatibility** – new CVs, CMIP7 JSON –LD structure.
- **New ODS 2.6 specifications** – provision to indicate additional uncertainty. information, point and station data, doi.
- **Climatologies.**
- **Uncertainty information** – ancillary variables vs files.
- **CMOR-like vs CMOR.**

obs4REF has been set up as an interim ESGF publishing project.
Once obs4MIPs is able to meet all the REF transition criteria, the datasets can be transferred to the obs4MIPs publishing project