

# The History and Future of Systematic Assessment of Earth System Models

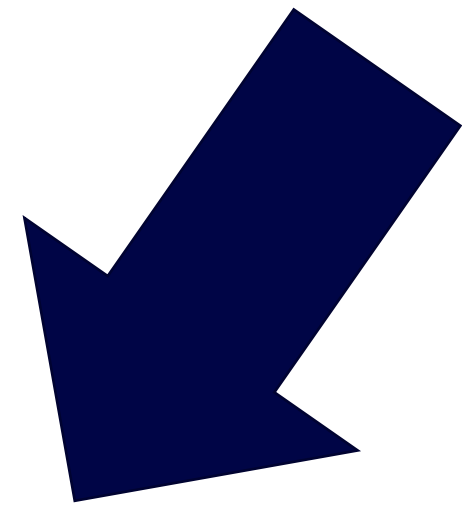
Birgit Hassler, DLR, and Forrest M. Hoffman, ORNL

UK Met Office Science Seminar

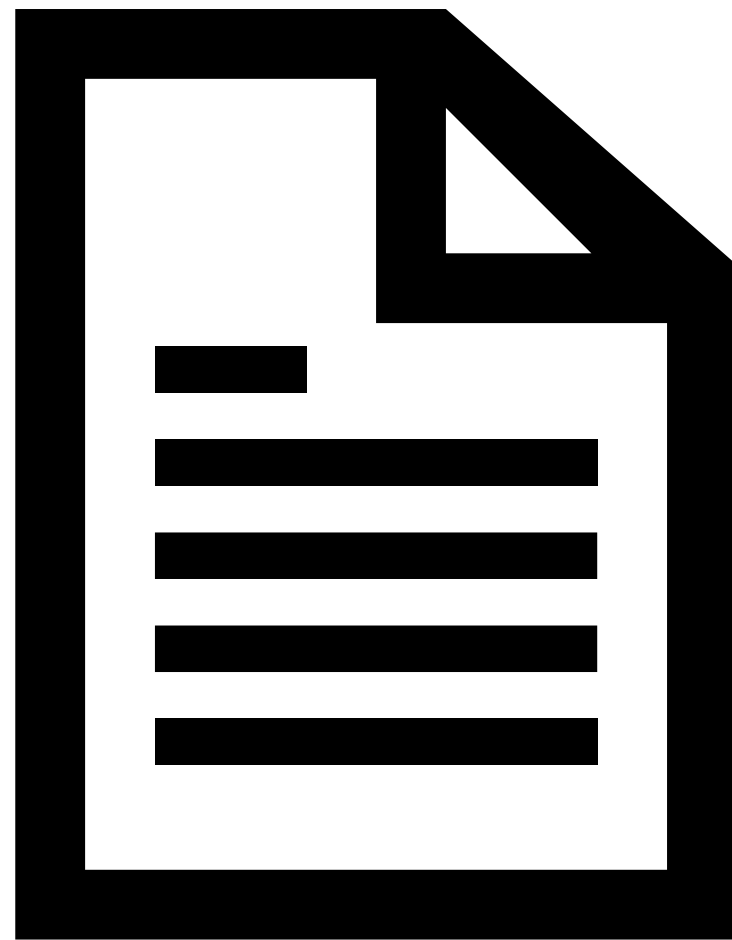
11 March 2025



# History and Future



Not just, but including:



„Systematic Benchmarking of Climate Models: Methodologies, Applications, and New Directions”. Manuscript submitted to Reviews of Geophysics.

CMIP

WCRP

# CMIP Model benchmarking Task Team



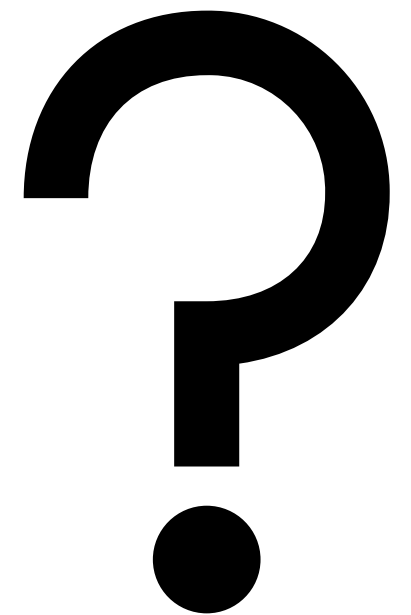
Rebecca Beadling, *USA*  
Ed Blockley, *UK*  
Jiwoo Lee, *USA*  
Valerio Lembo, *Italy*  
Jared Lewis, *Australia*  
Jianhua Lu, *China*  
Luke Madaus, *USA*  
Elizaveta Malinina,  
*Canada*  
Brian Medeiros, *USA*  
Wilfried Pokam Mba,  
*Cameroon*  
Enrico Scoccimarro, *Italy*  
Ranjini Swaminathan,  
*UK*

Task Team members from 2022 to 2024 (Phase 1).



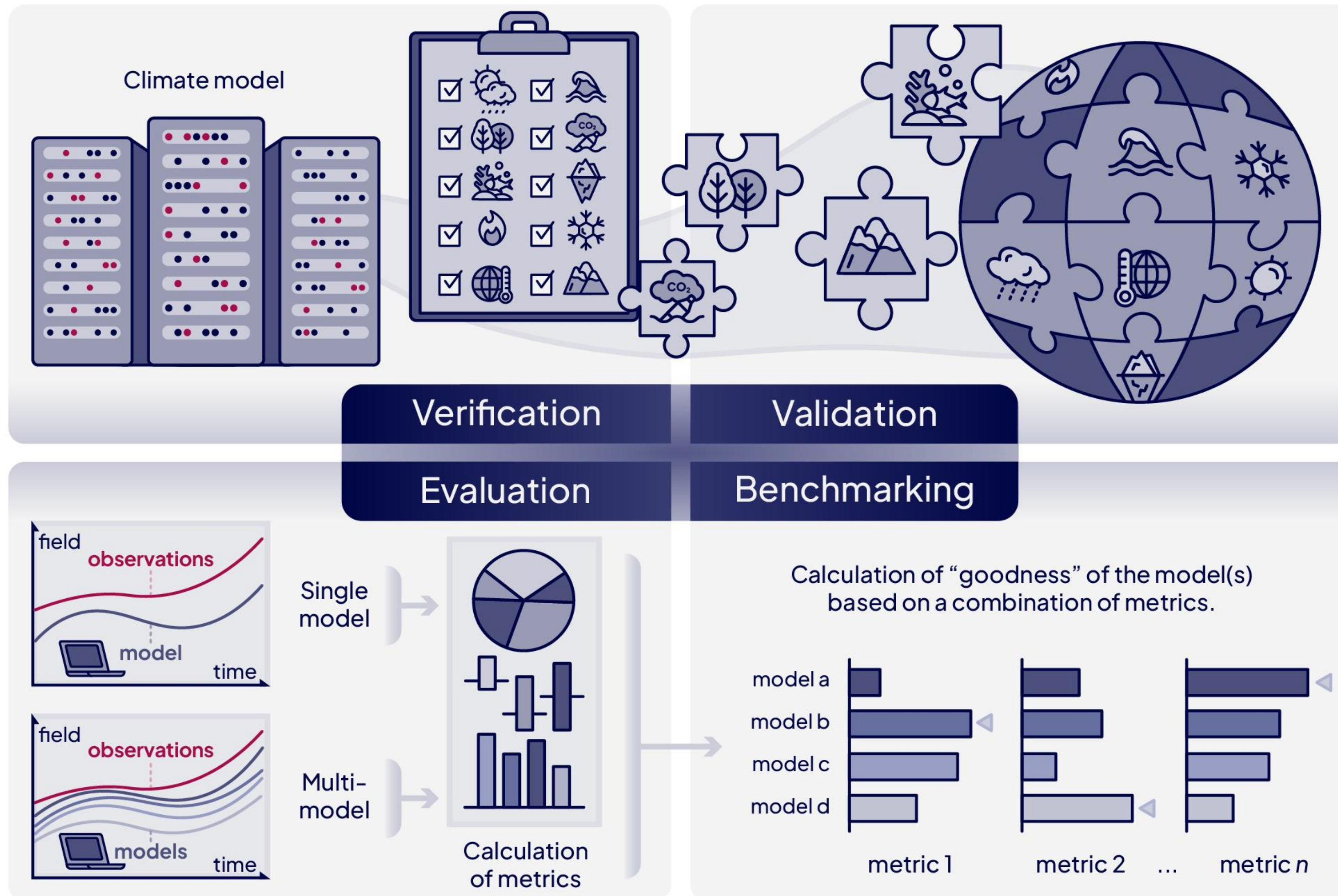
# Definition of model evaluation and benchmarking

- **What** is model evaluation and benchmarking?
- **How** was the evaluation and benchmarking done for different CMIP phases?
- **What** tools have been used to evaluate and benchmark the CMIP model ensemble?
- **What** are the challenges that model evaluation and benchmarking has to face, and developments it needs to undergo in the future?
- **What** would be a good framework to evaluate and benchmark CMIP AR7 Fasttrack simulations as soon as they become available on ESGF?



