Reviews of Geophysics®

REVIEW ARTICLE

10.1029/2025RG000891

Systematic Benchmarking of Climate Models: Methodologies, Applications, and New Directions

Birgit Hassler[1], Forrest M. Hoffman[2], Rebecca Beadling[3], Ed Blockley[4], Bo Huang[5], Jiwoo Lee[6], Valerio Lembo[7], Jared Lewis[8], Jianhua Lu[9], Luke Madaus[10], Elizaveta Malinina[11], Brian Medeiros[12], Wilfried Pokam[13], Enrico Scoccimarro[14], and Ranjini Swaminathan[15]

[1]Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany, [2]Oak Ridge National Laboratory (ORNL), Oak Ridge, TN, USA, [3]Department of Earth and Environmental Science, Temple University, Philadelphia, PA, USA, [4]Met Office, Exeter, UK, [5]Norwegian University of Science and Technology, Trondheim, Norway, [6]Lawrence Livermore National Laboratory (LLNL), Livermore, CA, USA, [7]Institute for Atmospheric Science and Climate, National Research Council of Italy (CNR-ISAC), Rome, Italy, [8]Climate Resource Pty Ltd, Melbourne, VIC, Australia, [9]School of Atmospheric Sciences, Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Sun Yat-Sen University, Zhuhai, China, [10]Jupiter Intelligence, Inc., San Mateo, CA, USA, [11]Canadian Centre for Climate Modelling and Analysis (CCCma), Environment and Climate Change Canada (ECCC), Victoria, BC, Canada, [12]NSF National Center for Atmospheric Research, Boulder, CO, USA, [13]University of Yaoundé I, Yaoundé, Cameroon, [14]CMCC Foundation - Euro-Mediterranean Center on Climate Change, Bologna, Italy, [15]Department of Meteorology, University of Reading, Reading, UK

Key Points:

- This review provides a description of the history and contemporary philosophy of climate model evaluation and benchmarking
- Key features of commonly used open-source community benchmarking software packages are presented
- Observational reference data, integrating uncertainty quantification and a smart selection of metrics are vital for meaningful evaluations

Correspondence to:

B. Hassler and F. M. Hoffman,
birgit.hassler@dlr.de;
hoffmanfm@ornl.gov

Citation:

Hassler, B., Hoffman, F. M., Beadling, R., Blockley, E., Huang, B., Lee, J., et al. (2026). Systematic benchmarking of climate models: Methodologies, applications, and new directions. *Reviews of Geophysics*, *64*, e2025RG000891. https://doi.org/10.1029/2025RG000891

Received 12 MAR 2025
Accepted 27 DEC 2025

Abstract

As climate models become increasingly complex, there is a growing need to comprehensively and systematically assess model performance with respect to observations. Given the increasing number and diversity of climate model simulations in use, the community has moved beyond simple model intercomparison and toward developing methods capable of benchmarking a large number of simulations against a suite of climate metrics. Here, we present a detailed review of evaluation and benchmarking methods and approaches developed in the last decade, focusing primarily on scientific implications for Coupled Model Intercomparison Project (CMIP) simulations and CMIP6 results that contributed to the Intergovernmental Panel on Climate Change (IPCC) Sixth Assessment Report (AR6). Based on this review, we explain the resulting contemporary philosophy of model benchmarking, and provide clear distinctions and definitions of the terms model verification, process validation, evaluation, and benchmarking. While significant progress has been made in model development based on systematic evaluation and benchmarking efforts, some climate system biases still remain. The development of open-source community software packages has played a fundamental role in identifying areas of significant model improvement and bias reduction. We review the key features of several software packages that have been commonly used over the past decade to evaluate and benchmark global and regional climate models. Additionally, we discuss best practices for the selection of evaluation and benchmarking metrics and for interpreting the obtained results, the importance of selecting suitable sources of reference data and accurate uncertainty quantification.

Plain Language Summary

Global and regional climate models are increasingly becoming more advanced and complex. Observational data used to assess the ability of models to reproduce realistic climate is becoming more diverse and available over longer time spans. Both of these factors make the comparison of observations with model data a complex task. Analysis methods or diagnostics are developed to evaluate and benchmark model data both with and without observations. These diagnostics have shown many improvements in the abilities of the current generation of models to reproduce the current climate, but some problems remain in which model output and observations do not agree well. In order to identify these areas of model-observation disagreement more efficiently, scientists have developed software packages that are freely available to the global community. The characteristics of some of the most commonly used packages are described here in addition to best practices on how to interpret the results obtained. We explain why the choice of observations and the selection of appropriate diagnostics is vital to obtain meaningful results. Climate models are constantly evolving and improving; therefore, it is critical that evaluation frameworks also continue to advance concurrently.

1. Introduction

Earth's climate is the result of intricate and rich interactions between the system's realms: atmosphere, biosphere, ice, land, and ocean. One cannot fully understand the physical and biogeochemical dynamics that give rise to the mean-state climate, its temporal and spatial variability, and its response to past, present, and future perturbations without considering the system in its entirety. The rise in atmospheric greenhouse gas concentrations, land-use change, and changes in atmospheric aerosols associated with post-1850 global industrialization represent a significant perturbation to Earth's climate system. The vast scale and coupled nature of Earth's climate means that the full impact of such anthropogenic perturbations cannot be studied by traditional laboratory methods but must be studied using tools that encompass the full system. Developing a detailed understanding of how climate change is evolving and will evolve throughout the rest of the 21st century and beyond is vital to societal decision-making, including the development of appropriate climate adaptation and mitigation strategies.

General Circulation Models developed in the late 1960s (Edwards, 2011; Manabe & Bryan, 1969; S. H. Schneider & Dickinson, 1974) provide the tools necessary to probe climate dynamics and the climate system response to various perturbations. These numerical simulations, known as climate models, are rooted in fundamental scientific principles that represent the physical processes and exchanges of energy and matter between the atmosphere, sea ice, land and ocean. While these "physical" ("physics-based") climate models provide the means to understand the physics of the system, over the past several decades as the field of climate modeling has advanced, more comprehensive configurations, known as Earth System Models (ESMs) have been adopted by climate modeling centers. In addition to representing the physical climate system (being "physics-based"), ESMs include additional processes such as interactive atmospheric composition, biogeochemistry and carbon exchange between Earth system components. Both physical and ESM-based climate models provide the virtual infrastructure to develop a comprehensive understanding of climate dynamics, serving as the primary tools for understanding the climate system, its variability, and its response to anthropogenic forcing. In this manuscript we focus our discussions on climate models used for the World Climate Research Programme's (WCRP) Coupled Model Intercomparison Project (CMIP; Meehl et al., 1997). These include both physical climate models and climate models based upon ESMs. For easier understanding we will refer to both model types as "climate models" from hereon.

Output from climate models has become central to the Intergovernmental Panel on Climate Change (IPCC) assessment reports, and model-derived future projections form the backbone for the development of climate adaptation and mitigation policies. Given their role in political and societal decision-making, the need to systematically assess the simulation realism of climate models is paramount. As climate science and the field of climate modeling advanced, recognition of this need led to the development of CMIP; providing the framework necessary for systematic evaluation and comparison of model simulations, ushering in a new era in climate change research (Meehl et al., 2023). Initiated in 1995, and now entering its seventh phase (CMIP7; Dunne et al., 2025), the results from CMIP have become inextricably linked to the IPCC assessment reports (Meehl, 2023) and have led to significant advancements in climate science.

CMIP has evolved from a single experiment performed by 21 global climate models in CMIP1 (Lambert & Boer, 2001; Meehl et al., 1997), to an enormous coordinated international effort involving over 100 different climate models, hundreds of experiments, and the public archiving of over 15 petabytes of model output (Eyring et al., 2021; Eyring, Gleckler, et al., 2016). The "historical" experiment introduced in CMIP5-6 (Eyring, Bony, et al., 2016; Taylor et al., 2012), in CMIP3 known as "climate of the 20th Century experiment (20C3M)" (Meehl et al., 2023), provides the framework necessary to appropriately assess model performance relative to the real world given that the prescribed forcing is designed to be as consistent as possible with observed atmospheric composition changes and time-evolving land cover (in CMIP5 & beyond). Small differences in external forcings can have a considerable impact on climate model simulations (e.g., Holland et al., 2024) and thus the specifications of prescribed forcing in the historical simulations are critical when evaluating model fidelity across multiple models. For a more detailed description of CMIP history see Durack et al. (2025).

The historical simulation is one of the common experiments that must be completed as part of the "entry card" for models to participate in CMIP6 or other organized Model Intercomparison Projects (MIP) endorsed by CMIP (CMIP-Endorsed MIPs; Eyring, Bony, et al., 2016), e.g. the Aerosols and Chemistry Model Intercomparison Project (AerChemMIP, Collins et al., 2017), the Detection and Attribution Model Intercomparison Project (DAMIP, Gillett et al., 2016) and the Cloud Feedback Model Intercomparison Project (CFMIP, Webb et al., 2017). The historical simulation is also critical as it provides the branch point for additional experiments aimed to provide projections of future climate throughout the 21st century and beyond based on scenarios of future energy consumption and carbon emissions associated with societal change (ScenarioMIP; O'Neill et al., 2016). The other experiments required for entry into CMIP6 are known collectively as the Diagnostic, Evaluation and Characterization of Klima (DECK) experiments. These include idealized experiments, such as an instantaneous step to four times pre-industrial CO_2 levels and a transient increase of 1% CO_2 per year, that are designed to understand the system response to external forcings, rather than to reproduce observed changes in climate.

Leveraging output from historical (or 20th Century) CMIP simulations, climate model evaluation and benchmarking, originally undertaken "in-house" at individual modeling centers, has evolved into an important field of its own involving international coordination, thousands of scientific publications, and has driven the development of new research centers, software capabilities, and infrastructure to facilitate rapid assessment of model performance (Eyring et al., 2019; Eyring, Bony, et al., 2016; Eyring, Gleckler, et al., 2016; Gleckler et al., 2016; Neelin et al., 2023; Waliser et al., 2020). Such efforts are driven by the need to document and understand systematic biases present in model simulations in order to steer model development efforts toward improved simulations and increase confidence in model results, particularly increase confidence in model-derived future projections. The topic of climate model evaluation has also become a prominent feature of IPCC reports (Flato et al., 2013) given the focus on climate projections providing estimates of near- and long-term climate change. Such projections are subsequently used to understand potential climate change impacts at both global and regional scales.

In order for the user community to make efficient use of the large number of model simulations available within CMIP, there is a need for benchmarking of model outputs to understand the relative strengths and weaknesses of the available simulations. This benchmarking is not about ranking models or finding the "best" model, but about ensuring that consistent information is available, allowing users to make informed decisions about which model-derived products they should use and for what purpose. Such an approach recognizes the fact that the suitability of a model simulation is very much dependent on the purpose that one intends to use it for—including the regions and variables of interest and the specific question(s) that the user is trying to address.

As CMIP rapidly grew from CMIP5 into CMIP6—both in the complexity and number of participating models and experiments, and in the volume of model output made available for community distribution—the urgent need to develop a framework to allow model evaluation and benchmarking to be performed more efficiently and routinely (Eyring, Gleckler, et al., 2016) was recognized. It was noted that there must be efforts to transition from individual ad-hoc (and often "in-house") model evaluation efforts to a more community-coordinated approach leveraging available computational hardware already integrated into the CMIP workflow (i.e., the Earth System Grid Federation (ESGF) system; D. Williams, 2015; Petrie et al., 2021). The community was encouraged to contribute diagnostic codes, observations, and observation-based products to on-going efforts including CREATE-IP (Potter et al. (2018), previously ana4MIPs) and obs4MIPs (Teixeira et al., 2014), which could aid in routine and rapid model evaluation, all while leveraging existing ESGF-node infrastructure. The exact details of the vision presented in Eyring et al. (2019) (referred to hereafter as EY19) was successful in many parts but ultimately was not realized in the complete manner originally outlined. The development of individual evaluation and benchmarking tools progressed tremendously from CMIP5 to CMIP6, more contributions from those tools were included in the latest IPCC report: a code repository for IPCC figures was created (IPCC-WG1, 2023) and also code quality controls were put in place. However, the ability to process model output automatically alongside ESGF to rapidly and routinely evaluate CMIP6 output as the simulations came online could not be fully accomplished. This quasi-operational evaluation framework fell short not due to technical plausibility, but due to complexities associated with the data quality of the submitted simulations. Deviations from the pre-defined, strict structure of file format and metadata information, even apparently trivial ones, meant that implementation of fully automated and rapid evaluation directly after data publication was not possible for CMIP6. Furthermore, although all modeling centers had agreed on providing their simulations in this pre-defined format, they were not all able to do so. While the exact implementation did not pan out, the EY19 vision ushered in dedicated efforts and increased motivation to develop community-oriented diagnostic tools to aid in rapid and routine model evaluation efforts. EY19's call to action to start the transition from individual ad hoc efforts to community-driven, community-oriented, and open-source model evaluation tools, was realized.

This paper provides a current and retrospective overview of the state-of-play for climate model evaluation and benchmarking efforts, the application of such efforts to the scientific and broader community, and a discussion on the advances and growing challenges. Additionally, a number of open-source evaluation and benchmarking tools that were used for the evaluation of CMIP6 simulations are described and characterized. Many of the tools described here have been developed and significantly advanced since EY19, indicating the recognition by the community of the need for realizing the goal of routine, rapid, and robust climate model evaluation. We do not provide a detailed follow-on vision to EY19 here but instead provide a comprehensive retrospective overview of evaluation and benchmarking efforts which are now available to be leveraged. We first present a discussion of the terminology and philosophy of model evaluation and benchmarking (Section 2) before following with a comprehensive overview of available tools and community-driven efforts (Section 3). Applications of model benchmarking as part of the model development process and within the growing community of external end users are discussed (Section 4). Improvements in reducing long-standing model biases have been achieved by systematically evaluating models for different model generations (see Section 5). However, some of the long-standing model biases persist even though the community has worked on reducing them. Examples of these persisting biases are given in Section 6. In Section 7 the prerequisites of observations and their uncertainties are briefly outlined since they play a very important role in every evaluation and benchmarking effort. Finally, in Section 8 we summarize the provided definitions, tool characteristics and availabilities, possible applications, and indicate where the advancements of model evaluation and benchmarking could aid rapid model evaluation efforts in CMIP7.

2. Philosophy of Model Evaluation and Benchmarking

Traditionally, approaches for evaluation, benchmarking and assessment of Earth system models have centered on fidelity to observed phenomena or comparative performance in their ability to accurately model physical processes (i.e., a model accurately captures modes of variability). Broadly speaking, model evaluation approaches are designed to help us understand qualitatively, or more often quantitatively, uncertainty in model simulations related to the following aspects:

- **Internal Variability**: the climate system's internal fluctuations, reflecting the fact that the modeled system is intrinsically chaotic.
- **Model Structure**: different model formulations (process-representation in and structure of the model) arising due to factors including but not limited to different grid resolutions, vertical coordinates or parameterization choices, different levels of interactivity for specific modules (e.g., fixed or interactive ice sheets, aerosols, etc.), and perturbations of parameters that go on to determine differences in model behavior.
- **Boundary Conditions**: uncertainties in the external data used to drive the models such as forcings or orography.
- **Future Projections**: specification of scenarios for the future including future forcings, emissions and socio-economic pathways will always be a key source of uncertainty (Lehner et al., 2020).

Here, we attempt to define key terms relating to climate model output analysis and climate model evaluation by systematically framing the purpose, outcomes and limitations of such analyses. It should be noted that these terms are not universally applicable across the atmospheric sciences. There are different terminologies in use for numerical weather predictions or decadal predictions that do not fully conform with the definitions introduced here. Additionally, the definitions below are not entirely exclusive, but can overlap with some aspects of one of the other terms (see e.g. the definition of "Benchmarking"). However, all four key terms are important enough for climate model analyses and assessments that they deserve their own definition:

- **Model Verification**: the process of assessing model consistency in terms of correct implementation of the included processes as articulated in the model and experimental design. This is a basic first step to ensure code translates correctly to simulations by adhering to the basic physical, chemical, and biological principles such as Newton's laws or the laws of thermodynamics (in both atmosphere and ocean). Sometimes, model verification is performed as the model simulations are being produced (such as monitoring the conservation of total energy, total atmospheric mass, etc.), and the focus is often on the artifacts introduced by the numerical discretization scheme (e.g., Lauritzen et al., 2022) or by changes to software or hardware used for the simulations (e.g., the Ensemble Consistency Test in A. H. Baker et al. (2015) and the Time Step Consistency Test in Wan et al. (2017));

Table 1
Characteristics of Validation, Evaluation and Benchmarking of Climate Model Simulations

Feature	Process validation	Evaluation	Benchmarking
Can it be quantitative?	Yes	Yes	Yes
Can it be qualitative?	Yes	Yes	Yes
Can it include observations?	Maybe	Yes	Yes
Must it include observations?	No	Mostly	No
Can it determine fitness for a purpose?	Yes	Yes	Yes
Can it be used for future simulations?	Yes	No	No
Is it scale dependent (global, regional etc.)?	No	No	No
Can it be used in a multi-model context?	No	Yes	Yes
Is it realm-specific?	No	No	No
Is it experiment-dependent?	No	Yes	Maybe
Is it suitable for monitoring during model development?	Yes	Yes	No
Is it used for impacts assessments or policy formulation?	No	Maybe	Yes
Can it involve more than one model components or domains?	Yes	Yes	Mostly not
Can it include performance metrics?	No	Maybe	Yes
Can it be used for process diagnostics?	Yes	Yes	Indirectly

- **Process Validation**: the process of determining how well a model represents processes in the real world, particularly for the intended uses of the model. Process Validation therefore goes beyond Model Verification where the focus is to check if a model captures (e.g., the physical or biogeochemical) processes as we encode them, but not necessarily as they are in the real world. Process validation can include a broad range of aspects from ensuring correct units and sign of the data produced, to the interactions between model components or variables and process representations, and may or may not include observations.
- **Evaluation**: the process of assessing simulations against one or more observational data sets. The necessity for observations means evaluation can only be done for the historical period, and only for variables or processes for which observations or reanalysis data are available. Model evaluation can be done for a single model or in a multi-model context. Incomplete observational records, including limited time series length, unobserved variables, biases due to specific instruments, and uncertainties in spatial and temporal coverage can make evaluation challenging for certain processes and realms of the climate system that are under-observed.
- **Benchmarking**: the process where model simulations are evaluated with observations, reanalysis data or with other models often resulting in a statement made about the "goodness" of the simulation or model based on a predetermined set of standards or criteria (e.g., observations or other standards).The evaluation process normally occurs whenever new schemes are added to the model, but model benchmarking occurs only after all the pieces are assembled and climate simulations are produced.

Note that the term "assessment" is used in this manuscript from time to time. It describes mostly more generally the activities of Process Validation, Evaluation and Benchmarking, that is the process of looking at a model simulation in more detail to be able to decide if it is suitable for the intended purpose ("performance assessment"). An assessment can therefore include one or more of the aforementioned activities.

Since the terms validation, evaluation and benchmarking are used most often, and probably also most easily, confused in their usage, we provide characteristics of these terms in Table 1. This overview aims to clarify differences in their characteristics.

Our definitions, schematically shown and explained in Figure 1, are similar in some respects but also differ slightly from prevailing terminology for model evaluation and benchmarking (Best et al., 2015; Grewe et al., 2012; Luo et al., 2012), notably in the inclusion of observations for assessing simulations. However, we note that even if a model performs credibly when compared to observations, it may not provide reliable future projections (McAvaney et al., 2001; Notz, 2015). We also need to ensure that the model's response to perturbations remains credible, therefore comparisons with observations need to be made in the context of internal variability

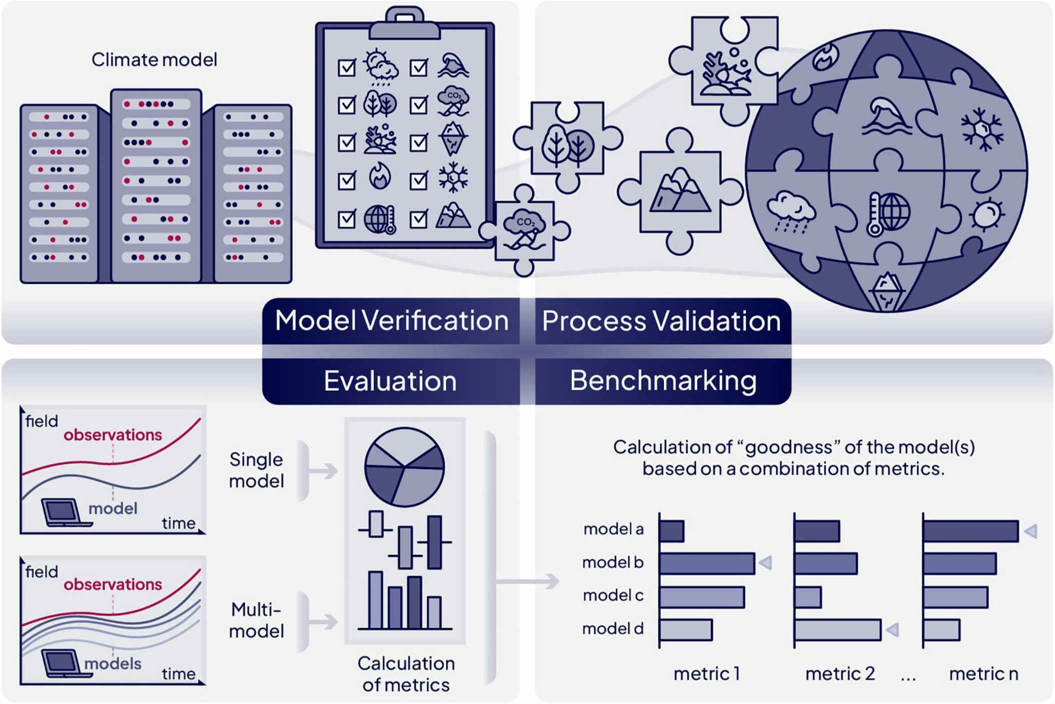


Figure 1. Schematic definition of the terms Model Verification, Process Validation, Evaluation and Benchmarking for use in the climate model context. Note that although some form of ranking can be performed during benchmarking, based on the chosen metric and selected observations, this ranking is generally not valid for all metrics, all realms and all possible observational references. Thus, it is important to realize that a "high" ranking of one model for a given comparison does not mean that this model performs well in other climate realms (Hassler et al., 2025).

and observational uncertainty. Furthermore, a model that performs better on a specific metric or process, may underperform with other metrics. The inferred model skill, and ultimately the determined rank, is highly variable and uncertain, and depends not just on the chosen metric or evaluated realm, but also very strongly on the chosen observational data set (Schwalm et al. (2013); see more details about the importance of observations in model evaluation and benchmarking in Section 7). No individual evaluation technique or performance measure can therefore be considered superior. It is rather the combined use of many techniques, performance measures and observations that provides a comprehensive overview of model performance (Flato et al., 2013).

This implies that even scores obtained through benchmarking models have to be correctly interpreted (Knutti & Rugenstein, 2015) and that peculiarities of individual model uncertainties must be addressed. Compensating errors must also be taken into account since they can be masking issues specific to individual processes. The analysis of model performance should be guided by the intended purpose, as exemplified in the treatment of uncertainty on Equilibrium Climate Sensitivity discussed in Chapter 4 of the IPCC 6th Assessment Report (J.-Y. Lee et al., 2021). We envisage that Table 1 will serve as a guide that enables model developers, end users and the wider scientific community to understand the advantages and limitations of these methods and make the best choice for their desired objectives.

Based on the above discussion, we recommend that model evaluation and benchmarking efforts or frameworks include but are not limited to the following:

1. An evaluation of key variables simulated by models with standardized observations and reanalysis data or previous model runs (e.g., CMIP6). This involves community participation in identifying such key variables.
2. An evaluation of whether fundamental processes in the Earth System are adequately represented in models, that is, represented well enough that the resulting simulations can be trusted as being realistic and "fit for purpose." This involves both identifying the processes of interest and developing metrics that appropriately assess their representation.
3. A standard set of performance metrics and diagnostics to facilitate (1) and (2).

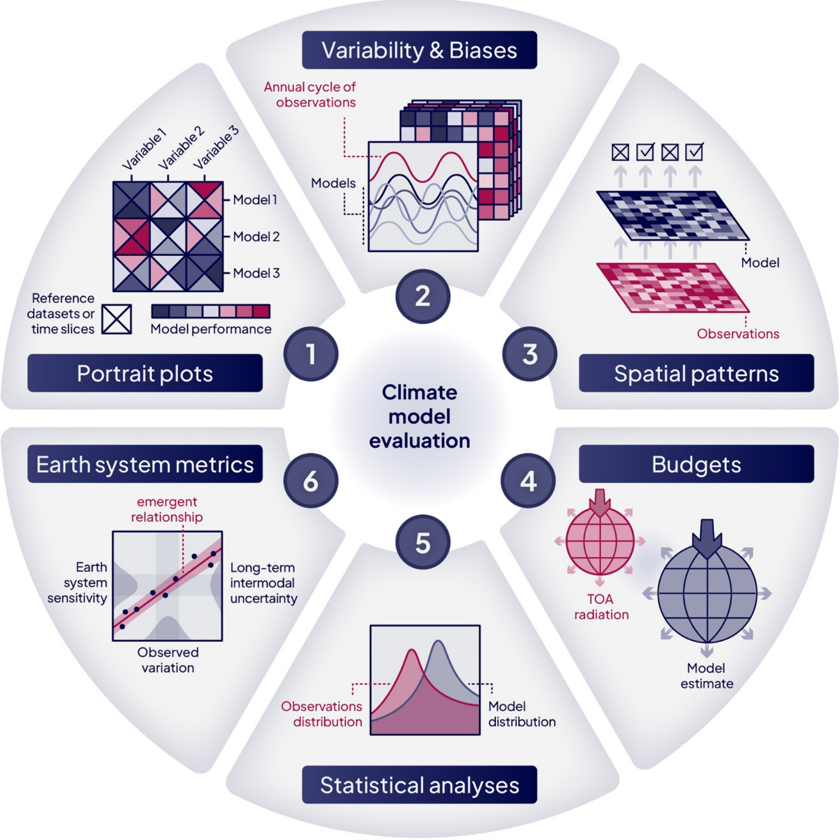


Figure 2. Different approaches that are most commonly used for the evaluation and benchmarking of climate models. Most of the approaches can be applied to different realms (e.g., atmosphere, ocean, land and land ice, ocean and sea ice), and each approach can include more than one diagnostic or metric. Approach 1 uses the so-called Portrait Diagram to simultaneously display several performance metrics (see Section 2), which is very versatile in its application across different domains, analyzed variables and number of included observations or time periods. Approach 2 represents all diagnostics that are based on analyses of biases and variability. Approach 3 includes all diagnostics that focus on spatial analyses, for example spatial correlations or physical connections between neighboring regions/realms. Approach 4 includes any budget assessments. These diagnostics are commonly applied globally, but can also be applied regionally if boundary conditions and fluxes across boundaries are clearly defined. Approach 5 represents all other statistical approaches for model evaluation, for example the analyses of distributions. Approach 6 finally includes all diagnostics that aim for describing Earth system and its interconnections and changes as a whole, for example emergent constraints or equilibrium climate sensitivity (ECS) (Lembo et al., 2024).

4. Observational, reanalyses, satellite or experimental data products to facilitate (1) and (2).
5. Appropriate software, hardware and data infrastructure specifications to facilitate (1) and (2). This will require addressing issues related to hardware (such as memory capacity, GPU capabilities, speed, and disk storage requirements), software (such as handling unstructured grids, licensing restrictions for evaluation tools), data (such as licensing, documentation, missing data or uncertainties in measured quantities).
6. Flexibility to incorporate new scientific targets easily, for example evaluate Machine Learning (ML)/Artificial Intelligence (AI) based model development or analyzing tipping points or overshoot scenarios.
7. Capability to address differences in simulations such as forced versus unforced simulations and prescribed versus concentration/emission-driven simulations.

Figure 2 shows different approaches that are commonly used for evaluating and benchmarking climate model simulations. Each approach represents an overarching evaluation topic (e.g., evaluation metrics) and can often be applied to all different realms of a climate model (e.g., atmospheric parameters, land parameters, etc.) and on global and regional scales.