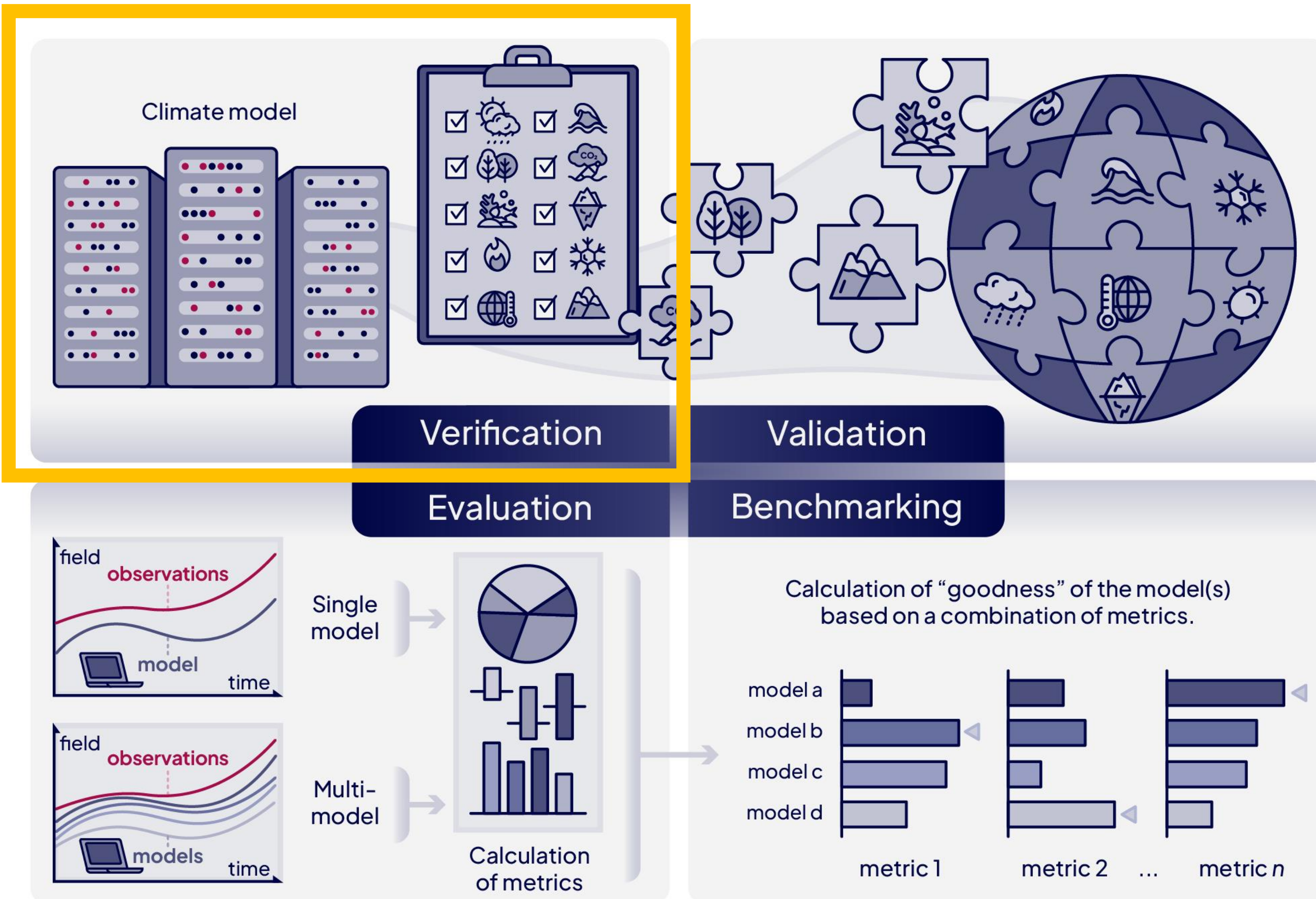


# Definition of model evaluation and benchmarking





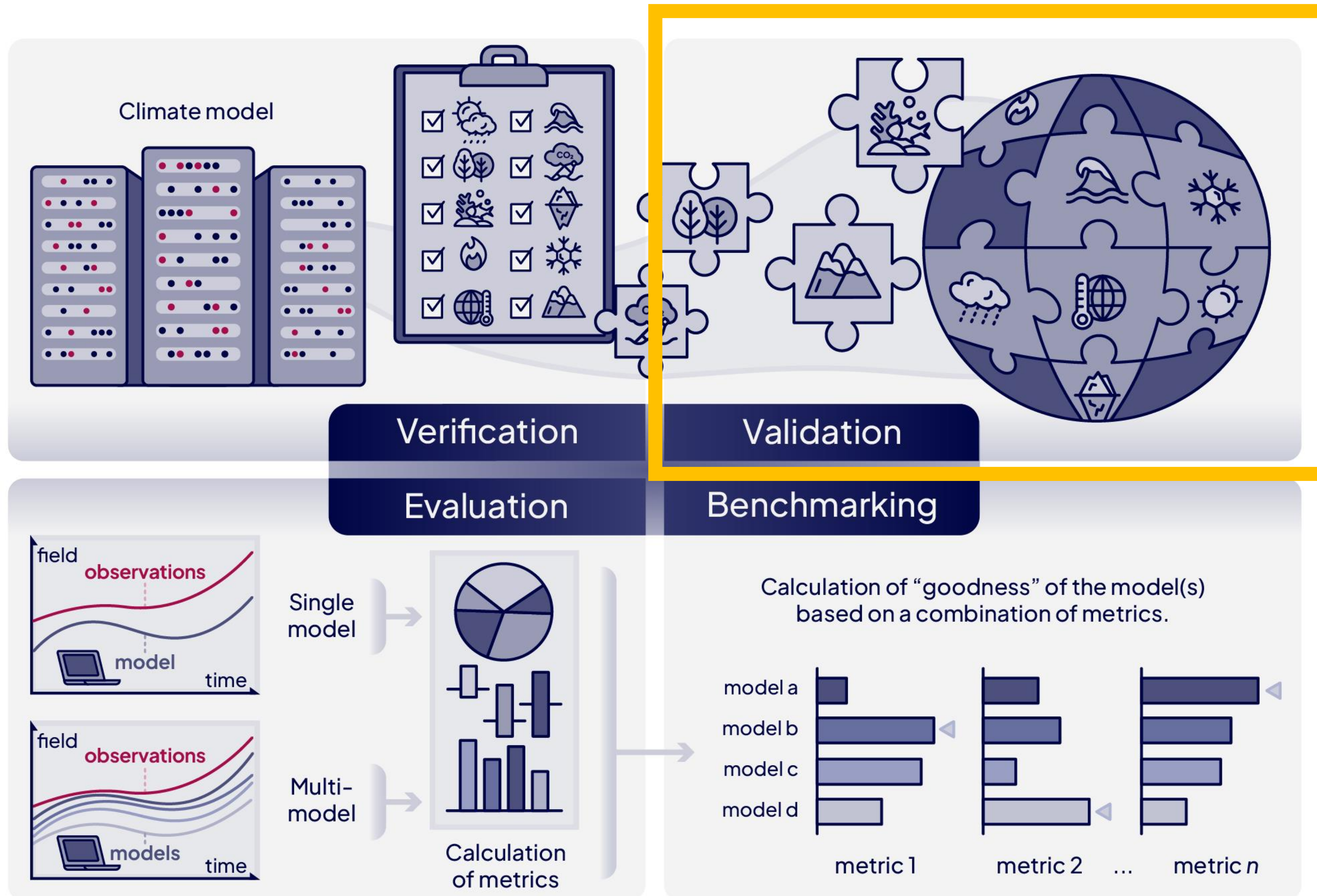
# Definition of model evaluation and benchmarking



- **Verification:** the process of assessing model consistency in terms of correct implementation of the included processes as expected by the model and experiment design.



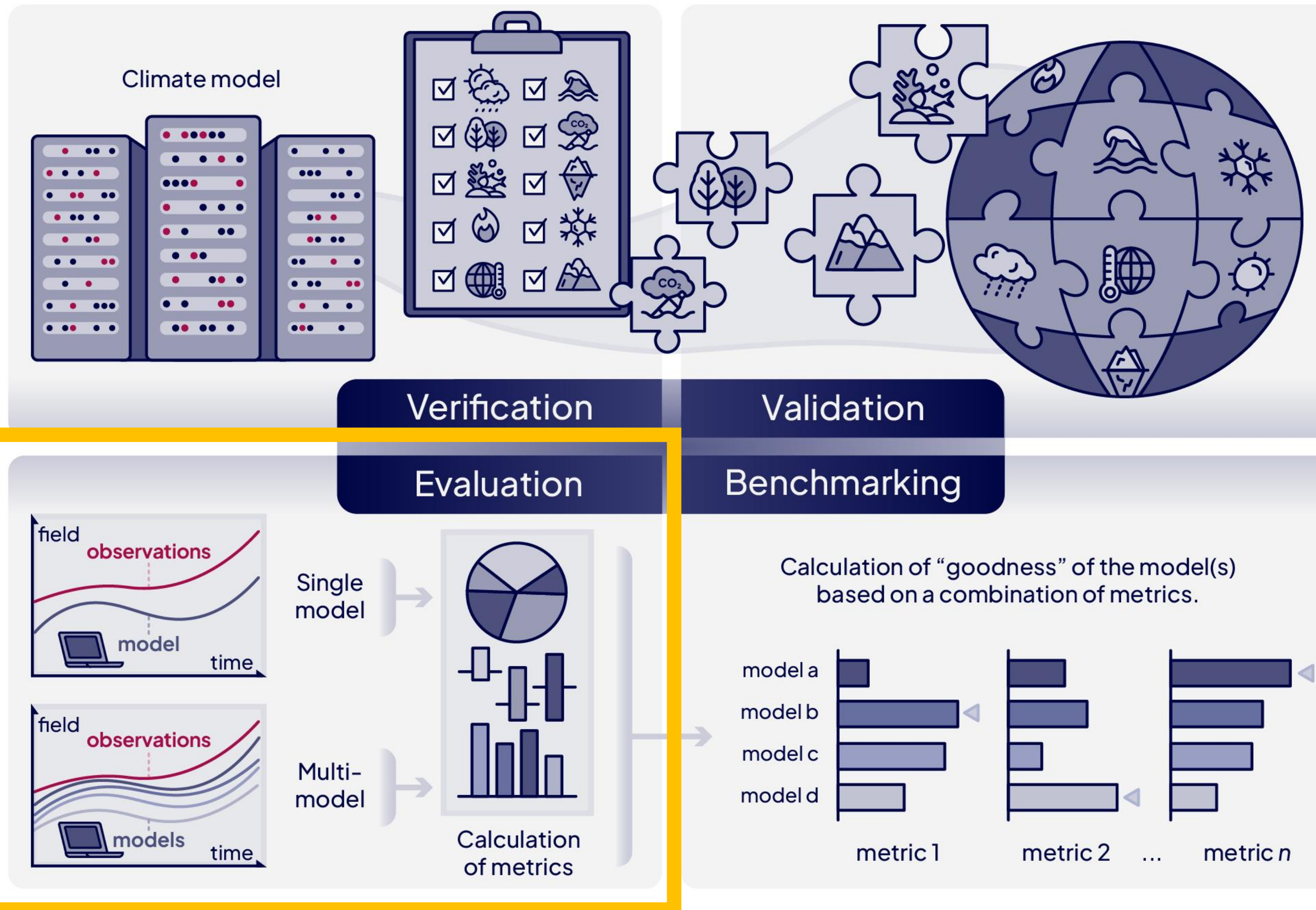
# Definition of model evaluation and benchmarking



- **Validation:** the process of determining the degree to which a model accurately represents processes in the real world, particularly for the intended uses of the model.