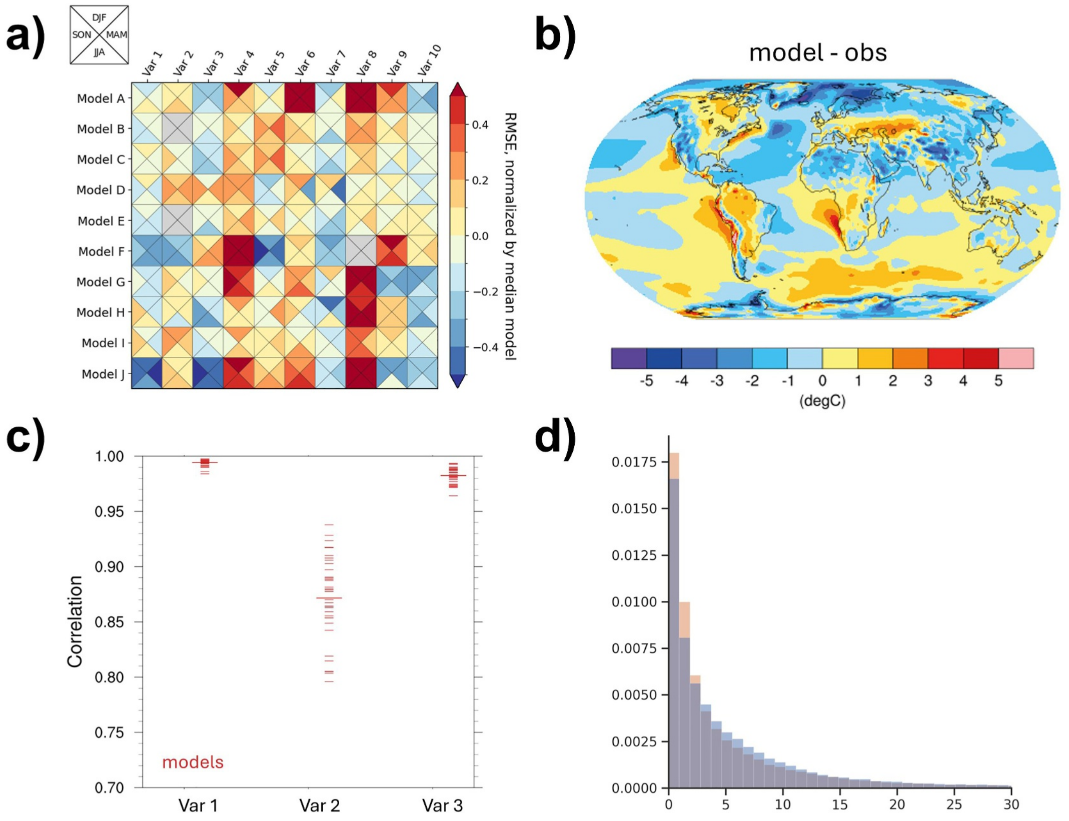


Figure 3. Examples of metrics and diagnostics for different evaluation and benchmarking approaches. (a) Example of a Portrait Diagram (Approach 1 in Figure 2), (b) example of a bias analysis (Approach 2 in Figure 2), (c) example of a pattern correlation analysis (Approach 3 in Figure 2), and (d) example of a distribution (Approach 5 in Figure 2). Examples generated using PCMDI Metrics Package (a) and ESMValTool (b–d), see Table A1 (Hassler, 2025).

Portrait Diagrams (see Approach 1 in Figure 2), often also referred to as performance metric plots, condense the evaluation of different variables down to a few numbers displayed as colored squares or triangles. They are metrics used to provide information about relative performance of different models and were originally applied to atmospheric variables only (Gleckler et al., 2008). Since their beginnings Portrait Diagrams have become a popular visualization for different realms and are included in many evaluation studies and evaluation and benchmarking tools (e.g., Bock et al., 2020; Collier et al., 2018; Eyring et al., 2021; J. Lee et al., 2024). Portrait Diagrams usually show the Root Mean Square Error (RMSE) calculated for different variables in comparison to observational data sets. The RMSE is normalized by the median of all models, and the magnitude of the RMSE is indicated by the color in the plot (see Figure 3a). Each triangle can then represent the RMSE based on different observational data sets or different seasons of the same variable and observation pairing.

Approach 2 in Figure 2 represents a wide variety of metrics and diagnostics that are based on first-order characteristics of variables, like their means (e.g., Stevenson et al., 2020; Tsujino et al., 2020), climatologies (e.g., Eman et al., 2024; Huang et al., 2020), variabilities and biases (e.g., J. C. A. Baker & Spracklen, 2022; Cesana et al., 2023; Donohoe et al., 2024; Hsu et al., 2021; Stevenson et al., 2020; Q. Zhang et al., 2023). We think of them as first-order since right after the general overview of a Portrait Diagram, these would be the analyses that are performed to understand the characteristics of simulations. The metrics and diagnostics of this approach are very versatile in their application and can help evaluate variables of different realms and also cover global or regional scales. An example of a global bias analysis is shown in Figure 3b.

For some variables, their spatial distribution or spatial connections are very important characteristics. This has been recognized by the evaluation community by introducing metrics and diagnostics that specifically analyze whether climate models can reproduce these spatial characteristics, for example, through pattern correlations (e.g., Bjarke et al., 2023; Bock et al., 2020; Fasullo, 2020; Wu et al., 2020). Figure 3c shows an example of a pattern correlation analysis for different variables. Thick horizontal lines represent the multi-model mean for that variable, the thinner lines represent the individual models. The whole group of spatial pattern metrics and diagnostics that are again very versatile in their application regarding realm and spatial scale, are summarized in Figure 2 as Approach 3.

Another very important and well-used diagnostic approach are budget assessments (Approach 4 in Figure 2). While traditionally many budget assessments in the atmospheric realm were focused on energy budgets at the top of the atmosphere (Lembo et al., 2019), often considering the influence of clouds (e.g., Dolinar et al., 2015; D. Li et al., 2023; Mayer et al., 2016), or energy budgets at the surface (e.g., D. Li et al., 2023; Wild et al., 2015), other budget. also moved into the focus of the scientific community, such as the evaluation of hydrological budgets (e.g., Freedman et al., 2014), land surface fluxes (e.g., J. Li et al., 2021), sea ice mass (e.g., Keen et al., 2021; S. Li et al., 2021) or sea ice concentration budgets (e.g., Nie et al., 2023).

Statistical analyses take a further step in the depth of model evaluation and benchmarking. These analyses quantify emergent relationships in the simulated output using statistical methods (including ML/AI approaches). Like the other approaches of Figure 2, statistical analysis can be applied globally or regionally. Taking precipitation as our example, there are many analyses that focus on global metrics, like globally distributed trends (e.g., Vicente-Serrano et al., 2022), but there are also many studies focusing on specific regional trends (e.g., J. Li et al., 2019; L.-L. Li et al., 2022; Peña-Angulo et al., 2020; Rivera & Arnould, 2020; Xin et al., 2020). A common characteristic of statistical approaches is their emphasis on distributions/histograms (e.g., Ahn et al., 2023; Ebtehaj & Bonakdari, 2023), probability density functions (e.g., Almazroui et al., 2021; Jönsson et al., 2023; Martinez-Villalobos et al., 2022; Sharma et al., 2022; Song et al., 2021), and cumulative distribution functions (e.g., Yang et al., 2018). Selective sampling of parts of these distributions are used for quantifying and evaluating extremes (e.g., John et al., 2022; Srivastava et al., 2020). Spectral analysis is another common approach, especially to evaluate variability in simulations (e.g., Ahn et al., 2022; Holt et al., 2022). One example for a diagnostic of this approach is shown in Figure 3d where histograms of one variable for two different models are shown.

”Earth System Metrics” include metrics and diagnostics that describe the behavior of the whole Earth system with all connected subsystems. A typical metric example is the equilibrium climate sensitivity (ECS) that describes the long-term temperature rise that is expected to result from a doubling of atmospheric CO_2 concentration (Knutti & Hegerl, 2008). It is an established metric that can quantify the joint effect of forcing and feedback, and is calculated regularly with newly available simulations (e.g., Meehl et al., 2020; Nijsse et al., 2020; Schlund, Lauer, et al., 2020). Other examples for Earth System Metrics are the transient climate response (TCR) and the transient climate response to cumulative emissions of carbon dioxide (TCRE). Both these metrics also describe the sensitivity of a given model to increases in CO_2 and are therefore used as metrics for characterizing how well the different parts of models are connected (e.g., Jones & Friedlingstein, 2020; Meehl et al., 2020; Spafford & MacDougall, 2020; Tokarska et al., 2020; R. G. Williams et al., 2020). These examples of Earth System Metrics are based on the Earth energy budget and demonstrate how the approaches of Figure 2 are often closely connected to each other. They do not apply to a specific realm or region. Another example of Earth System Metrics are emergent constraints with which a quantity related to the future climate is put in perspective with an observable quantity in the past or present-day climate (e.g., Allen & Ingram, 2002). Besides being applicable to global quantities, emergent constraint metrics can also be used to describe regional phenomena (e.g., P. Dai et al., 2024; Simpson et al., 2021).

The six approaches presented in Figure 2 are only separated to aid in explaining their general characteristics. In reality, most evaluation or benchmarking studies apply metrics and diagnostics from several of the approaches to diagnose the simulations and variables of interest, for example the diurnal cycle of precipitation (Covey et al., 2016), or the progress of model development over different phases of CMIP (Bock et al., 2020).

The growing prevalence and availability of ML/AI tools has led to a number of applications for model evaluation and benchmarking. Many of the tasks within evaluation and benchmarking (Figure 3) lend themselves to the strengths of ML techniques, namely classification and regression. Examples include regression algorithms that better handle outliers and collinearity than ordinary least squares, such as ridge regression (e.g., Ceppi & Nowack, 2021). Similar approaches can be extended to understand causal inference and constraining uncertainties in climate projections (e.g., Nowack et al., 2020). Classification approaches are also common in model evaluation by, for example, defining observed regimes and quantifying the ability of models to capture the observed regimes. This can be done with clustering methods, as has been shown for cloud regimes (e.g., I. Davis

& Medeiros, 2024; Tselioudis et al., 2021), or with more sophisticated ML methods like self-organizing maps (e.g., Gibson et al., 2017; Nigro et al., 2011).

3. Tools for Monitoring Climate Simulations

Ahead of evaluation and benchmarking, the verification and validation of the numerical implementation of climate models entails a multi-faceted approach. As noted in Section 2, verification attempts to ensure that the model is implemented correctly. While some verification can be accomplished by considering basic characteristics of the simulation (e.g., energy conservation), subtle changes during model development or in the computational environment can be difficult to detect.

The chaotic nature of the climate system poses challenges for verification of software and hardware changes. As is well understood, small changes in the initial conditions may introduce perturbations that grow in time and may influence the numerical solution that can appear as a difference in the time-averaged behavior of the system (i.e., the "climate") that persist for a substantial amount of simulated time. Similarly, changing computational environments may introduce unexpected changes, and ruling out significant differences may require very long simulations (Guarino et al., 2020). The use of ensembles of simulations helps to characterize this variability. Similar perturbations can be introduced by software changes (e.g., a compiler version change) or hardware (e.g., running the model on a different HPC system). While these changes sometimes expose errors or technical issues in a model, it is difficult to ascertain whether differences from a "baseline" simulation are statistically significant or not. One tool that has been developed to address this issue is the Ensemble Consistency Test (A. H. Baker et al., 2015). In this approach, a series of short simulations are produced to generate a distribution of geophysical quantities that are then statistically compared with a reference distribution; statistically significant changes in the distribution indicate that the change has shifted the simulated climate and, therefore, flag the change as needing additional attention. The recent development of the Ensemble Consistency Test approach suggests that many climate-changing software or hardware changes can be detected in only a few time steps (Milroy et al., 2018). An alternative approach, known as the Time Step Consistency test, compares the time step sensitivity of a known model configuration to some change to evaluate whether the prognostic variables show a significant deviation by measuring the deviation between a 2 s time step and a 1 s time step (Wan et al., 2017). Having objective tests that measure whether changes are significant is crucial when moving complex climate model codes to new computational systems and architectures, and also for evaluation of code modifications that do not replicate the reference case bit-for-bit but nevertheless are not expected to modify the simulated climate (i.e., refactoring).

4. Implementations of Model Evaluation and Benchmarking

4.1. Climate Model Development Community

As was noted above, and illustrated in Figure 1, climate models (and their individual component models) undergo verification and validation processes during development and tuning. There are different practices across modeling centers, but some commonalities naturally emerge. For example, it is typical for components of a model (e.g., an individual parametrization) to be developed separately from other components, and possibly even separately from a climate model altogether, and later be implemented, evaluated, and possibly adopted. An example of this is the Gent-McWilliams parametrization of mesoscale mixing in the ocean, as recounted by Gent (2011): the parametrization was first developed from theoretical considerations (Gent & McWilliams, 1990), and afterward implemented into the GFDL ocean model by Danabasoglu et al. (1994), and later was incorporated into the first version of the Community Climate System Model (Boville & Gent, 1998). Other parts of climate models are developed in situ, so to speak, always being part of a particular model. For example, the turbulence parametrization used in ECHAM was developed within that model in the 1990s (Brinkop & Roeckner, 1995) and has been used in all subsequent versions (Stevens et al., 2013). Evaluations occur throughout the process, but evaluating new approaches should be thought of as distinct from later model benchmarking activities which occur only when all the pieces are assembled and climate simulations are produced.

Once a model is developed to the point of producing climate simulations, a great deal of evaluation is undertaken. A key component of that evaluation takes place during the model "tuning" process. Model tuning is effectively the process of "calibrating" the model's climate by parameter estimation to targets which can be observed or modeled quantities or for some condition to be met (Schmidt et al., 2017). Hourdin et al. (2017) provided background on the concept, including providing some perspective on methodologies and challenges. Mauritsen et al. (2012)

shows examples of alternative tuning of one climate model that is particularly instructive. Using a variety of global annual-mean diagnostics, several different tunings of their model produce a relatively narrow range of climates that all appear equally plausible. They show, however, that regional differences emerge between their different tunings, the representation of tropical intra-seasonal variability is sensitive to the tuning choices, and the climate sensitivity to increased CO_2 varies by more than 0.5 K. Recently, Duffy et al. (2024) used a perturbed physics ensemble of another model to show that cloud feedback (and thereby climate sensitivity) can vary widely within a climate model depending on parameter settings. Elsaesser et al. (2025) used a similar perturbed physics ensemble to train a neural network-based emulator to calibrate the model parameters to match a set of observational references and generate a second "calibrated physics ensemble." This work illustrates how model tuning is evolving toward more systematic and objective methods that might produce simulations in better agreement with observations, and provides an example of how ML approaches can be incorporated into the model development cycle. The tuning process, therefore, plays a decisive role in determining the climate of a model, and methods of evaluation are key factors during this stage of development.

Critically evaluating the simulated climate's fidelity is, as we have just established, a crucially important step in the model development cycle, and indeed model evaluation is essential for climate models to make accurate and trustworthy climate predictions (T. Schneider et al., 2024). As discussed by Hourdin et al. (2017) and others though, it is infeasible to evaluate every aspect of a climate model, and the many different valid ways how a model can be successfully tuned (e.g., Mauritsen et al., 2012) makes such an endeavor very difficult. So it falls upon the modeling centers and the broader community to find meaningful methods that are broadly applicable for evaluating the simulated climate as well as more nuanced and specialized diagnosis of regions and phenomena of interest. In doing so, sometimes it has become useful to differentiate between metrics and diagnostics. Gleckler et al. (2008) provided one approach to this, defining metrics as scalar quantities that represent the distance between model results and observations (or any other reference), as shown in approach 1 of Figure 2. Diagnostics, on the other hand, encompass a much broader array of evaluations that can be judged qualitatively or quantitatively, as illustrated by the other approaches of Figure 2. Recently a lot of attention has been devoted toward developing "process-oriented diagnostics" that aim to guide model development by quantifying errors in process representations within the models (as opposed to broad geographic biases, e.g.) (Neelin et al., 2023) or even special observational data sets for "process-oriented" evaluation (e.g., Kaps et al., 2023). These generally fall into the statistical analyses of Figure 2, though sometimes the methods are not strictly statistical.

4.2. Open-Source Evaluation and Benchmarking Tools

The introduction of consolidated software packages that calculate and visualize climate model metrics and diagnostics has provided modeling centers and individual researchers access to powerful evaluation tools at a relatively minimal cost. Leveraging these tools during model development helps to guide decision-making. This is highly practical as the evaluation/benchmarking tools can be applied to different versions/variants of a model during its development cycle or in the tuning process. Having "off the shelf" (and independent) evaluation tools provides the advantage of examining a model's performance more comprehensively than otherwise might be practical. Some evaluation packages provide holistic assessments as output (e.g., Gleckler et al., 2008; Reichler & Kim, 2008; Taylor, 2001). For example, as introduced by Gleckler et al. (2008), Waugh and Eyring (2008), and Pincus et al. (2008), diagnostic outputs such as Portrait Diagrams can bring many metrics together, sometimes using multiple observational baselines, in a concise display of model skill. Similarly, Taylor diagrams (Taylor, 2001) provide a compact display of the pattern correlation and bias of particular fields. The normalized mean squared error provides similar information as a Taylor diagram and can be visualized as a bar chart (e.g., Simpson et al., 2020). Along with these concise displays of metrics, many of these benchmarking tools provide a browsable catalog of diagnostics that range from maps and time series of basic fields (e.g., long-term mean, global mean) to climate indices to more elaborate process-oriented diagnostics (see e.g. ESMValTool gallery (ESMValTool Development Team, 2025), PMP mean climate result browser (PCMDI Metrics Package, 2025), ILAMB land comparison (ILAMB 2.6, 2025)).

There has been increasing interest in bringing different tools together to make model evaluation and intercomparison more accessible and leverage the tools more efficiently. One example is ESMValTool (Eyring et al., 2020; Eyring, Righi, et al., 2016; Righi et al., 2020) which includes comprehensive metrics and diagnostics to evaluate model performances against observed quantities or other reference values (e.g., Bock & Lauer, 2024; Bock et al., 2020; Gier et al., 2020, 2024; Lauer et al., 2017, 2023) or conservation properties and theoretical considerations about the intrinsic nature of the climate system (e.g., Lembo et al., 2019; Mauritsen et al., 2012). Recently, also more modern approaches like emergent constraint analyses (e.g., Schlund, Lauer, et al., 2020; Zechlau et al., 2022), causal model evaluation (e.g., Galytska et al., 2023; Karmouche et al., 2023), extreme event analyses (e.g., Malinina & Gillett, 2024; Paçal et al., 2023), and machine learning methods (e.g., Schlund, Eyring, et al., 2020; Swaminathan et al., 2024) have either been implemented in ESMValTool and successfully used for model evaluation or data have been processed by EMSValTool to facilitate further analyses outside the ESMValTool framework. ESMValTool also has been used in several chapters of IPCC AR6 (e.g., Eyring et al., 2021). Most importantly, ESMValTool is constantly updated and improved to meet the upcoming challenges of multi-model evaluation and benchmarking, for example, the next phase of CMIP with a huge expected data volume (Lauer et al., 2025; Schlund et al., 2023, 2025), and therefore meets the criteria for a model benchmarking framework as listed in Section 2.

Another example is the PCMDI Metrics Package (PMP) (J. Lee et al., 2024) which focuses on delivering and expanding a diverse suite of performance metrics of the physical climate. It includes model performance assessment accessible via interactive visualization capabilities and a public archive of the summary statistics. Metrics from the PMP have been used to document the model performance of the E3SM and GFDL models (Smith et al., 2024; Zhao et al., 2018). PMP's metrics for extra-tropical modes of variability have been leveraged during the evaluation of models during their development cycle and/or inter-comparison with other CMIP models (e.g., J. Lee et al., 2019; Orbe et al., 2020; Smith et al., 2024; Sung et al., 2021). The CLIVAR ENSO metrics that were incorporated into the PMP workflow have been used to show the performance evolution of IPSL models between its CMIP5 and CMIP6 versions (Boucher et al., 2020; Planton et al., 2021).

A further example tool is the International Land Model Benchmarking (ILAMB) package (Collier et al., 2018), which offers a comprehensive assessment of models including land surface carbon cycle components by evaluating hydrological and biogeochemical variables and their relationships with driving variables. ILAMB is an open source package developed with significant input gathered through community workshops (Hoffman et al., 2017; Luo et al., 2012). Employing a suite of in situ, remote sensing, synthesis, and reanalysis data, ILAMB produces a hierarchical set of web pages containing statistical analyses and figures designed to provide insights into the strengths and weaknesses of models spatially and through time. For every variable, ILAMB generates graphical diagnostics (spatial contour maps, time series line plots, and Taylor diagrams; Taylor, 2001) and determines model performance for the time period mean, bias, RMSE, spatial distribution, interannual coefficient of variation, seasonal cycle, and long-term trend. Model performance scores are calculated for each metric and variable and are scaled based on the degree of certainty of the observational data set, the scale appropriateness, and the overall importance of the constraint or process to model predictions, following a customizable weighting rubric. Scores are aggregated across metrics and data sets to produce a single score for each variable for every model or model version. ILAMB checks functional relationships between prognostic variables and one or more driver variables through variable-to-variable comparisons (e.g., gross primary production vs. precipitation) and scores model performance in capturing these emergent relationships. The International Ocean Model Benchmarking (IOMB) package (Fu et al., 2022), which employs the same code base as ILAMB, provides diagnostics to evaluate ocean physical and marine biogeochemical fields and produces a similar hierarchy of web pages containing statistical analyses and figures. ILAMB and IOMB were used to benchmark CMIP5 and CMIP6 models, as well as the mean of the CMIP5 and CMIP6 models. This combined analysis was illustrated in Chapter 5 of the Sixth Assessment Report (Canadell et al., 2021).

The growth in the number of different diagnostic tools and the diversity of implemented metrics has led the community to seek methods for combining the resulting workflows and diagnostics for comprehensive analysis capabilities. To execute this myriad of heterogeneous tools, users must manually coordinate model output data and run each tool in turn. A fully automatic evaluation can be challenging in such a setup. While various tools have been developed, the Coordinated Model Evaluation Capabilities (CMEC) is an effort that was initiated in the United States to unify the user interface and operation of various analysis packages for more efficient systematic evaluation of climate models. A key goal for CMEC is to achieve (a) interoperability through a generally robust and lightweight wrapper, (b) workflow standards that provide a common syntax for executing tools, and (c) bring diagnostic outputs together in an integrated visualization interface. Several tools that were developed mainly in the United States including PMP, ILAMB, and the NOAA Model Diagnostics Task Force (NOAA MDTF) suite of packages have become compliant with the CMEC by adopting the interface standard.

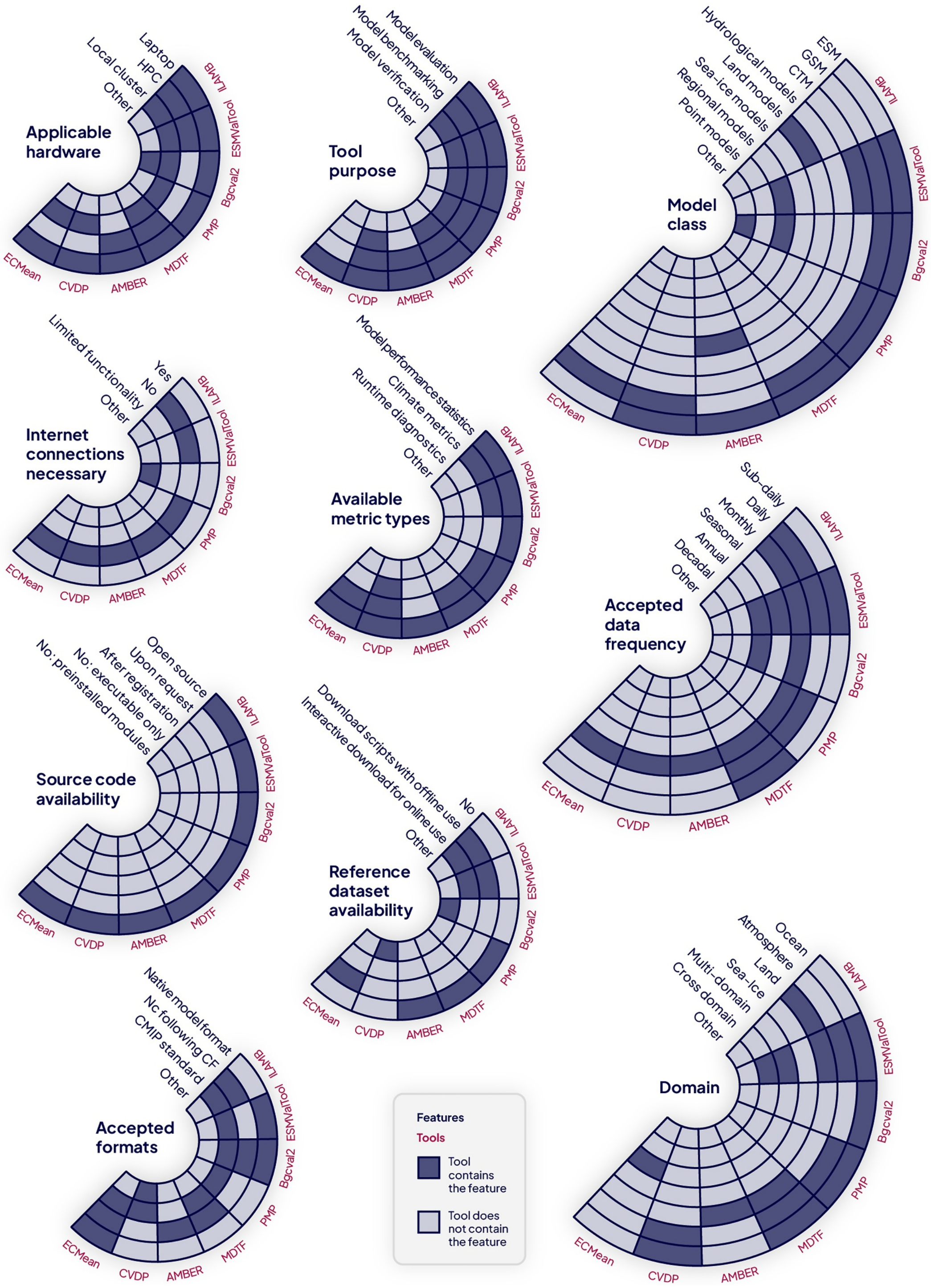


Figure 4. Schematic of different community open-source evaluation and benchmarking tools with their respective characteristics (Swaminathan et al., 2025).

As part of their mandate, the CMIP Model Benchmarking Task Team (WCRP CMIP, 2025a), that was created during the CMIP reorganization in preparation for their next phase (CMIP7), has collected and collated information about different open-source evaluation and benchmarking tools that have been used in the community for analyses of CMIP6 simulations. The list as it stands is now available via the CMIP website (WCRP CMIP, 2025b), and can easily be extended with additional tools that are not yet listed. Figure 4 illustrates many of these open-source evaluation and benchmarking tools, including ESMValTool, PMP, ILAMB and IOMB that were described earlier. More details about these software packages can also be found in Table A1 in Appendix A.

Even as software tools are developed and established for systematic, comprehensive model evaluation, it is necessary to continue to devise novel approaches to confront climate models with observations. One reason this is imperative is to continually check that the models have not been "overfit" to any particular time period or observational target. Updating the observational baselines provides one way to keep such overfitting in check, but takes time. Applying additional constraints by clever uses of observations provides more opportunity to expose model deficiencies. There are a myriad of examples from the literature that have demonstrated potential applicability of various metrics, and often have been applied as methods for model inter-comparison. A few examples are provided here:

- [**Tropical cyclone representation**] Roberts et al. (2020) used Tropical Cyclone tracking methods to verify the role of horizontal resolution in determining Tropical Cyclone representation in GCMs.
- [**Arctic sea ice projections**] Massonnet et al. (2012) used benchmarking of CMIP5 sea ice model simulations to constrain Arctic sea ice projections through a process of model selection. This model subset was subsequently used in IPCC AR5, and other model development and evaluation activities.
- [**Arctic sea ice loss**] Notz and Community (2020) used benchmarking of CMIP6 simulations to produce a "plausible subset" of CMIP6 models, considering the rate of Arctic sea ice loss per degree of global warming alongside standard sea ice metrics.
- [**Climate extreme indices**] Kim et al. (2020) evaluated CMIP6 GCMs in terms of their performance in simulating the standard climate extreme indices defined by the Expert Team on Climate Change Detection and Indices (ETCCDI).
- [**Extreme precipitation**] Scoccimarro and Gualdi (2020) provided an assessment of the CMIP6 extreme precipitation historical representation, following the same approach previously applied to CMIP5 defined in Scoccimarro et al. (2013) consolidating the usage of different metrics used to investigate the shape of the extreme wet end of the precipitation distribution under different climate conditions.
- [**Heatwave characteristics**] Hirsch et al. (2021), provided the first global evaluation of CMIP6 models in representing heatwave characteristics between 1950 and 2014, also applying the same methodology to CMIP5 models for comparison.