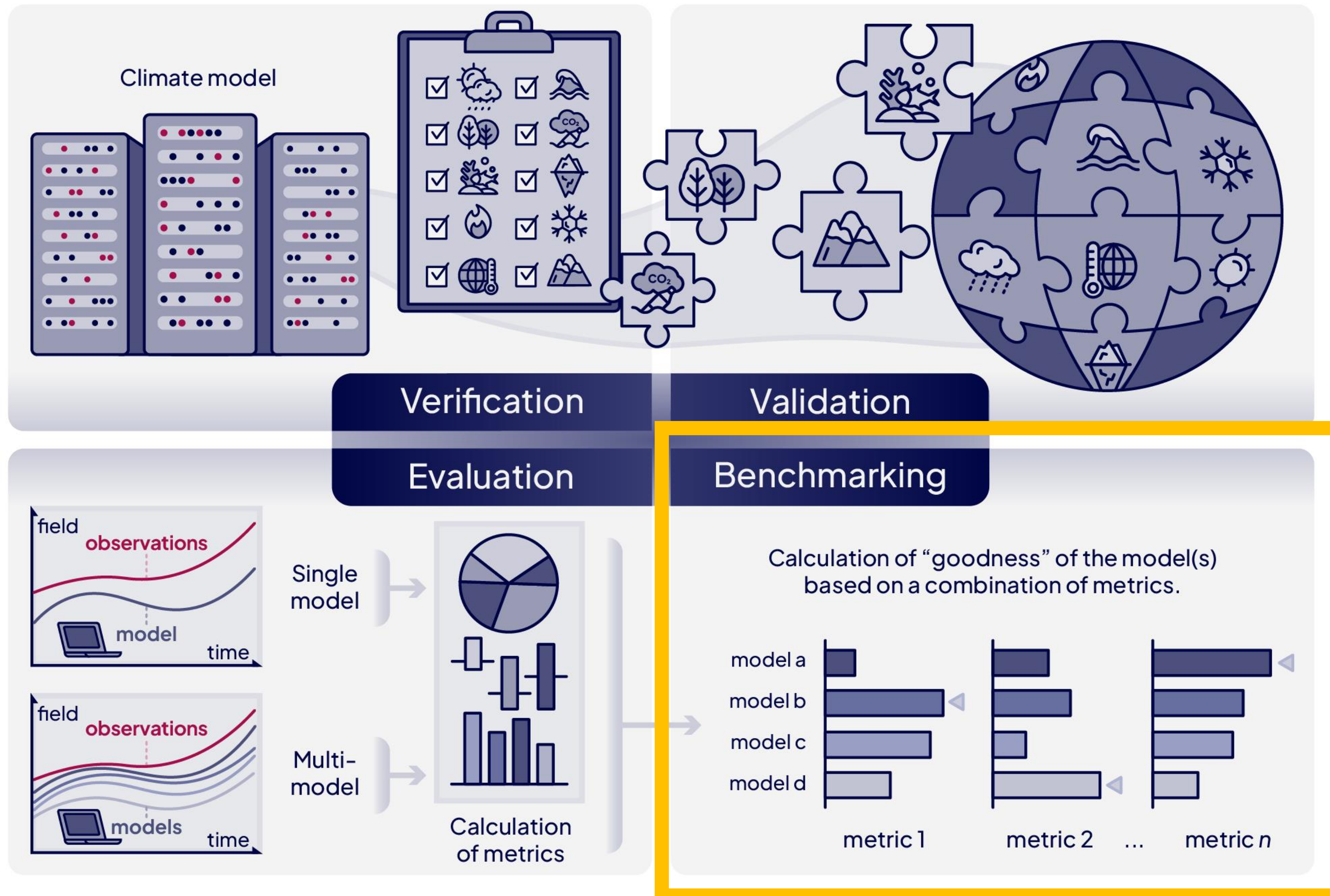
# Definition of model evaluation and benchmarking



- **Evaluation:** the process of assessing simulations against observations



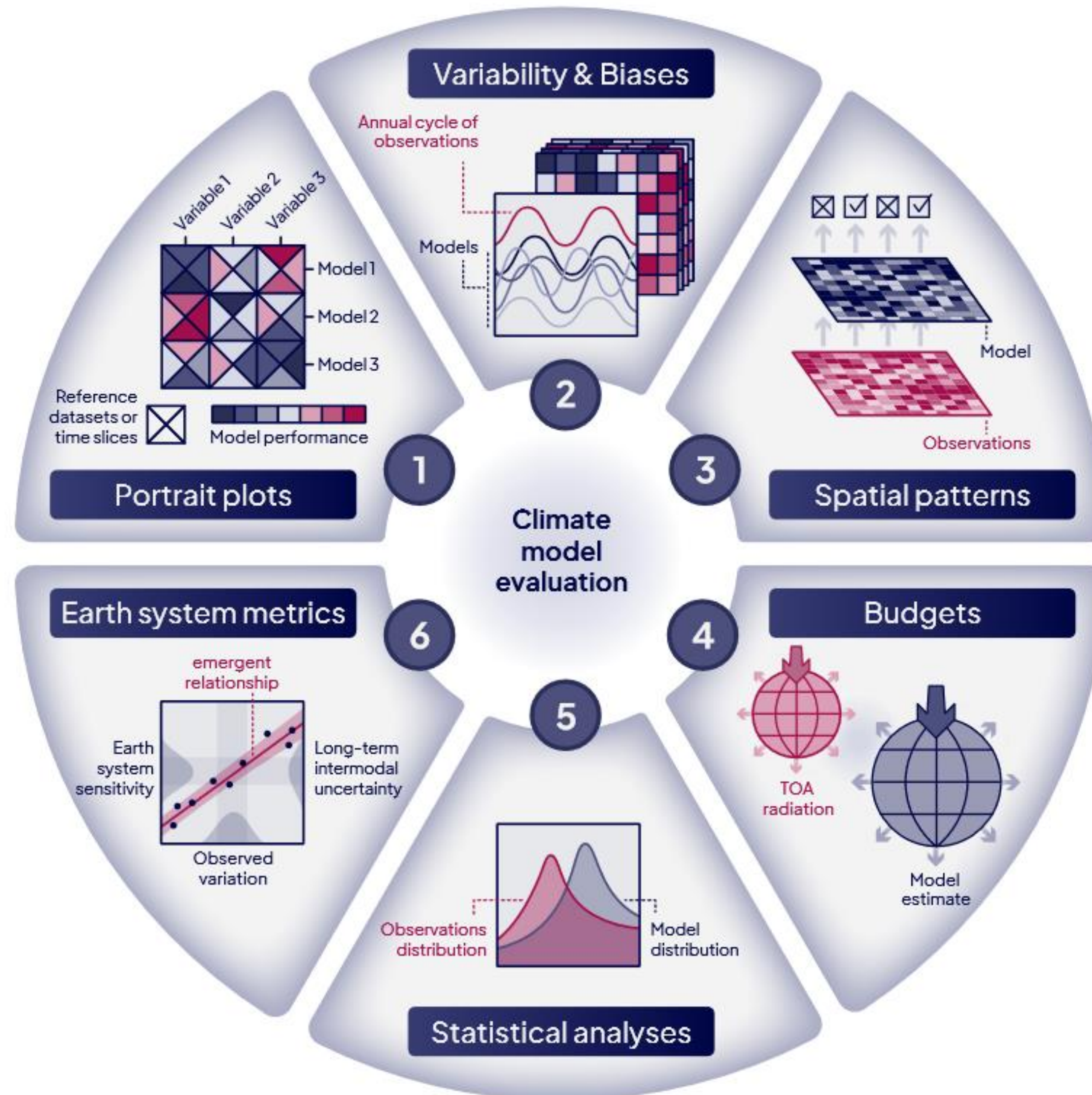
# Definition of model evaluation and benchmarking



- **Benchmarking:** the process where model simulations are evaluated with observations, often resulting in a statement made about the goodness of the simulation or model.



# Benchmarking and evaluation schemes



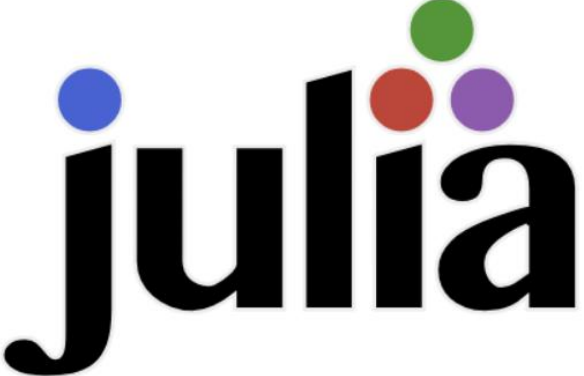




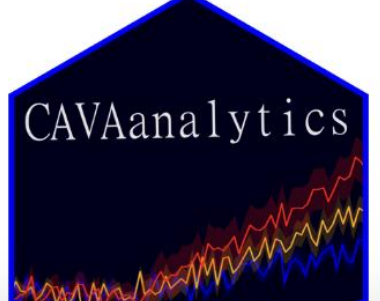




- **Six general schemes**, grouped according to their underlying evaluation principles
- Portrait plots, Variability & Biases, Spatial patterns, Budgets, Statistical analyses, Earth System metrics
- Most of them can be applied **regionally or globally**



# Tools for analysing climate data

CMIP have created a community database of tools which can be used for analyses with climate and CMIP data.