Despite efforts to mitigate overfitting, either by using multiple metrics, considering the physical process representation or structural uncertainties comprised in observations, there is still a risk of overfitting, and so scientists should be aware of this when interpreting results from model evaluation and benchmarking activities.

A crucial aspect of model evaluation regards the implementation of a weighting scheme for the performance metrics of a multi-model ensemble, taking into account the ability of models to reproduce current climate (Knutti et al., 2017), as well as analogies between ensemble members pertaining to the same "model family" (Sanderson et al., 2015). From a statistical point of view, while the conventional weighting scheme is applied to mean/variance of relevant observables (e.g., Knutti, 2010), recent approaches have involved weighting of models by combining performance and independence (Brunner et al., 2020), or comparing all moments of the distributions, as in the case of approaches using the Wasserstein distance (e.g., Vissio et al., 2020).

The approaches mentioned above provide new insights and additional facets to model evaluation that go well beyond the mean state climate. In particular, as climate models have advanced, the ability to capture climate variability, trends in climate variables, and climate extremes have emerged as key indicators of model fidelity. The first promising results from individual and large ensembles of climate models that show skill in producing regional climate features and extremes (e.g., Deser et al. (2020) or Das and Ganguly (2025)), and a thorough discussion on model skills in reproducing observed trends (Simpson et al., 2025), demonstrate significant advances in climate modeling capabilities over the years.

It should also be noted, that retrospective evaluations of climate projections by earlier generations of models—for example, how well climate models "predicted" or "projected" changes in a given climate metric against what actually occurred in the observed climate system—are also useful measures of model performance. For example, Stouffer and Manabe (2017) and Hausfather et al. (2020) provide examples of evaluations of the skill of early climate models (developed in the 1970s—early 2000s) in predicting the patterns and rates of changes in global temperature in response to rising atmospheric CO_2.

4.3. End Users

The applications for global climate model benchmarking methods and information have grown beyond climate scientists and model developers. Demand for reliable information about current and possible future climate conditions continues to grow as policymakers, governments, and industries increasingly evaluate exposure to climate-related risks. Global climate models remain a commonly used tool that these climate data consumers look to for climate information. Often with an eye toward reducing uncertainty about climate projections (Hawkins & Sutton, 2009), many of these users have turned to climate model benchmarking/evaluation efforts to guide their refinement of information from global climate models to better reflect their specific needs or requirements. Some specific examples of the use of model benchmarking methods outside the global climate modeling development community include:

- [**Reducing uncertainties**] Selecting specific global climate models for use in climate assessments, other than IPCC reports, based on their benchmarked performance for specific aspects of the climate in a particular region (sometimes referred to as model "culling" or "subsetting") (e.g., Brekke et al., 2008; Infanti et al., 2020; Massonnet et al., 2012; Pierce et al., 2009).
- [**Reducing uncertainties**] Weighting climate model projections with benchmarking results to produce a multi-model estimate with potentially lower uncertainty about a target metric (e.g., Knutti et al., 2017).
- [**Regional impacts**] Benchmarking results and methods used to determine which climate model outputs to use as inputs for user-specific impact models, and evaluate the output of those models (e.g., Wagener et al., 2022).
- [**Regional impacts**] Benchmarking outputs from lower-resolution global models used in combination with storyline-like high-resolution global or regional simulations to provide useful information for adapting and responding to extreme climate/weather events (De Dominicis et al., 2020).
- [**Input for AI/ML applications**] Benchmarking techniques used to evaluate sources of training data or compare output from reduced complexity or data-driven (AI- or ML-based) climate models (e.g., Nicholls et al., 2021; Ullrich et al., 2025; Watson-Parris et al., 2022).
- [**Constraining future projections**] Benchmarking results used to constrain future projections of key environmental quantities via the application of emergent constraints (e.g., Brient, 2020; Cox et al., 2018).
- [**Constraining future projections**] Benchmarking results from climate models used to develop climate model weighting schemes and screening methods to constrain future hydrology and inform water resource planning in the Colorado River Basin in the Western U.S. (Lukas et al., 2020).

Notably, in many documented uses of model benchmarking methods or data, the specific benchmarking data used are computed specifically by the user for their application, instead of relying on a common repository of model benchmarking results. The use of these derived benchmarking data sets can make it difficult to compare results from different studies and requires additional effort from model developers. This presents an opportunity for a coordinated model benchmarking effort that targets common use cases of benchmarking methods and data. For instance, given the regional scope for many climate benchmarking consumers, there is an opportunity to develop a common framework of regionally-based benchmarking results for global climate simulations addressing commonly-used climate variables. There is also a high demand for benchmarks that categorize the performance of climate models with respect to extreme events when using climate model data to evaluate climate exposure and risk. The ease of availability and endorsement of such benchmarking results by the climate modeling community will likely encourage a broad set of consumers of climate change information to use this benchmarking data and also encourage the interplay between climate modeling, adaptation, and mitigation activities (Lu, 2024).

5. Model Improvements Demonstrated by Systematic Evaluation and Benchmarking

The expanded and systematic application of model-observation intercomparison strategies has revealed notable improvements in the fidelity of today's climate models and a reduction of long-standing cross-generational biases. Continued evaluation efforts across model generations also provides reminders of biases that have not been eliminated or that may recur and helps modeling centers to identify processes to prioritize in future model development. In this section, we describe a few examples of model performance benchmarking between different CMIP generations (i.e., CMIP5 and CMIP6), using some of the evaluation tools discussed in this paper and review specific model biases that have been reduced as a result of benchmarking.

To illustrate the use of evaluation/benchmarking methods to assess cross-generational bias reduction, here we apply the PCMDI Metrics Package to check the performance of a suite of CMIP5 models alongside the

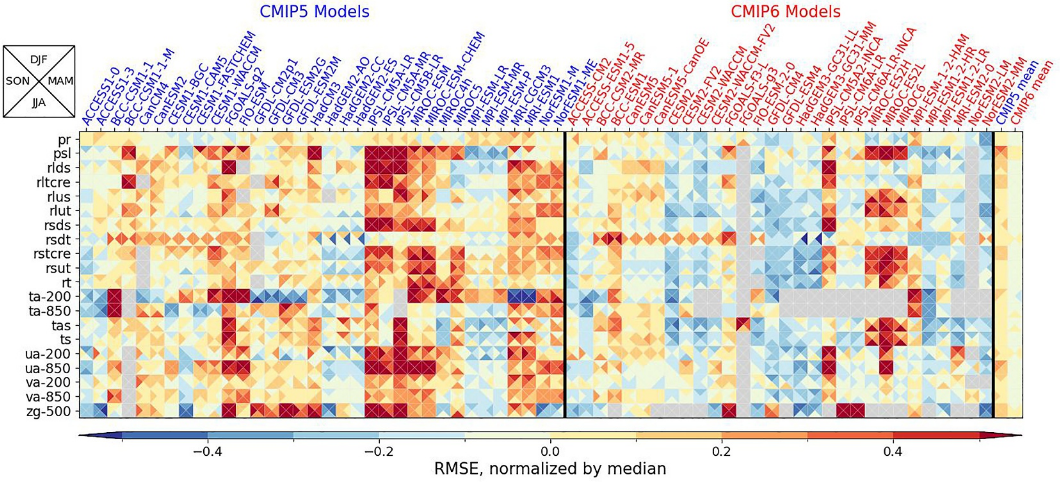
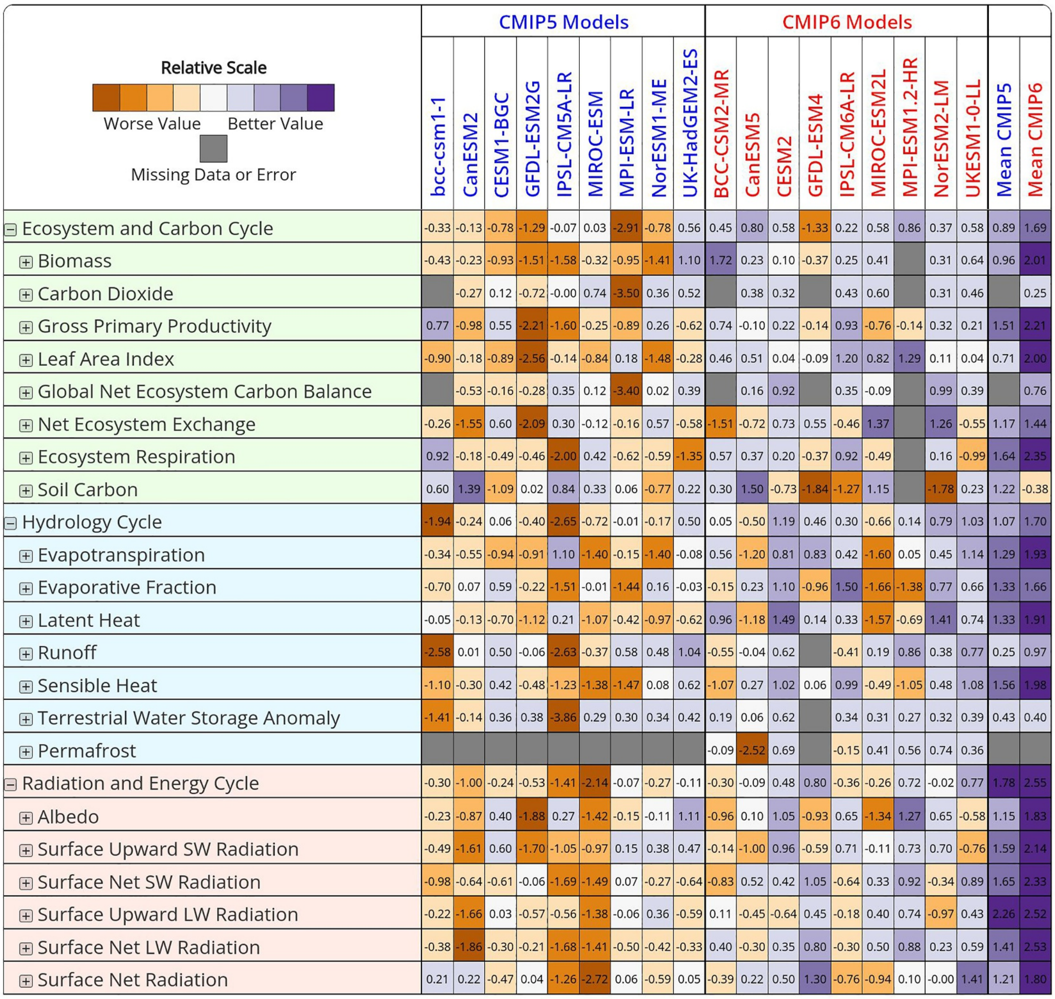


Figure 5. Overview scores for CMIP5 (left-hand side) and CMIP6 (right-hand side) models generated by the PCMDI Metrics Package, for the seasonal climatology of multiple atmospheric variables evaluated against observational data sets over the period of 1981–2005. Included here are a subset of models from institutions that participated in both CMIP5 and CMIP6 historical experiments, in order to trace changes from one multi-model ensemble to the next. CMIP5 models are labeled in blue and CMIP6 in red. Root Mean Squared Error (RMSE) values are calculated against reference data sets for each season and normalized by the median value of each row. Thus the normalized RMSE value indicates performance relative to other models within a given row, with negative values indicating a better agreement with observations. Detailed analysis information can be found in J. Lee et al. (2024) (J. Lee, 2025).

corresponding CMIP6 models from the same modeling centers. The Portrait Diagram summarizing various atmospheric performance metrics (Figure 5), indicates an overall improvement in the seasonal climatology in general. In many cases, the CMIP6 models' Root Mean Squared Errors (RMSE) are smaller than those from the CMIP5 models in most cases in this example (colors shifting from warmer hues for CMIP5 to cooler hues for CMIP6). However, it should be noted that the evolution of such statistics can be sensitive to the selection of reference data sets, analysis period, ensemble member, and subset of models. Similarly, most CMIP6 land carbon cycle models exhibited improvement over their CMIP5 counterparts as shown in an overview calculated with the ILAMB package (Figure 6). The CMIP5 multi-model-mean performs better across all the variables with regard to the metrics than any single CMIP5 model. Likewise, the CMIP6 multi-model-mean performs better than any single CMIP6 model across all the variables. However, there are a few variables for which a single model may outperform the mean of CMIP5 or CMIP6 models (e.g., CMIP5 biomass, CMIP6 carbon dioxide, CMIP5/6 soil carbon, CMIP6 global net ecosystem carbon balance, CMIP5 runoff, and CMIP6 terrestrial water storage). The multi-model mean of the CMIP6 suite of models performs better, for nearly every carbon cycle variable considered in Figure 6, than any single model considered and better than the multi-model mean of the CMIP5 suite of models. Another way of quantifying performance progress between different CMIP model generations is to look at the pattern correlation for different variables. In these diagnostics, calculated with ESMValTool, the closer the correlation value is to 1 the better the performance. Focusing on a subset of atmospheric variables, Figure 7 shows the progress between CMIP5 (blue lines) and CMIP6 (red lines). For most variables shown, the CMIP6 multi-model-mean has a higher correlation coefficient than the CMIP5 multi-model-mean, and very often the spread between the correlation coefficients is reduced from CMIP5 and CMIP6. This is a clear indication that the models reproduce a more realistic climate for these variables (closer to the observational reference data set) than the model generation before.

For some variables the alternative observational data sets (marked as gray circles in Figure 7) show a weaker correlation than some of the individual models (marked as thin horizontal lines), for example Northward Wind at 850 hPa. This is noteworthy since it indicates that some models represent “reality” better than actual observations. For all shown variables the analyzed time period for both model simulations and observations are identical which makes it unlikely that internal variability is the cause for this phenomenon. Also, the used versions of the observational data sets are not too far out-of-date to explain the weaker correlation. The most likely explanation for this phenomenon is that there are uncertainties connected to the observations that are not taken into account


Relative Scale
Worse Value
Better Value
■ Missing Data or Error

	CMIP5 Models									CMIP6 Models										
	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	Mean CMIP5	Mean CMIP6
⊟ Ecosystem and Carbon Cycle	-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
⊟ Hydrology Cycle	-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	■	■
⊟ Radiation and Energy Cycle	-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

Figure 6. Overview scores for CMIP5 (left-hand side of table) and CMIP6 (right-hand side of table) climate model land models generated by ILAMB for multiple metrics against different observational reference data sets. The evaluation time period depends upon the time period available for each of the observational data sets that contributes to the aggregate score for each variable. Scores are relative to other models within each row, with positive scores (blue to purple) indicating a better agreement with observations and negative scores (yellow to brown) indicating a worse agreement with observations. Models included are only those from institutions that participated in both CMIP5 and CMIP6 carbon cycle experiments, in order to trace changes from one ensemble to the next. CMIP5 models are labeled in blue and CMIP6 in red at the top of the table (Hoffman, Collier, et al., 2025).

when the spatial patterns of the climatologies of the variables are evaluated. For a more in-depth discussion about the importance of carefully choosing observations for evaluation and benchmarking activities and an appropriate treatment of observational uncertainties can be found in Section 7.

5.1. Reductions in Specific Model Biases

Here, we describe a few examples of specific model biases that have largely improved in climate models following routine quantitative assessments over the past decade.

5.1.1. Vegetation Phenology

Early work in evaluating and intercomparing land carbon cycle models identified persistent biases in the timing of the seasonal variation of vegetation growth with respect to satellite-derived leaf area index (LAI) (Randerson et al., 2009, Figure 1). Through comparison with the observational data set from MODIS LAI, the authors showed that the timing of maximum leaf area lagged behind observations by one to 2 months. Since the regional timing

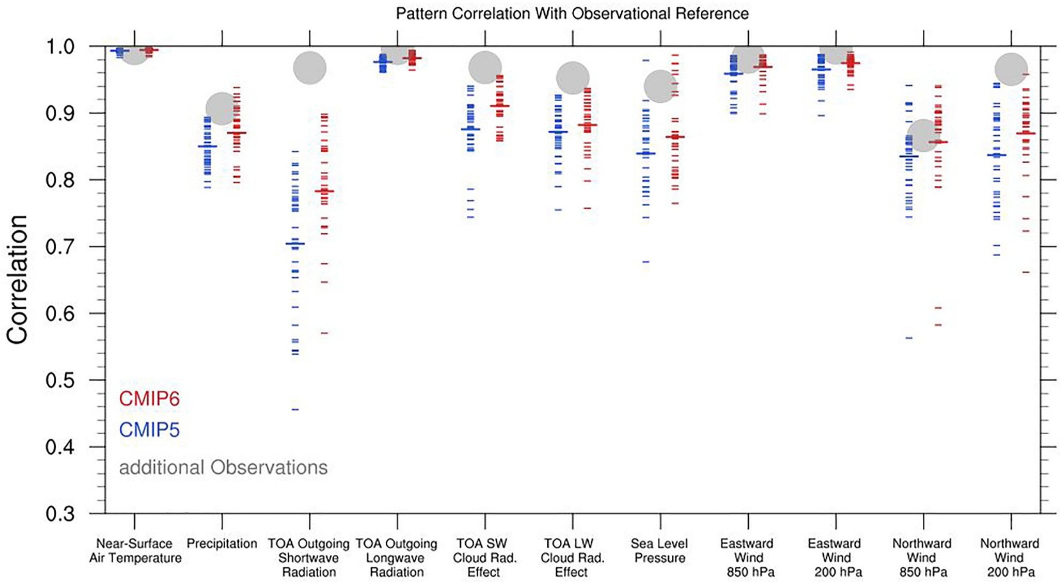


Figure 7. Pattern correlation for different atmospheric variables for CMIP5 (blue colors) and CMIP6 (red colors) models generated by ESMValTool, based on the period 1985–2004. Thick horizontal lines show the correlation with the multi-model mean and the thinner lines indicate the correlations for each individual model. Gray circles indicate a second observational data set. Models included are only those from institutions that participated in both CMIP5 and CMIP6 historical experiments, in order to trace changes from one ensemble to the next (Hassler & Bock, 2025).

delays were systematic in both of the land carbon models being studied, the authors identified the likely cause as an underestimate of the carbohydrate pools carried over from one growing season to the next. Subsequent modifications to the phenology scheme in the Community Land Model - Carbon-Nitrogen (CLM-CN) model significantly improved the timing of maximum LAI.

5.1.2. Biomass in the Amazon Basin

In the same study as discussed above, both of the land models substantially overestimated aboveground live biomass in the Amazon Basin compared to estimates from satellite observations (Randerson et al., 2009, Figure 5). While both models predicted too much biomass, they adequately reproduced the observed spatial patterns of aboveground live biomass in the basin. Part of the magnitude bias was attributed to the use of a preindustrial land cover map with higher forest cover fractions in the southern border of the Amazon Basin, but the authors attributed most of the bias to low autotrophic respiration in one of the models and excessive allocation of net primary production to wood in both models. Improvements in model predictions of respiration and carbon allocation in tropical trees have yielded reductions in the bias of aboveground live biomass in the Amazon Basin.

5.1.3. East Asian Summer Monsoon

The East Asian summer monsoon (EASM) is an important component of the Asian climate system, carrying moisture from the Indian and Pacific Oceans to East Asia, and exhibits intense interannual variability, resulting in severe droughts and floods (Yihui & Chan, 2005). Several studies and multi-model benchmarking efforts have focused on model reproduction of the EASM (Boo et al., 2011; Kang et al., 2002). The common biases said to influence this variability are a weakened western North Pacific anticyclone and the Meiyu-Baiu-Changma rainband. Yu et al. (2023) used atmospheric variables from ERA5 (Hersbach et al., 2020), precipitation from the Climate Prediction Center Merged Analysis of Precipitation (Xie et al., 2007) and monthly Sea Surface Temperature data from the Hadley Centre Sea Ice and Sea Surface Temperature (HadISST) data set (Rayner et al., 2003) to show that, relative to CMIP5, CMIP6 models have stronger westerly winds over the Indo-China peninsula and reduced biases in the western Pacific subtropical high as well as the Meiyu-Baiu-Changma rain belt. Such improvements led to a better simulation of the EASM dynamics in CMIP6 models.

5.1.4. Cloud and Water Vapor Processes

Representation of cloud processes and water vapor in climate models plays an important role in accurately modeling the surface warming response due to global warming. Evaluation studies (Dolinar et al., 2015) performed on CMIP5 models have helped identify key sources of errors and uncertainties in these processes such as cloud vertical distribution and overlap (Stephens et al., 2002), effects of over-tuning cloud processes to observations (Lauer & Hamilton, 2013) and interactions between large-scale circulation and clouds (Su et al., 2014). More recent studies highlight changes made within the CMIP6 suite of models, leading to improvements in the representation of these processes (Bock et al., 2020; Jiang et al., 2021; Schlund, Lauer, et al., 2020).

6. Remaining Biases in Climate Model Simulations

A variety of biases exhibited by many or most climate models have been difficult to reduce or completely eliminate despite frequent model evaluation and bias characterization. Sources of such long-standing biases remain difficult to identify, and model developers often perform model tuning exercises or apply bias removal techniques to reduce their impacts on simulations of contemporary climate and future projections under different greenhouse gas emissions scenarios. Some examples include:

- Precipitation: Biases in precipitation in climate models vary widely and are sensitive to the representation of clouds, radiation, surface processes, and multi-scale circulation (e.g., Ahn et al., 2023; A. Dai, 2006). In monsoon regions, biases in the location, timing, and spatial and seasonal distributions have significant implications for understanding impacts of climate change and potential mitigation strategies (Hegerl et al., 2015; Pathak et al., 2019; Pincus et al., 2008). Despite efforts to improve the representation of linked processes, regional biases in the strength and timing of precipitation remain in current generation models.
- 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 the observed world. Tian and Dong (2020) discuss the progress made in the reduction of this continuous source of bias across generations of models and hypothesize that it may still take decades to completely eliminate this error in simulations.
- Warming bias in the tropical troposphere: Excessive tropospheric warming in the tropics has been well documented in climate models (Bengtsson & Hodges, 2011; Douglass et al., 2008; McKitrick et al., 2010). Updated comparisons with radiosondes and satellites, as well as reanalysis data, show that models continue to exhibit statistically significant warming trend differences in both the averaged lower and mid-tropospheric temperature series, and that such biases are not only restricted to the tropics but appear to be a global phenomenon (Casas et al., 2023; Po-Chedley et al., 2021).
- Arctic sea ice sensitivity to global warming: Arctic sea ice decline in CMIP models is lower than that observed in the satellite record. Internal variability is huge in the polar regions and so we would not necessarily expect the models to match reality (Rosenblum & Eisenman, 2016). However, systematic benchmarking activity has shown that CMIP models do underestimate the sensitivity of Arctic sea ice to global warming. The reduction in Arctic sea ice extent per degree of global warming, or cumulative anthropogenic CO_2 emissions, is lower in CMIP historical simulations than observed (Notz & Community, 2020). This underestimation is most likely caused by deficiencies in Arctic amplification and atmospheric and oceanic northward transport of heat.

The development of additional performance metrics and the collection of new observational data sets are required to identify and understand the sources of these long-standing model biases. Furthermore, continued systematic and increasingly comprehensive assessments of model performance are needed to inform model development efforts aimed at reducing biases.

7. Reference Data and Uncertainty Quantification

To benchmark a climate model or its components, observational or observationally constrained data are commonly used and are often considered to be "the absolute truth." However, these observational data sets are never free of uncertainties and systematic biases. The sources of the observational inaccuracies are diverse and include uncertainties in the measuring system, biased sampling schemes, uncertainties introduced by different processing steps, spanning from the raw measurement (e.g., irradiances) to a gridded and equally temporally spaced product, and any resampling, regridding, interpolation or homogenization that occurs before a comparison can be made with model simulations (e.g., Merchant et al., 2017; Von Clarmann et al., 2020; Zumwald et al., 2020). Additional sources of uncertainties and difficulties arise as soon as several different observational sources are combined into one data set, for example, the HadCRUT5 data set (Morice et al., 2021), the HadISST data set (Rayner et al., 2003), or the SWOOSH data sets (S. M. Davis et al., 2016).

While some information on uncertainties and biases can be found alongside the observational data set or can be deduced from the corresponding publications or documentations, clearly defined, generally applicable, and traceable uncertainties are normally not provided with the gridded satellite observations or even with point measurements such as ground-based station data. Although not yet adopted as a common practice, the community has started to use multi-observational means, calculated from multiple observational data sets that provide measurements for the same variable, to characterize the observational uncertainties of a specific variable (e.g., Lauer et al., 2023; Pathak et al., 2023). Such a method cannot replace a full and statistically rigorous uncertainty propagation method, but it can help characterize the range of observational uncertainty for a given variable and indicate when or where model simulations cannot be robustly constrained.

Another type of product commonly used in climate model benchmarking are reanalysis data sets. Examples of the most commonly used reanalysis data sets in the atmospheric community include the fifth generation ECMWF reanalysis (ERA5) (Hersbach et al., 2020), the Japanese 55-year Reanalysis (JRA-55) (Kobayashi et al., 2015), and the Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA2) (Gelaro et al., 2017). More recently, reanalysis products that focus specifically on atmospheric composition were developed and are widely used in the community to evaluate model simulations, for example, the Copernicus Atmosphere Monitoring Service (CAMS) reanalysis (Inness et al., 2019) or the MERRA-2 Stratospheric Composition Reanalysis of Aura Microwave Limb Sounder (M2-SCREAM) (Wargan et al., 2023). Ocean reanalyses give a four dimensional description of the ocean combining ocean models, atmospheric forcing fluxes and observations (Storto et al., 2019). Well-used examples of oceanic reanalysis include the global eddy-resolving physical ocean and sea ice reanalysis at 1/12° horizontal resolution (GLORYS12) (Jean-Michel et al., 2021) and the Simple Ocean Data Assimilation ocean reanalysis (SODA) (Carton et al., 2018). One of the largest advantages of reanalyses over many observational data sets is their full spatial coverage, often spanning the whole globe, and their long temporal coverage. While some reanalyses cover only the period from the 1980s (e.g., MERRA-2) or even from the 2000s (e.g., CAMS) to the 2020s, other products cover over half a century (e.g., ERA5 or JRA-55). Even if the gridded reanalysis data sets provide an ensemble (e.g., ERA5), commonly only the ensemble means/medians are used in practice for most evaluations, leading to the omission of the uncertainty intervals of the benchmark data. A disadvantage of reanalyses is that numerical models are used to provide continuous space—time coverage, constrained in some but not all regions through various data assimilation methods by actual observations that carry their own uncertainties (e.g., Fujiwara et al., 2017). Additionally the underlying models are typically structurally similar to the models being benchmarked with those reanalysis products.

A related challenge is the selection of the most suitable data sets for the evaluation of models. Different kinds of models are intended to capture different processes at different spatial and temporal scales. Observational data collected at a single location or at high frequency may not be suitable for comparison with a simulation for a climate model grid cell because the location may not be representative of that grid cell or the data may contain variations not intended to be captured by a given model. In addition, analysts may select data sets based on the convenience of their use, their ease of accessibility, or the format of the data. To assist the research community in accessing and utilizing a diversity of observational data products for the assessment of models, projects such as obs4MIPs (Waliser et al., 2020) and CREATE-IP (Potter et al., 2018) were instituted to adapt observational data to the well-established format used by the modeling community and to provide a clearinghouse of the resulting products through ESGF, which is the same portal used for distribution of CMIP model output. However, the data produced in projects like obs4MIPs and CREATE-IP are not exhaustive and suffer from a lack of routine updates. As a result, the benchmarking community often prioritizes synthesizing their own data sets, which may be incomplete and less accessible for reuse. Renewed interest and effort in growing the obs4MIPs data collection is opening up contribution policies so that it can become more useful for benchmarking climate models.