 <p><b>Python</b></p> <p>CATEGORY Coding language</p> <p>DESCRIPTION Python is a high-level, general-purpose programming language.</p> <p>COMMUNITY SUPPORT? Mailing list Slack Other (specify)</p>	 <p><b>R</b></p> <p>CATEGORY Coding language</p> <p>DESCRIPTION R is a free software environment for statistical computing and graphics.</p> <p>COMMUNITY SUPPORT? Mailing list FAQs Other (specify)</p>	 <p><b>Julia</b></p> <p>CATEGORY Coding language</p> <p>DESCRIPTION Julia is a high-level, general-purpose dynamic programming language. Its features are well suited for numerical analysis and computational science.</p> <p>COMMUNITY SUPPORT? Forums GitHub/GitLab Other (sp)</p>	 <p><b>ESGF</b></p> <p>CATEGORY Data access platform</p> <p>DESCRIPTION The Earth System Grid Federation (ESGF) Peer-to-Peer (P2P) enterprise system is a collaboration that develops, deploys and maintains ...</p> <p>COMMUNITY SUPPORT? Mailing list GitHub/GitLab</p>	 <p><b>PANGEO</b></p> <p>CATEGORY Data access platform</p> <p>DESCRIPTION Pangeo Catalog is an open-source project to enumerate and organize cloud-optimized climate data stored across a variety of providers. In ...</p> <p>COMMUNITY SUPPORT? Forums GitHub/GitLab</p>
 <p>Climate Data Store (CDS) Climate data at your fingertips</p>	 <p>ipcc data distribution center</p>	 <p>CAVAanalytics</p>	 <p>xMIP</p>	 <p>PANGEO</p>





[www.wcrp-cmip.org/tools](http://www.wcrp-cmip.org/tools)



# Benchmarking and evaluation tools

Filter Sortieren



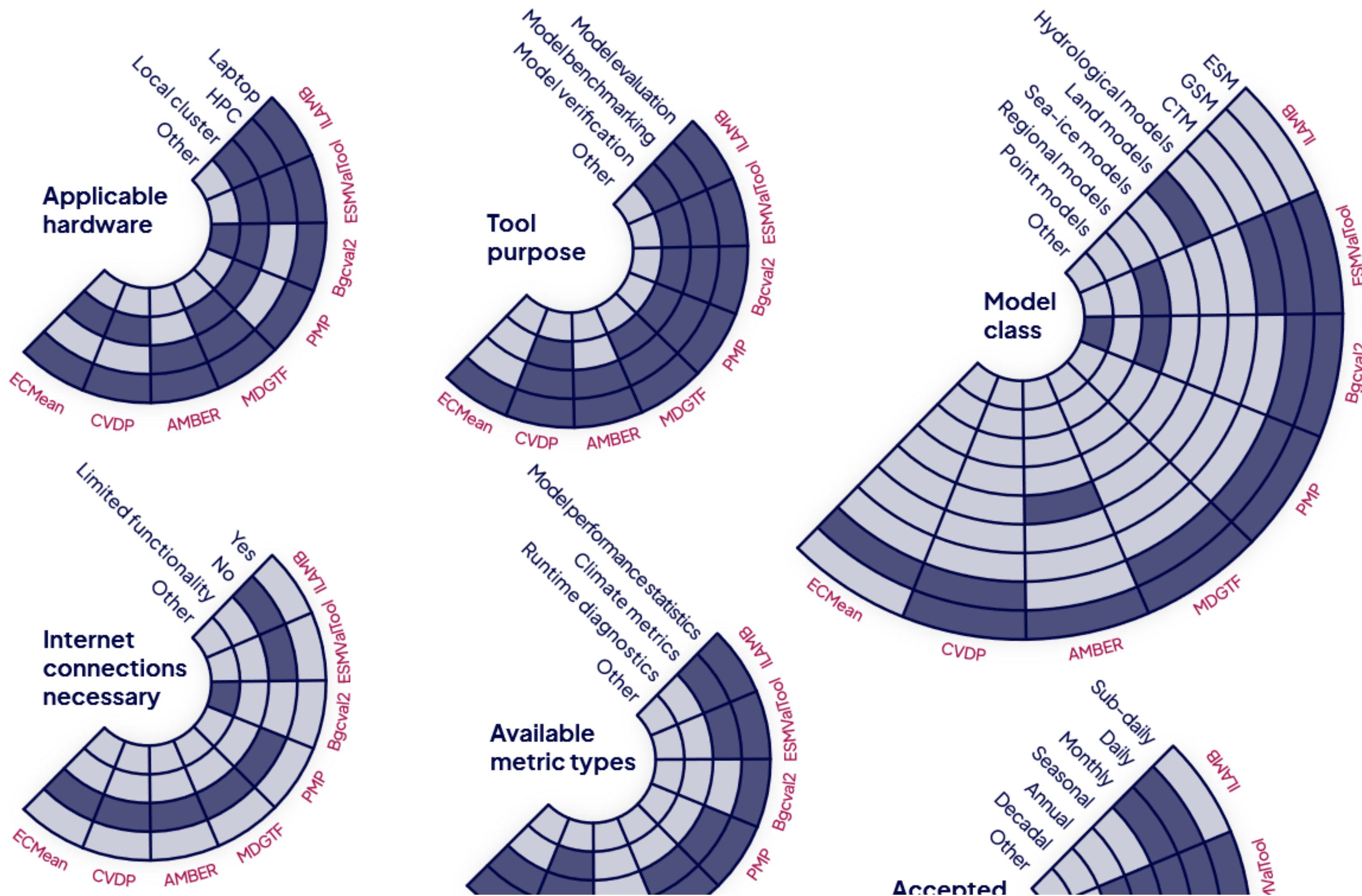
 <p><b>RUBISCO</b></p> <p>ILAMB</p> <p>Category Evaluation and benchmarking tools</p> <p>Description The International Land Model Benchmarking (ILAMB) project is a model-data intercomparison and integration project designed to assess the performance of land ...</p> <p>Website <a href="https://www.ilamb.org/">https://www.ilamb.org/</a></p>	 <p>ESMValTool Earth System Model Evaluation Tool</p> <p>ESMValTool</p> <p>Category Evaluation and benchmarking tools</p> <p>Description ESMValTool is an open-source community-developed diagnostics and performance metrics tool for the evaluation and analysis of Earth System Models.</p> <p>Website <a href="https://www.esmvaltool.org/">https://www.esmvaltool.org/</a></p>	 <p>bgcval2</p> <p>bgcval2</p> <p>Category Evaluation and benchmarking tools</p> <p>Description Python based Software toolkit for monitoring on-going simulations of the ocean, and the marine component of earth System models.</p> <p>Website <a href="https://github.com/valeriupredoi/bgcval2">https://github.com/valeriupredoi/bgcval2</a></p>	 <p>PCMDI Metrics Package (PMP)</p> <p>Category Evaluation and benchmarking tools</p> <p>Description The PCMDI Metrics Package (PMP) is an open source Python software that provides "quick-look" objective comparisons of Earth System Models (ESMs) with one another and available...</p> <p>Website <a href="http://pcmdi.github.io/pcmdi_metrics/">http://pcmdi.github.io/pcmdi_metrics/</a></p>
--	--	--	---

- Open-source
- Useable for CMIP data analyses

- 11 tools in the database so far
- Available at: <https://wcrp-cmip.org/tools/model-benchmarking-and-evaluation-tools/>



# Summary of tool characteristics

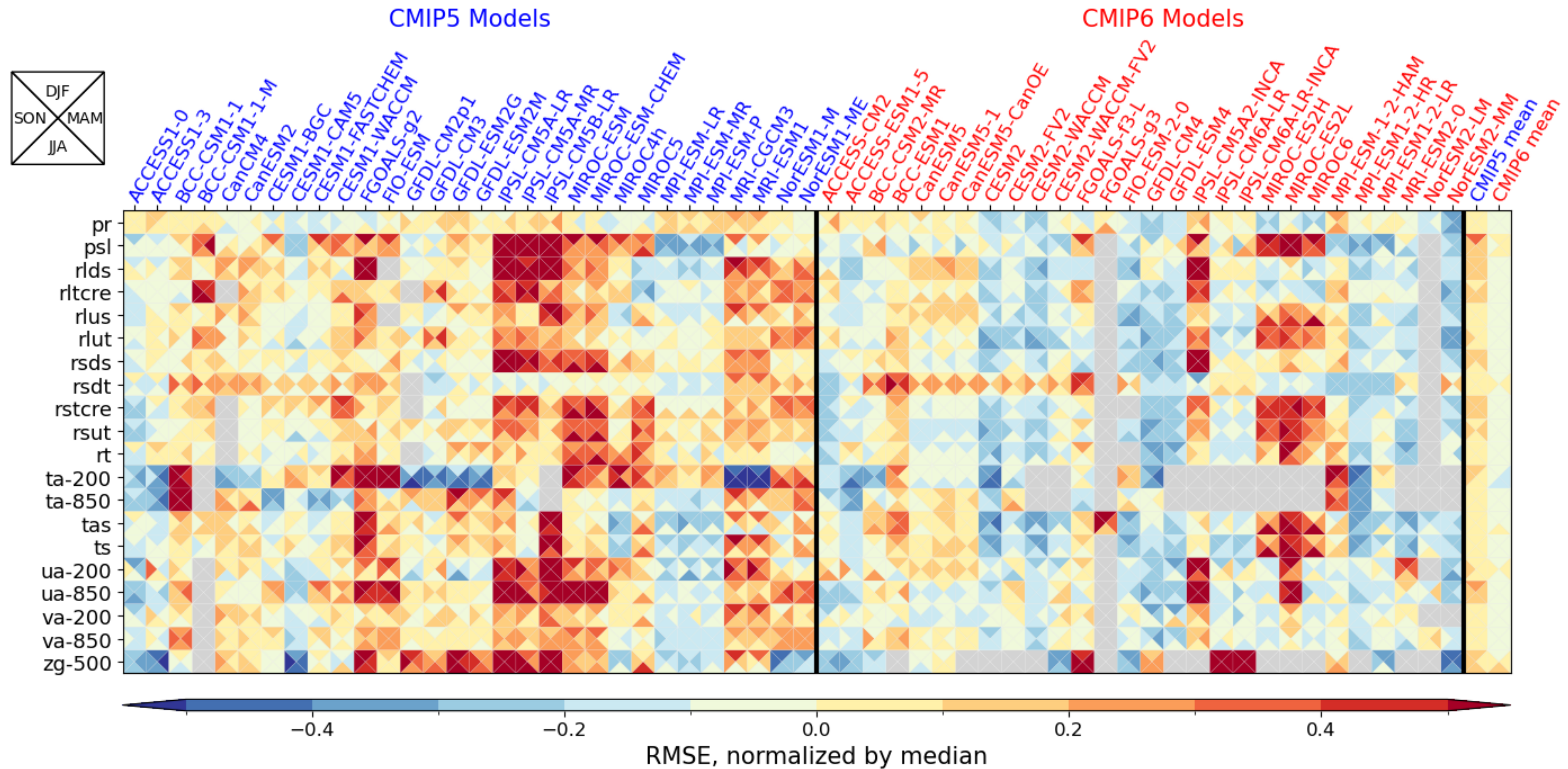


- Eight open-source tools are summarized by their ten main characteristics



# Improvements through different CMIP phases

- Benchmarking atmospheric variables from CMIP5 and CMIP6 models





# Improvements through different CMIP phases

	CMIP5 Models								CMIP6 Models								Mean CMIP5	Mean CMIP6		
	bcc-csm1-1	CanESM2	CESM1-BGC	GFDL-ESM2G	IPSL-CM5A-LR	MIROC-ESM	MPI-ESM-LR	NorESM1-ME	UK-HadGEM2-ES	BCC-CSM2-MR	CanESM5	CESM2	GFDL-ESM4	IPSL-CM6A-LR	MIROC-ESM2L	MPI-ESM1-2-HR			NorESM2-LM	UKESM1-0-LL
<b>Ecosystem and Carbon Cycle</b>	-0.33	-0.13	-0.78	-1.29	-0.07	0.03	-2.91	-0.78	0.56	0.45	0.80	0.58	-1.33	0.22	0.58	0.86	0.37	0.58	0.89	1.69
⊕ Biomass	-0.43	-0.23	-0.93	-1.51	-1.58	-0.32	-0.95	-1.41	1.10	1.72	0.23	0.10	-0.37	0.25	0.41		0.31	0.64	0.96	2.01
⊕ Carbon Dioxide		-0.27	0.12	-0.72	-0.00	0.74	-3.50	0.36	0.52		0.38	0.32		0.43	0.60		0.31	0.46		0.25
⊕ Gross Primary Productivity	0.77	-0.98	0.55	-2.21	-1.60	-0.25	-0.89	0.26	-0.62	0.74	-0.10	0.22	-0.14	0.93	-0.76	-0.14	0.32	0.21	1.51	2.21
⊕ Leaf Area Index	-0.90	-0.18	-0.89	-2.56	-0.14	-0.84	0.18	-1.48	-0.28	0.46	0.51	0.04	-0.09	1.20	0.82	1.29	0.11	0.04	0.71	2.00
⊕ Global Net Ecosystem Carbon Balance		-0.53	-0.16	-0.28	0.35	0.12	-3.40	0.02	0.39		0.16	0.92		0.35	-0.09		0.99	0.39		0.76
⊕ Net Ecosystem Exchange	-0.26	-1.55	0.60	-2.09	0.30	-0.12	-0.16	0.57	-0.58	-1.51	-0.72	0.73	0.55	-0.46	1.37		1.26	-0.55	1.17	1.44
⊕ Ecosystem Respiration	0.92	-0.18	-0.49	-0.46	-2.00	0.42	-0.62	-0.59	-1.35	0.57	0.37	0.20	-0.37	0.92	-0.49		0.16	-0.99	1.64	2.35
⊕ Soil Carbon	0.60	1.39	-1.09	0.02	0.84	0.33	0.06	-0.77	0.22	0.30	1.50	-0.73	-1.84	-1.27	1.15		-1.78	0.23	1.22	-0.38
<b>Hydrology Cycle</b>	-1.94	-0.24	0.06	-0.40	-2.65	-0.72	-0.01	-0.17	0.50	0.05	-0.50	1.19	0.46	0.30	-0.66	0.14	0.79	1.03	1.07	1.70
⊕ Evapotranspiration	-0.34	-0.55	-0.94	-0.91	1.10	-1.40	-0.15	-1.40	-0.08	0.56	-1.20	0.81	0.83	0.42	-1.60	0.05	0.45	1.14	1.29	1.93
⊕ Evaporative Fraction	-0.70	0.07	0.59	-0.22	-1.51	-0.01	-1.44	0.16	-0.03	-0.15	0.23	1.10	-0.96	1.50	-1.66	-1.38	0.77	0.66	1.33	1.66
⊕ Latent Heat	-0.05	-0.13	-0.70	-1.12	0.21	-1.07	-0.42	-0.97	-0.62	0.96	-1.18	1.49	0.14	0.33	-1.57	-0.69	1.41	0.74	1.33	1.91
⊕ Runoff	-2.58	0.01	0.50	-0.06	-2.63	-0.37	0.58	0.48	1.04	-0.55	-0.04	0.62		-0.41	0.19	0.86	0.38	0.77	0.25	0.97
⊕ Sensible Heat	-1.10	-0.30	0.42	-0.48	-1.23	-1.38	-1.47	0.08	0.62	-1.07	0.27	1.02	0.06	0.99	-0.49	-1.05	0.48	1.08	1.56	1.98
⊕ Terrestrial Water Storage Anomaly	-1.41	-0.14	0.36	0.38	-3.86	0.29	0.30	0.34	0.42	0.19	0.06	0.62		0.34	0.31	0.27	0.32	0.39	0.43	0.40
⊕ Permafrost										-0.09	-2.52	0.69		-0.15	0.41	0.56	0.74	0.36		
<b>Radiation and Energy Cycle</b>	-0.30	-1.00	-0.24	-0.53	-1.41	-2.14	-0.07	-0.27	-0.11	-0.30	-0.09	0.48	0.80	-0.36	-0.26	0.72	-0.02	0.77	1.78	2.55
⊕ Albedo	-0.23	-0.87	0.40	-1.88	0.27	-1.42	-0.15	-0.11	1.11	-0.96	0.10	1.05	-0.93	0.65	-1.34	1.27	0.65	-0.58	1.15	1.83
⊕ Surface Upward SW Radiation	-0.49	-1.61	0.60	-1.70	-1.05	-0.97	0.15	0.38	0.47	-0.14	-1.00	0.96	-0.59	0.71	-0.11	0.73	0.70	-0.76	1.59	2.14
⊕ Surface Net SW Radiation	-0.98	-0.64	-0.61	-0.06	-1.69	-1.49	0.07	-0.27	-0.64	-0.83	0.52	0.42	1.05	-0.64	0.33	0.92	-0.34	0.89	1.65	2.33
⊕ Surface Upward LW Radiation	-0.22	-1.66	0.03	-0.57	-0.56	-1.38	-0.06	0.36	-0.59	0.11	-0.45	-0.64	0.45	-0.18	0.40	0.74	-0.97	0.43	2.26	2.52
⊕ Surface Net LW Radiation	-0.38	-1.86	-0.30	-0.21	-1.68	-1.41	-0.50	-0.42	-0.33	0.40	-0.30	0.35	0.80	-0.30	0.50	0.88	0.23	0.59	1.41	2.53
⊕ Surface Net Radiation	0.21	0.22	-0.47	0.04	-1.26	-2.72	0.06	-0.59	0.05	-0.39	0.22	0.50	1.30	-0.76	-0.94	0.10	-0.00	1.41	1.21	1.80

Overall improvements in:

- Vegetation phenology
- Biomass in the Amazon basin
- East Asian Summer Monsoon
- Cloud and Water Vapour Processes



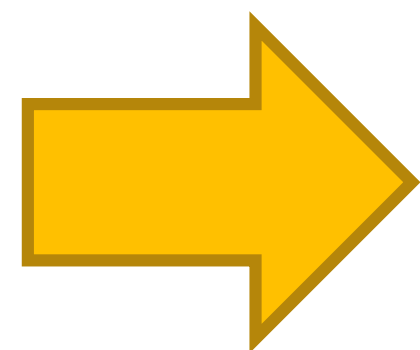
## But still: long-standing biases

- **Precipitation:** Biases in precipitation in climate models vary widely and are sensitive to the representation of clouds, radiation, surface processes, and multi-scale circulation.
- **Double ITCZ:** The double ITCZ (Intertropical Convergence Zone) is characterized by the double zonally elongated narrow belt of high precipitation in the tropics, which is present in model simulations but not in observations.
- **Warming bias in the tropical troposphere:** Excessive tropospheric warming in the tropics has been well documented in climate models.
- **Arctic sea ice sensitivity to global warming:** Arctic sea ice decline in CMIP models is lower than that observed in the satellite record, and CMIP models do underestimate the sensitivity of Arctic sea ice to global warming .



# What to consider when interpreting benchmarking results

- Choice of observation/reanalysis data (easy access, familiarity, global coverage, right format...)
- Choice of metric or diagnostics (variables or regions considered, analysed time period, consideration of seasonal aggregation...)
- Consideration of uncertainties?
- Choice of region
- Choice of application



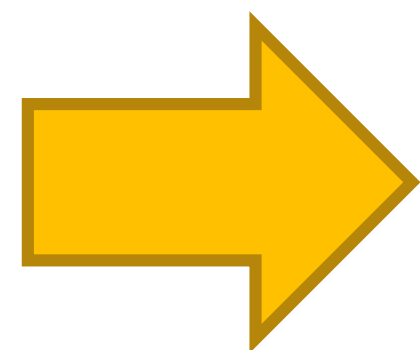
**Ranking** of models based on benchmarking must be treated with caution!