Scientists recognize such data challenges and in recent decades initiatives such as the ESA Climate Change Initiative (CCI) have been proposed (Plummer et al., 2017). ESA CCI has been conceived to bridge the long-standing gap between observational and climate modeling communities, still existing, despite being progressively closed by an increasing number of researchers working at the interface of the two communities (e.g., the Cloud

Feedback Model Intercomparison Project (CFMIP) Observation Simulator Package (COSP) (Bodas-Salcedo et al., 2011)). More specifically, the goal of the ESA CCI project is to bring the observational (primarily space-borne) and climate research communities together to create reliable and temporally complete data for essential climate variables. Another example of cross-fertilization is the creation and regular improvements of the Global Space-based Stratospheric Aerosol Climatology (GloSSAC) data set (Kovilakam et al., 2020). While these initiatives help to create more harmonized data, they are still constrained in time and often limited to certain observation types (e.g., satellite remote sensing platforms).

All of the issues described above influence the results of different metrics and diagnostics used for model evaluation and benchmarking. For example, by selecting only one data set as the reference, the results may be biased and the conclusions about model performance might be drawn without more thorough consideration. It is therefore necessary to keep the observational data sets used and their uncertainties in mind when interpreting results of model evaluation analyses. Additionally, it is important to be aware of the fact that each chosen metric and diagnostic is tailored toward a specific scientific question and that the achieved model scores are only a representation of the model performance for that specific metric in combination with the chosen observation(s).

8. Discussion/Conclusion

With each successive phase of CMIP comes a wave of analyses of climate model simulations. These analyses originate from model development activities, model intercomparison studies, and increasingly from studies that extend beyond the traditional climate modeling activities. This proliferation of investigation has required the development of software tools tailored toward the purposes of model benchmarking. These tools vary in form and function, from bespoke analysis for an individual publication to packages used during model development to large and organized software activities aimed toward streamlined multi-model comparisons.

There is a large and continuous effort to improve and move the code of evaluation tools toward modern coding practices. In recent years, it appears that there is an emerging convention to use Python for developing these tools. The use of Python for model benchmarking has mirrored the larger data science trend toward Python-based tools and open-source software. Although there is momentum behind the adoption of Python as a common language for climate model analysis, many individual researchers and existing code bases rely on other languages such as R, NCL and Matlab (among many others). Similarly, while it has become common practice to provide analysis code that supports individual studies (e.g., using GitHub repositories), not all researchers provide their code openly. Even when the code is available, there is no assurance of code quality, portability, scalability, or documentation, all of which can inhibit the use of the code. These impediments often lead to researchers implementing their own version of particular diagnostics of interest. Without the original code, or at least the numerical results from the original analysis, it can be difficult to compare new results with old ones.

A number of collaborative community-driven projects have stepped forward to help alleviate the pressure of producing and maintaining model benchmarking codes. These open-source efforts, such as ESMValTool, PMP, and the MDTF package developed by NOAA collect individual contributions into common frameworks. These collections allow model development teams and individual researchers the ability to reproduce well-documented diagnostics for the various model benchmarking activities we have outlined above. Similarly, communities of practice have started to emerge to support researchers and promote ecosystems of tools that reinforce the move toward a common “stack”; the Pangeo community is one influential example.

Even with these projects and resources, there are still bumps in the road for any of the model benchmarking activities we have described. There is always some resistance in adopting new tools. Sometimes there are challenges with documentation or software environments, but it can also be a cultural challenge to introduce new diagnostics that are not familiar to a group. There can also be skepticism around whether new data sets or diagnostic codes will be maintained moving forward. A key challenge for multi-model comparisons is to compute diagnostics for all available models or simulations in a reusable form, especially when this would entail downloading large data sets. And while it is already helpful to have the diagnostics code freely available for all members of the community, unstable internet connections, power outages and just overall fewer large data storage capabilities put especially members of the Global South community at a clear disadvantage.

Some of the challenges can be addressed through community-driven efforts that remove, for example, much of the computational burden of reproducing diagnostics. There are already some efforts underway to provide online tools for accessing pre-computed metrics across CMIP models using documented benchmark observations, realized for example, in ILAMB or PMP. Both provide an interactive interface linking performance metrics with their underlying diagnostics and direct access to the statistics. Another example for addressing some of the challenges is the automatic download function of ESMValTool that allows the user to specify exactly which data would be needed for a diagnostic and if they are not available on the local storage system they are automatically downloaded from the closest (or easiest to access) ESGF node. Additionally, provenance records, including information about data and software versioning, origin and processing, are now generated with most of the open-source evaluation tools to clearly track the used data, their handling, and modification steps.

By providing well-documented analysis code, pre-computed results from CMIP, and access to curated observational data, researchers and model development centers can focus energy on the analysis of new results or model versions or extending current understanding of the climate system instead of re-implementing or re-computing previous results. The practice of maintaining these code and data repositories, including routinely updating observations, is a critical issue that must be supported and valued. Developing new diagnostics, contributing code to these common collections, and maintaining the code should similarly be seen as a foundational contribution to the model benchmarking endeavor. Ideally, there would be regular meetings and discussions between the different evaluation and benchmarking software providers to coordinate the ongoing and planned developments. However, there is nothing like this in place yet.

Following these examples and the vision of routine evaluation for new CMIP simulations from EY19, and recognizing the needs of the community for easy access to CMIP evaluations and benchmarks, the CMIP Panel followed the suggestion from the CMIP Model Benchmarking Task Team to develop a Rapid Evaluation Framework (REF) in time for the Assessment Fast Track simulations (Dunne et al., 2025; Hoffman, Hassler, et al., 2025). The idea is to have the simulations evaluated with publicly available and well-formatted observational data (through obs4MIPs; Waliser et al. (2020)) as soon as they are published on ESGF, and the results then published on a publicly available website for the community to see. Additionally, preprocessed data provided by the different diagnostics can be downloaded directly from ESGF which will facilitate easier data access for community members with low internet bandwidth or intermittent internet access. With this the REF contributes to the CMIP efforts of providing a sustainable structure for regularly updating climate data for a variety of users (Hewitt et al., 2025). For this framework to be established in a relatively short period of time, it is planned for it to leverage the ESGF infrastructure for computing and storage capabilities and simulation access, the CMEC framework for a pre-defined interface, and the diagnostic calculation capabilities in existing open-source evaluation tools. The exact diagnostic definitions that will be available in the first instance of the REF were finalized in a community survey (CMIP Model Benchmarking Task Team, 2024). Diagnostics and metrics will only be based on monthly mean values, and they will be divided in five different thematic groups: atmosphere, ocean and sea ice, land and land ice, impacts and adaptation, and Earth system. With this it follows the CMIP7 data request division into exactly the same thematic groups (Dingley et al., 2025; Fox-Kemper et al., 2025; Y. Li et al., 2025; McPartland et al., 2025; Ruane et al., 2025). The REF is designed to be modular so that it will be easy to include additional diagnostics or even additional evaluation tools. To increase its usability in the community, it will also be available as containerized version so that, for example, modeling centers can use it in their simulation production pipeline to assess their simulations before they are even submitted to ESGF. More information about the REF can be found in the REF GitHub repository (Lewis et al., 2025).

Model evaluation and benchmarking has evolved over time, through the different phases of CMIP, in close interconnection with the development of climate models. With each newly added model component (e.g., the capability to calculate emission-driven rather than concentration-driven simulations (Sanderson et al., 2024)), change in resolution (e.g., km-scale model simulations (Schär et al., 2020)), change in complexity, and change in general model development technique (e.g., implementation of ML-based model components (Eyring et al., 2024)), existing methods of evaluating the models need to be adjusted, new methods need to be invented, or new observational data sets need to be made available or even created. AI and ML methods are increasingly being used in climate modeling is different ways—in specific components, in parameterizations or as hybrid models. With this novelty in modeling, there is also the accompanying challenge of ensuring robust evaluations around questions such as whether climate models with AI components continue to adhere to laws of conservation, simulate long term stability and perhaps most importantly can be explained by our physical understanding of the

Earth system. We therefore envisage that model evaluation and analysis will play a significant role in our ability to trust and deploy AI in climate models for variety of end users.

All of the above advances make the need for routinely evaluating simulations, with open-source and community-developed diagnostics more and more relevant and urgent. Many important and pioneering efforts toward this ambition have been implemented and brought to life already for CMIP6 (e.g.., EY19), and more advances are planned for the implementation for CMIP7. The open-source evaluation and benchmarking tools play a very important role in this endeavor to provide timely and transparent scientific results for the community to assess the latest available climate simulations.

Appendix A: Open-Source, Community Evaluation, and Benchmarking Tools

Table A1

Table A1
Open-Source, Community Evaluation, and Benchmarking Tools That Are Available in the Tool Collection of the CMIP Website

		Functionality			
Tool name	Primary focus	E	B	Lead development institution(s)	Reference
AMET	Provides a method for evaluating meteorological and air quality model predictions	X		EPA (USA)	Appel et al. (2011)
ASoP	Precipitation	X		Met Office (UK)	Klingaman et al. (2017)
Atmospheric Radiation Measurement (ARM)-DIAGS	Facilitate the use of long-term high-frequency measurements from the ARM program in evaluating the simulation of clouds, radiation, and precipitation	X		LLNL (USA)	C. Zhang et al. (2021)
Automated Model Benchmarking R Package (AMBER)	Land and hydrology	X	X	ECCC (Canada)	Seiler et al. (2021)
Coordinated Model Evaluation Capabilities (CMEC)	Framework for collective operations of different packages	X		LLNL (USA)	Ordonez (2023)
CVDP, CVDP-LE	Climate modes of variability	X	X	NCAR (USA)	Phillips et al. (2014)
ESMValTool	A community diagnostic and performance metrics tool for evaluation of Earth system models in CMIP and other intercomparison projects	X	X	DLR (Germany) & Met Office (UK)	Righi et al. (2020), Eyring et al. (2020), Lauer et al. (2020), and Weigel et al. (2021)
Freva	Data search and analysis platform developed by the atmospheric science community for the atmospheric science community.	X	X	FU Berlin (Germany)	Kadow et al. (2021)
GCMeval	Tool to help with evaluation of climate models from the CMIP5 and CMIP6 ensembles	X	X	Norwegian Meteorological Institute (Norway)	Parding et al. (2020)
Integrated Assessment Models of global climate change (IAM)	Inter-linkages between the human and the natural system	X		Potsdam Institute for Climate Impact Research (Germany)	Schwanitz (2013)
International Land Model Benchmarking package (ILAMB)	Focusing primarily on biogeochemistry and hydrology, the package examines model performance through a variety of statistical error measures by comparison with contemporary observational data and produces a wide range of graphical output for exploring model uncertainty	X	X	ORNL (USA)	Collier et al. (2018)

Table A1
Continued

		Functionality			
Tool name	Primary focus	E	B	Lead development institution(s)	Reference
International Ocean Model Benchmarking package (IOMB)	Evaluates the fidelity of ocean biogeochemistry models and provides scores and graphical diagnostics	X	X	UC Irvine, ORNL (USA)	Fu et al. (2022)
Model Diagnostics Task Force (MDTF)—Diagnostics Package	A portable framework for running process-oriented diagnostics (PODs) on weather and climate model data. Each POD targets a specific physical process or emergent behavior, with the goals of determining how accurately the model represents that process	X		NOAA (USA)	Neelin et al. (2023)
Model Evaluation Tools (MET)	Numerical weather forecast verification and evaluation	X		NCAR, NOAA, USAF (USA)	Brown et al. (2021)
PCMDI Metrics Package (PMP)	Evaluate physical climate with focus on atmosphere, climatology, climate variability (ENSO, extra-tropical modes of variability, monsoon, MJO), extreme, and cloud feedback	X	X	LLNL (USA)	J. Lee et al. (2024)
RCMES	Regional model evaluation tool developed for the CORDEX community	X	X	NASA-JPL (USA)	H. Lee et al. (2018)
SITool	Evaluate the model skills in simulating the bi-polar sea ice concentration, extent, edge location, thickness, snow depth, and sea ice drift	X	X	Université catholique de Louvain (Belgium)	Lin et al. (2021)
TheDiaTo v1.0	A Thermodynamic Diagnostic Tool for model diagnostics of the thermodynamics in the climate systems (energy, water mass, entropy, enthalpy, energetics)	X	X	Hamburg University (Germany)	Lembo et al. (2019)

Note. E, Evaluation; B, Benchmarking.

Acknowledgments

Work of Birgit Hassler is supported by the project "Earth System Models for the Future" (ESM2025; grant agreement no. 101003536) under the European Union's Horizon 2020 Research and Innovation Programme. Work of Forrest Hoffman was supported by the Reconciling Uncertainties in Biogeochemical Interactions through Synthesis and Computation (RUBISCO) Science Focus Area, which is sponsored by the Regional and Global Model Analysis (RGMA) activity of the Earth & Environmental Systems Modeling (EESM) Program in the Earth and Environmental Systems Sciences Division (EESSD) of the Office of Biological and Environmental Research (BER) in the US Department of Energy Office of Science. Oak Ridge National Laboratory (ORNL) is managed by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). Work of Jiwoo Lee is performed under the auspices of the DOE by Lawrence Livermore National Laboratory (LLNL) under Contract No. DE-AC52-07NA27344 and effort was supported by the RGMA program of the U.S. DOE's Office of Science (OS), BER program. Work of Enrico Scoccimarro is partially funded by the EU project Blueadapt Grant agreement ID: 101057764. Valerio Lembo acknowledges funding from the European Union's Horizon Europe research and innovation programme under OptimESM (Grant agreement No. 101081193). J. Lu is supported by the National Key Scientific and Technological Infrastructure Project "Earth System Science Numerical Simulator Facility" (Earthlab) of China. Ranjini Swaminathan's work is supported by the UK Research Infrastructure Natural Environment Research Council (UKRI-NERC) funded TerraFIRMA: Future Impacts, Risks and Mitigation Actions in a changing Earth system Grant (NE/W004895/1). Ed Blockley was supported by the Met Office Hadley Centre Climate Programme funded by DSIT. R. Beadling was supported under NSF Division of Polar Programs Grant NSF2319828 and NOAA award Number NA24OARX431C0057-T1-01 Brian Medeiros acknowledges support by the U.S. DOE, OS, Office of BER, RGMA component of the EESM Program under Award Number DE-SC0022070 and National Science Foundation (NSF) IA 1947282, and by the National Center for Atmospheric Research (NCAR), which is a major facility sponsored by the NSF under Cooperative Agreement No. 1852977. We would like to thank the Research and Reference Services team at ECCC Library Services for their assistance in conducting thorough literature reviews, including Mary Orlik, Kaley Young and Lindsay Adoranti. Support, including community facilitation and consultation, and commissioning of Figures 1, 2, and 4

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

All CMIP5 and CMIP6 output used for Figures 3 and 5–7 are available freely and publicly from the Earth System Grid Federation (ESGF; ESGF Federated Nodes, 2025). More information on the simulations, variables, and analysis code used to generate the figures can be found in the following Zenodo archives: Hassler (2025) (Figure 3), J. Lee (2025) (Figure 5), Hoffman, Collier, et al. (2025) (Figure 6), Hassler and Bock (2025) (Figure 7). Information for Table A1 is publicly available in the Tool Table from the CMIP Model Benchmarking Task Team (Model Benchmarking Task Team, 2025). All tools referenced in Table A1 are open-source, and publicly available. Please see the references within Table A1 for each tool's source code. The source code for the software tools (ESMValTool, PMP and ILAMB) that have been used to produce Figures 3 and 5–7 is fully open-source and is available in the following Github repositories, respectively: Andela et al. (2025), J. Lee et al. (2025), and Collier (2025).