# What to consider when interpreting benchmarking results

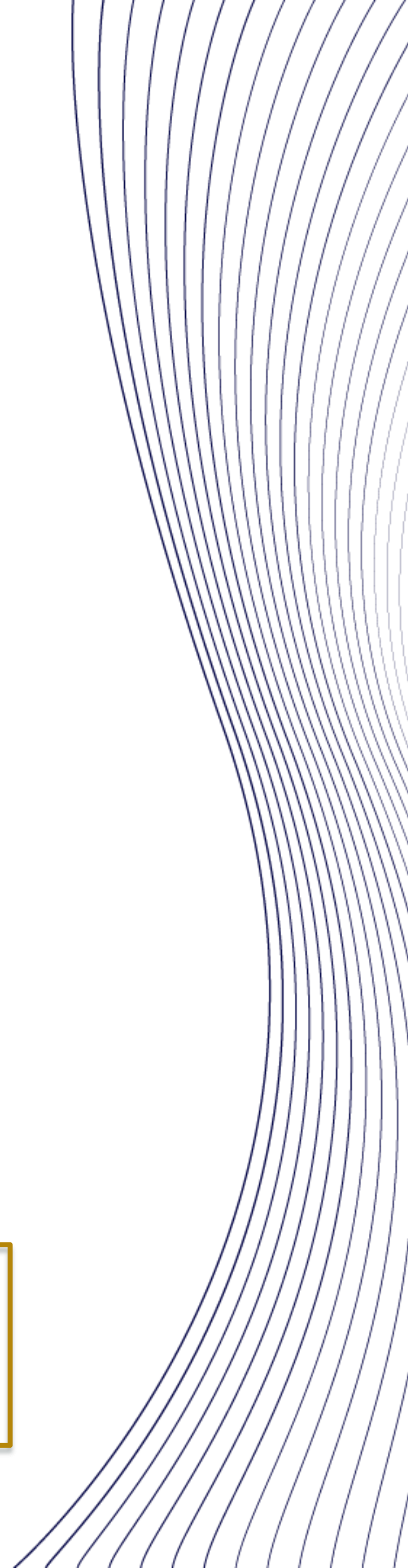
- Choice of variables
- Choice of metrics
- Choice of application

Manuscript submitted to  
Reviews of Geophysics just last  
week!



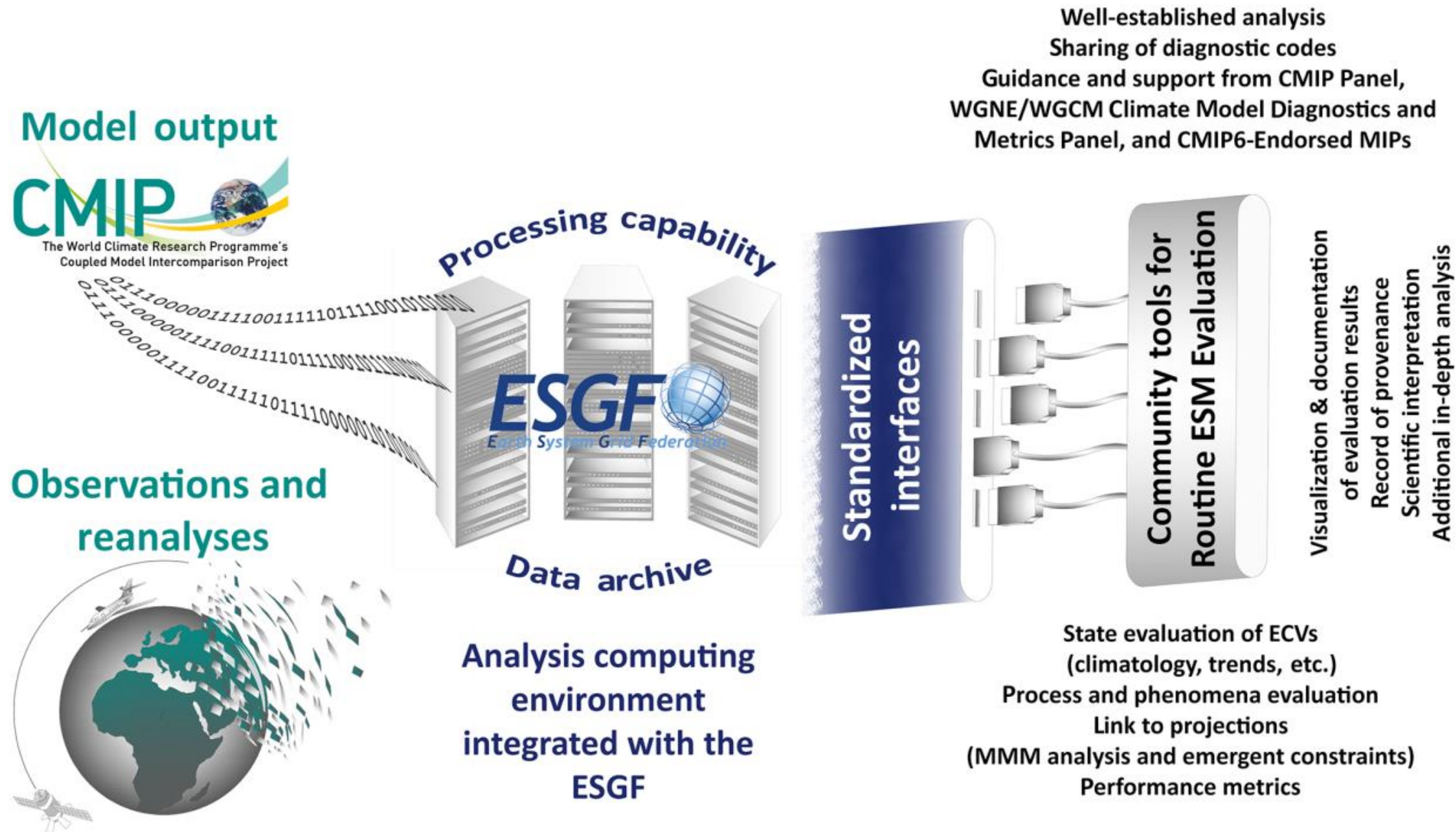
**Ranking** of models based on benchmarking must be treated with caution!

Global





# Routine evaluation for CMIP6



- Anticipated workflow for routine evaluation of CMIP6 simulations
- Did not fully work as anticipated...



# The CMIP Rapid Evaluation Framework (REF)

- Objective of the Model Benchmarking Task Team
- Conceptual design was developed at a workshop held at DLR in May 2024
- Approved by CMIP Panel in July 2024



# Consultation to date

## May 2024 Model Evaluation Survey

Three areas of focus:

1. how did respondents engage with CMIP?
2. What was their satisfaction with CMIP models & observations?
3. What are their aspirations for model attributes & benchmarking



This provided context for the May workshop where the propose for the REF was initiated.

**September 2024** Diagnostics Survey – analysed by IPO and Fresh Eyes on CMIP members

**October 2024** Observations for diagnostics session at ESA Climate Modellers Users Group Co-location meeting

**Planned:** Drop-ins for modelling centres and observation dataset providers during the REF Hackathon w/c 10th March - find out more here: <https://wcrp-cmip.org/event/ref-hackathon/>



# Overview

- **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 AR7 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.
- **Wider community use beyond CMIP:** The REF is designed to be a starting point for the community to develop and build upon, with applications across the WCRP Modelling Multiverse.
- **Diagnostics for CMIP AR7 FastTrack model evaluation:** The CMIP Model Benchmarking Task Team through consultation with the CMIP modelling community, have collected a set of informative diagnostics for five different climate model themes.

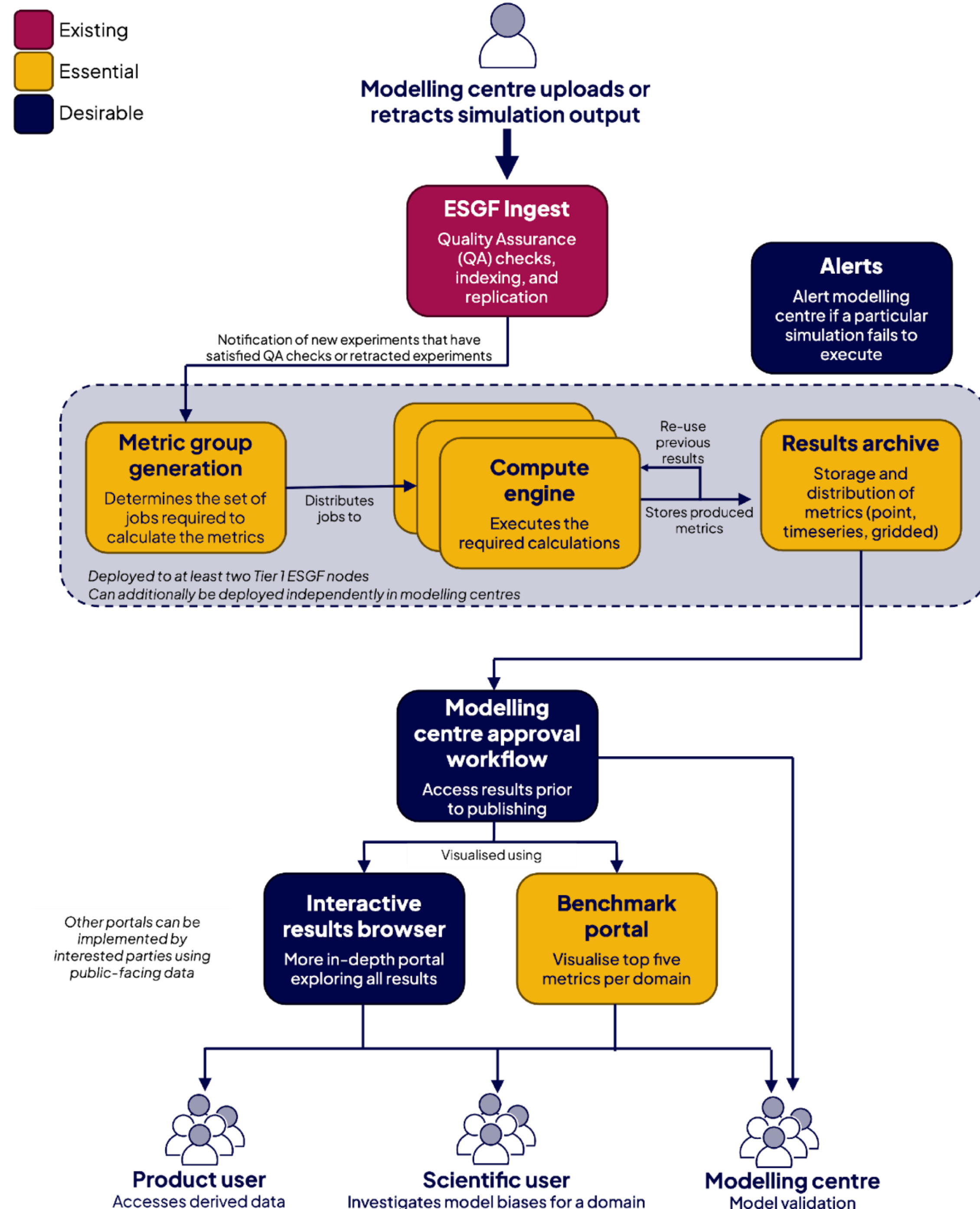


# Overview continued

The REF will be:

- **Open source**, capable of being used by, and further expanded by the model benchmarking and evaluation community
- **Open access**, available both online and through a portable containerised system for use by modelling centres in-house

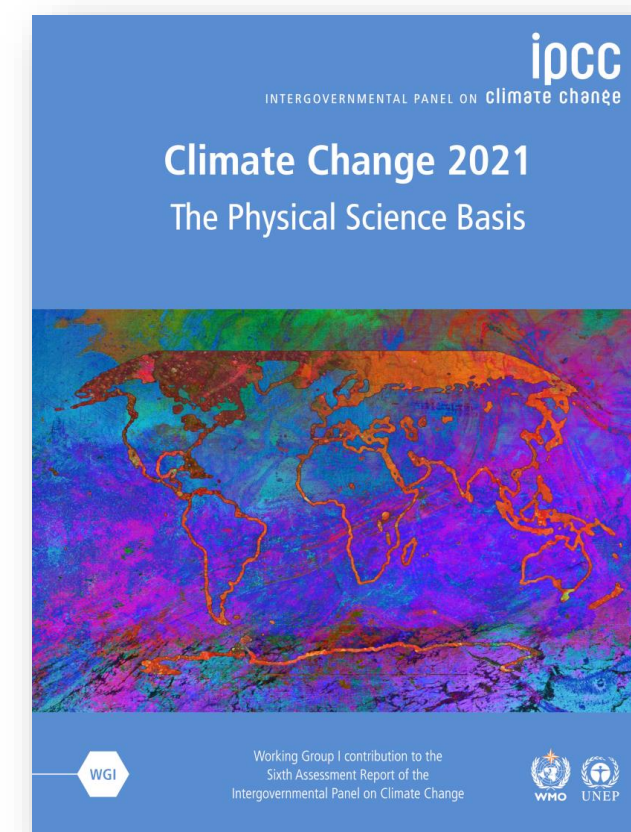
It will only be run for experiments that have satisfied QA checks via the ESGF QA process.



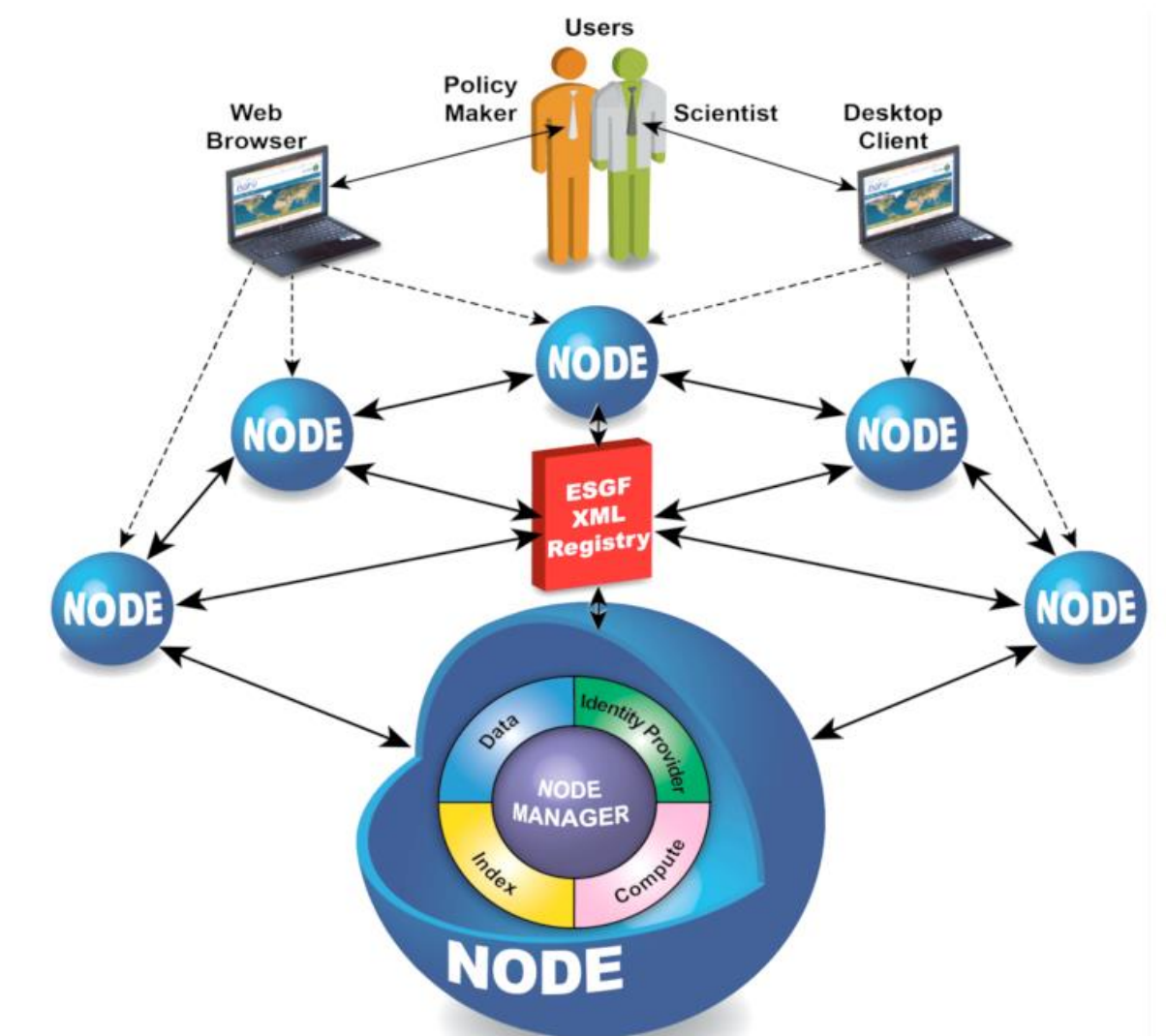


# Earth System Grid Federation (ESGF)

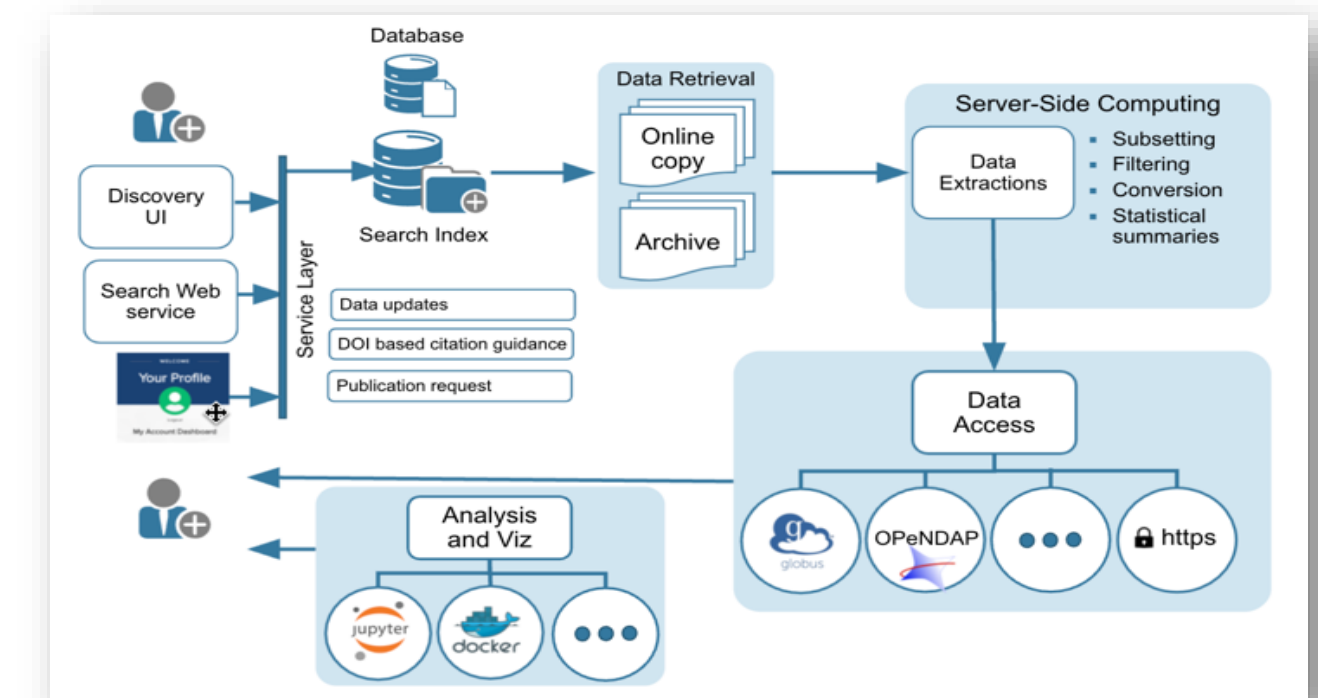
- **ESGF** is an *international consortium* and a *globally distributed peer-to-peer network* of data servers using a common set of protocols & interfaces to distribute climate model output and related input & observational data
- **Open Science data** are used by scientists all over the world to investigate consequences of climate change scenarios and feedbacks
- **Developing next generation architecture**, data discovery interfaces, server-side computing, and analysis platforms for the next set of models



## ESGF Conceptual Diagram



Model output data from ESGF are used for research that underpins IPCC Assessment Reports, like AR6

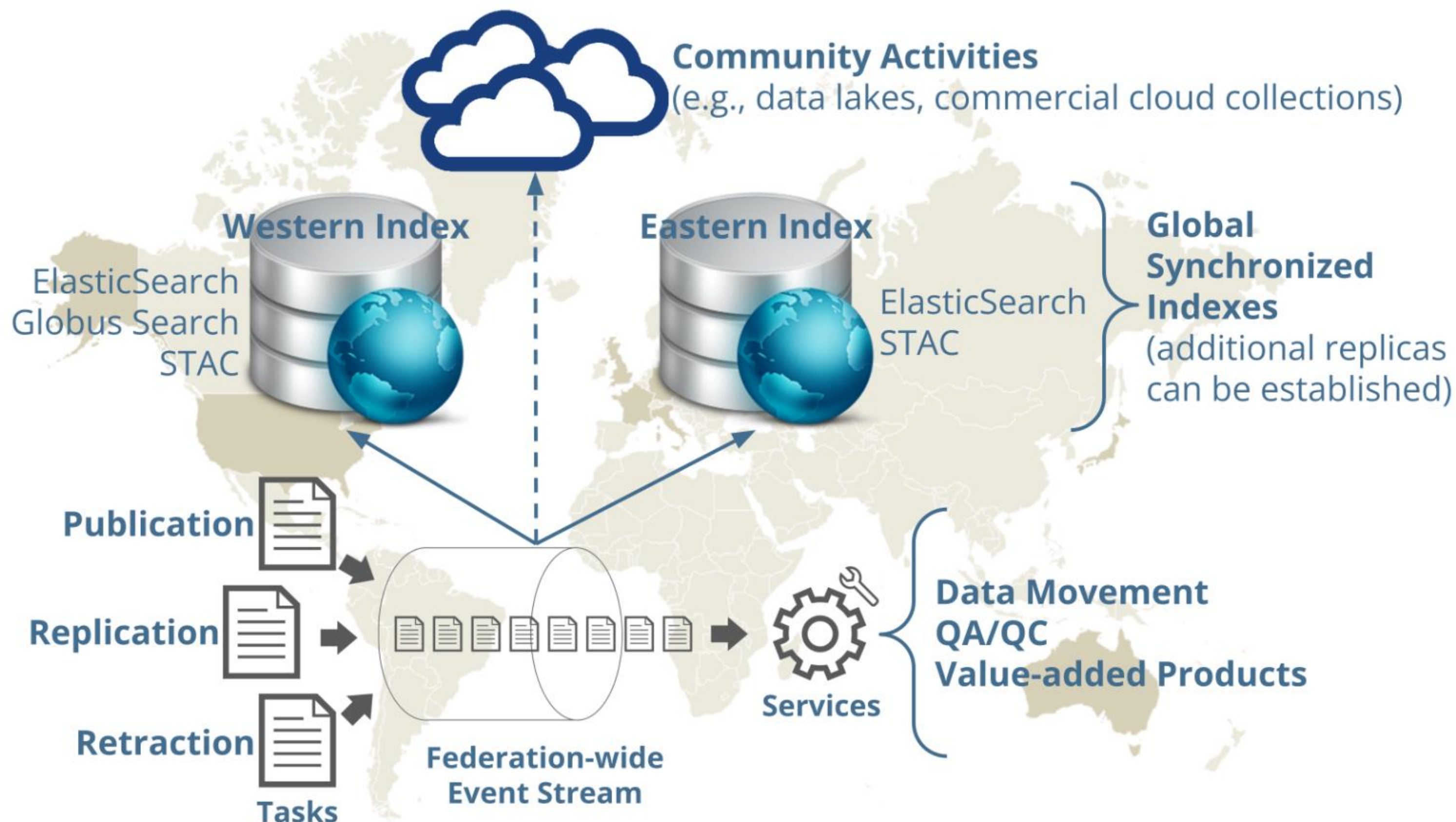




# ESGF-NG Core Architecture

The ESGF Next Generation Core Architecture will employ a *Federation-wide Event Stream* for ensuring synchronization of two global indexes.

The *Event Stream* will also facilitate automated execution of tasks and support community data activities.





# Current activities

- Funding secured for essential components for AR7
- Welcomed by the community – very positive response in the survey (September 2024)
- REF Benchmarking opportunity has been submitted to the Data Request and peer reviewed by all thematic author teams
- Diagnostics list for AR7 FT launched November 2024
- Community are encouraged to consider incorporating REF add-ons for enhancements in funding proposals.
- Expressions of interest open for Tier 1 nodes, reference data providers, modelling centres and scientists



# Timeline



# Milestones for REF development

- Milestone 1 – Community engagement to finalize the implemented metrics and diagnostics in a minimal version of the REF – **completed, launched 4 November**
- Milestone 2 - Provide recommendations for enhanced QA/QC package(s) **In progress - REF delivery team working with a dedicated ESGF-WIP QA QC Working Group**
- Milestone 3 - Prototype workflow across at least two participating ESGF nodes and test containerised version with at least 3 modelling centres **In progress**
- Milestone 4– Governance and operational structure for future evolution presented to and approved by WCRP JSC
- Milestone 5– Publication (peer reviewed) after the REF is in place and tested

- March 2025 Prototype operational and undergoes stress testing during hackathon.
- May 2025- Beta portal available for early adopter modelling centres
- October 2025- Public release of the metrics portal and REF for community evolution.

**Keep up to date with the latest REF progress by looking at the roadmap:**  
<https://cmip-ref.readthedocs.io/en/latest/roadmap/>







# Work packages



# Work packages

- CMIP IPO
- DOE
- NLeSC (Netherlands eScience Center)
- CR (Climate Resource)
- MBTT (Model Benchmarking Task Team)

**RACI**  
 R: Responsible (Does the work)  
 A: Accountable (Ultimately responsible)  
 C: Consulted (Provides input)  
 I: Informed (Kept in the loop)

In the MVP, WP1 would develop the engine to execute metric calculations

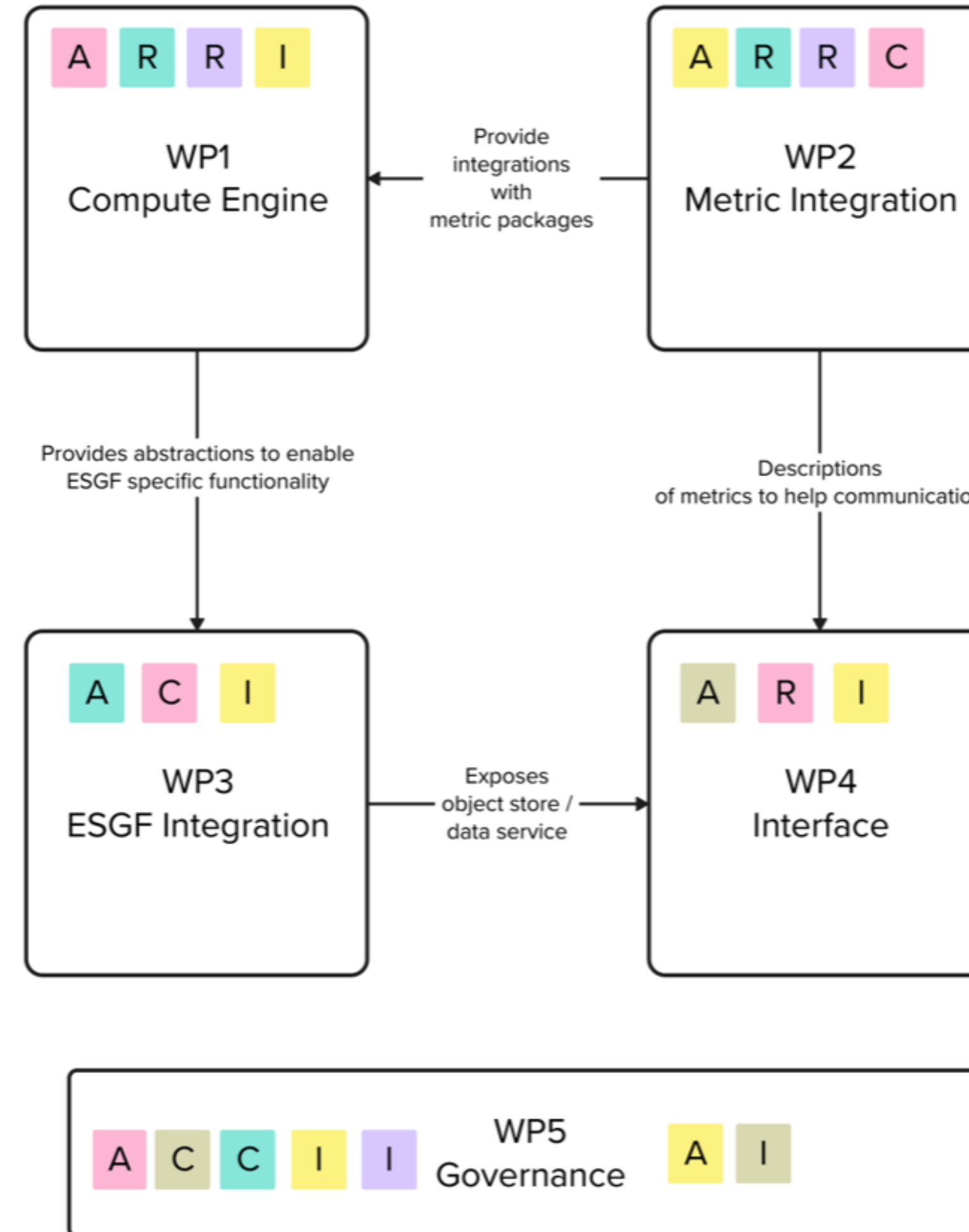
- Development of Open Source package that can update
- Provides an abstraction for running tasks that enables multiple target environments
- Includes a PoC local runner for testing/modelling centres

In the MVP, WP3 would createan ESGF deployment of the REF

- Development of required configuration for deployment to ESGF (K8 Helm chart/Ansible deployment)
- Production deployment of workers, event queue, object store and job database

Technical Director: Jared Lewis

Oversee the technical implementation of the REF  
 Define specification document that describes the metric calculation  
 Reports to the Model Benchmarking TT



In the MVP, WP2 would define howthe metrics are calculated

- Develop metric integrations for key packages using the CMEC driver
- Develop integration for timeseries calculations
- Define the CMIP AR7 Fast Track configuration
- Provide assistance packaging metrics into job environment

In the MVP, WP4 would build an portal to interrogate the results

- Benchmark portal for a select set of metrics
- Built on top of results served from ESGF

Science: Birgit Hassler/Forrest Hoffman in coordination with the CMIP MBTT

- Define of the user requirements of the REF (done)
- Decide on the list of metrics to implement as part of CMIP AR7 Fast Track (In progress)



# With thanks to everyone making the REF possible:

- Acknowledgement of funders

Being delivered by:





# Thank You



@wcrpcmip



wcrp-cmip



cmip-ipo@esa.int