References

Ahn, M.-S., Gleckler, P. J., Lee, J., Pendergrass, A. G., & Jakob, C. (2022). Benchmarking simulated precipitation variability amplitude across time scales. *Journal of Climate*, *35*(20), 3173–3196. https://doi.org/10.1175/JCLI-D-21-0542.1

Ahn, M.-S., Ullrich, P. A., Gleckler, P. J., Lee, J., Ordonez, A. C., & Pendergrass, A. G. (2023). Evaluating precipitation distributions at regional scales: A benchmarking framework and application to CMIP5 and 6 models. *Geoscientific Model Development*, *16*(13), 3927–3951. https://doi.org/10.5194/gmd-16-3927-2023

Allen, M. R., & Ingram, W. J. (2002). Constraints on future changes in climate and the hydrologic cycle. *Nature*, *419*(6903), 224–232. https://doi.org/10.1038/nature01092

Almazroui, M., Ashfaq, M., Islam, M. N., Rashid, I. U., Kamil, S., Abid, M. A., et al. (2021). Assessment of CMIP6 performance and projected temperature and precipitation changes over South America. *Earth Systems and Environment*, *5*(2), 155–183. https://doi.org/10.1007/s41748-021-00233-6

Andela, B., Broetz, B., de Mora, L., Drost, N., Eyring, V., Koldunov, N., et al. (2025). ESMValTool [Software]. *Zenodo*. https://doi.org/10.5281/zenodo.3401363

Appel, K. W., Gilliam, R. C., Davis, N., Zubrow, A., & Howard, S. C. (2011). Overview of the atmospheric model evaluation tool (AMET) v1.1 for evaluating meteorological and air quality models. *Environmental Modelling & Software*, *26*(4), 434–443. https://doi.org/10.1016/j.envsoft.2010.09.007

Baker, A. H., Hammerling, D. M., Levy, M. N., Xu, H., Dennis, J. M., Eaton, B. E., et al. (2015). A new ensemble-based consistency test for the Community Earth System Model (pyCECT v1.0). *Geoscientific Model Development*, *8*(9), 2829–2840. https://doi.org/10.5194/gmd-8-2829-2015

Baker, J. C. A., & Spracklen, D. V. (2022). Divergent representation of precipitation recycling in the Amazon and the Congo in CMIP6 models. *Geophysical Research Letters*, *49*(10), e2021GL095136. https://doi.org/10.1029/2021GL095136

Bengtsson, L., & Hodges, K. I. (2011). On the evaluation of temperature trends in the tropical troposphere. *Climate Dynamics*, *36*(3–4), 419–430. https://doi.org/10.1007/s00382-009-0680-y

Best, M. J., Abramowitz, G., Johnson, H. R., Pitman, A. J., Balsamo, G., Boone, A., et al. (2015). The plumbing of land surface models: Benchmarking model performance. *Journal of Hydrometeorology*, *16*(3), 1425–1442. https://doi.org/10.1175/JHM-D-14-0158.1

Bjarke, N., Barsugli, J., & Livneh, B. (2023). Ensemble of CMIP6 derived reference and potential evapotranspiration with radiative and advective components. *Scientific Data*, *10*(1), 417. https://doi.org/10.1038/s41597-023-02290-0

Bock, L., & Lauer, A. (2024). Cloud properties and their projected changes in CMIP models with low to high climate sensitivity. *Atmospheric Chemistry and Physics*, *24*(3), 1587–1605. https://doi.org/10.5194/acp-24-1587-2024

Bock, L., Lauer, A., Schlund, M., Barreiro, M., Bellouin, N., Jones, C., et al. (2020). Quantifying progress across different CMIP phases with the ESMValTool. *Journal of Geophysical Research: Atmospheres*, *125*(21), e2019JD032321. https://doi.org/10.1029/2019JD032321

Bodas-Salcedo, A., Webb, M. J., Bony, S., Chepfer, H., Dufresne, J.-L., Klein, S. A., et al. (2011). COSP: Satellite simulation software for model assessment. *Bulletin of the American Meteorological Society*, *92*(8), 1023–1043. https://doi.org/10.1175/2011BAMS2856.1

Boo, K.-O., Martin, G., Sellar, A., Senior, C., & Byun, Y.-H. (2011). Evaluating the East Asian monsoon simulation in climate models. *Journal of Geophysical Research*, *116*(D1), D01109. https://doi.org/10.1029/2010JD014737

Boucher, O., Servonnat, J., Albright, A. L., Aumont, O., Balkanski, Y., Bastrikov, V., et al. (2020). Presentation and evaluation of the IPSL-CM6A-LR climate model. *Journal of Advances in Modeling Earth Systems*, *12*(7), e2019MS002010. https://doi.org/10.1029/2019MS002010

Boville, B. A., & Gent, P. R. (1998). The NCAR Climate System Model, Version One. *Journal of Climate*, *11*(6), 1115–1130. https://doi.org/10.1175/1520-0442(1998)011⟨1115:TNCSMV⟩2.0.CO;2

Brekke, L. D., Dettinger, M. D., Maurer, E. P., & Anderson, M. (2008). Significance of model credibility in estimating climate projection distributions for regional hydroclimatological risk assessments. *Climatic Change*, *89*(3–4), 371–394. https://doi.org/10.1007/s10584-007-9388-3

Brient, F. (2020). Reducing uncertainties in climate projections with emergent constraints: Concepts, examples and prospects. *Advances in Atmospheric Sciences*, *37*(1), 1–15. https://doi.org/10.1007/s00376-019-9140-8

Brinkop, S., & Roeckner, E. (1995). Sensitivity of a general circulation model to parameterizations of cloud-turbulence interactions in the atmospheric boundary layer. *Tellus A*, *47*(2), 197–220. https://doi.org/10.3402/tellusa.v47i2.11501

Brown, B., Jensen, T., Gotway, J. H., Bullock, R., Gilleland, E., Fowler, T., et al. (2021). The Model Evaluation Tools (MET): More than a decade of community-supported forecast verification. *Bulletin of the American Meteorological Society*, *102*(4), E782–E807. https://doi.org/10.1175/BAMS-D-19-0093.1

Brunner, L., Pendergrass, A. G., Lehner, F., Merrifield, A. L., Lorenz, R., & Knutti, R. (2020). Reduced global warming from CMIP6 projections when weighting models by performance and independence. *Earth System Dynamics*, *11*(4), 995–1012. https://doi.org/10.5194/esd-11-995-2020

by Miranta Kouvari from Science Graphic Design, has been provided by the CMIP International Project Office, which is hosted by the European Space Agency, with staff provided by HE Space Operations Ltd. We would also like to thank Feba Francis for her help with collecting references for this manuscript, and André L. Marquez for input on observational data. We thank V. Eyring and K. Taylor for very helpful comments and suggestions on an early draft of the paper. We also thank S. Brands for very useful discussions about the definition of the terms "Model Verification," "Process Validation," "Evaluation" and "Benchmarking" during the revision process of the manuscript. We thank three anonymous reviewers for very helpful comments and suggestions during the review process. Open Access funding enabled and organized by Projekt DEAL.

Canadell, J., Monteiro, P., Costa, M., Cotrim da Cunha, L., Cox, P., Eliseev, A., et al. (2021). Global carbon and other biogeochemical cycles and feedbacks. In *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 673–816). Cambridge University Press.

Carton, J. A., Chepurin, G. A., & Chen, L. (2018). SODA3: A new ocean climate reanalysis. *Journal of Climate*, *31*(17), 6967–6983. https://doi.org/10.1175/JCLI-D-18-0149.1

Casas, M. C., Schmidt, G. A., Miller, R. L., Orbe, C., Tsigaridis, K., Nazarenko, L. S., et al. (2023). Understanding model-observation discrepancies in satellite retrievals of atmospheric temperature using GISS ModelE. *Journal of Geophysical Research: Atmospheres*, *128*(1), e2022JD037523. https://doi.org/10.1029/2022JD037523

Ceppi, P., & Nowack, P. (2021). Observational evidence that cloud feedback amplifies global warming. *Proceedings of the National Academy of Sciences*, *118*(30), e2026290118. https://doi.org/10.1073/pnas.2026290118

Cesana, G. V., Ackerman, A. S., Črnivec, N., Pincus, R., & Chepfer, H. (2023). An observation-based method to assess tropical stratocumulus and shallow cumulus clouds and feedbacks in CMIP6 and CMIP5 models. *Environmental Research Communications*, *5*(4), 045001. https://doi.org/10.1088/2515-7620/acc78a

CMIP Model Benchmarking Task Team. (2024). CMIP7 assessment fast track diagnostics list for the rapid evaluation framework [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.14284375

Collier, N. (2025). GitHub - Rubisco-sfa/ILAMB: Python software used in the International Land Model Benchmarking (ILAMB) project [Software]. *Github*. Retrieved from https://github.com/rubisco-sfa/ILAMB

Collier, N., Hoffman, F. M., Lawrence, D. M., Keppel-Aleks, G., Koven, C. D., Riley, W. J., et al. (2018). The International Land Model Benchmarking (ILAMB) system: Design, theory, and implementation. *Journal of Advances in Modeling Earth Systems*, *10*(11), 2731–2754. https://doi.org/10.1029/2018MS001354

Collins, W. J., Lamarque, J.-F., Schulz, M., Boucher, O., Eyring, V., Hegglin, M. I., et al. (2017). AerChemMIP: Quantifying the effects of chemistry and aerosols in CMIP6. *Geoscientific Model Development*, *10*(2), 585–607. https://doi.org/10.5194/gmd-10-585-2017

Covey, C., Gleckler, P. J., Doutriaux, C., Williams, D. N., Dai, A., Fasullo, J., et al. (2016). Metrics for the diurnal cycle of precipitation: Toward routine benchmarks for climate models. *Journal of Climate*, *29*(12), 4461–4471. https://doi.org/10.1175/JCLI-D-15-0664.1

Cox, P. M., Huntingford, C., & Williamson, M. S. (2018). Emergent constraint on equilibrium climate sensitivity from global temperature variability. *Nature*, *553*(7688), 319–322. https://doi.org/10.1038/nature25450

Dai, A. (2006). Precipitation characteristics in eighteen coupled climate models. *Journal of Climate*, *19*(18), 4605–4630. https://doi.org/10.1175/JCLI3884.1

Dai, P., Nie, J., Yu, Y., & Wu, R. (2024). Constraints on regional projections of mean and extreme precipitation under warming. *Proceedings of the National Academy of Sciences*, *121*(11), e2312400121. https://doi.org/10.1073/pnas.2312400121

Danabasoglu, G., McWilliams, J. C., & Gent, P. R. (1994). The role of mesoscale tracer transports in the global ocean circulation. *Science*, *264*(5162), 1123–1126. https://doi.org/10.1126/science.264.5162.1123

Das, P., & Ganguly, A. R. (2025). Finer resolutions and targeted process representations in Earth system models improve hydrologic projections and hydroclimate impacts. *npj Climate and Atmospheric Science*, *8*(247), 247. https://doi.org/10.1038/s41612-025-01134-5

Davis, I., & Medeiros, B. (2024). Assessing CESM2 clouds and their response to climate change using cloud regimes. *Journal of Climate*, *37*(10), 2965–2985. https://doi.org/10.1175/JCLI-D-23-0337.1

Davis, S. M., Rosenlof, K. H., Hassler, B., Hurst, D. F., Read, W. G., Vömel, H., et al. (2016). The Stratospheric Water and Ozone Satellite Homogenized (SWOOSH) database: A long-term database for climate studies. *Earth System Science Data*, *8*(2), 461–490. https://doi.org/10.5194/essd-8-461-2016

De Dominicis, M., Wolf, J., Jevrejeva, S., Zheng, P., & Hu, Z. (2020). Future interactions between sea level rise, tides, and storm surges in the world's largest urban area. *Geophysical Research Letters*, *47*(4), e2020GL087002. https://doi.org/10.1029/2020GL087002

Deser, C., Lehner, F., Rodgers, K. B., Ault, T., Delworth, T. L., DiNezio, P. N., et al. (2020). Insights from Earth system model initial-condition large ensembles and future prospects. *Nature Climate Change*, *10*(4), 277–286. https://doi.org/10.1038/s41558-020-0731-2

Dingley, B., Anstey, J. A., Abalos, M., Abraham, C., Bergman, T., Bock, L., et al. (2025). CMIP7 data request: Atmosphere priorities and opportunities. *EGUsphere*, *2025*, 1–54. https://doi.org/10.5194/egusphere-2025-3189

Dolinar, E. K., Dong, X., Xi, B., Jiang, J. H., & Su, H. (2015). Evaluation of CMIP5 simulated clouds and TOA radiation budgets using NASA satellite observations. *Climate Dynamics*, *44*(7–8), 2229–2247. https://doi.org/10.1007/s00382-014-2158-9

Donohoe, A., Fajber, R., Cox, T., Armour, K. C., Battisti, D. S., & Roe, G. H. (2024). Model biases in the atmosphere-ocean partitioning of poleward heat transport are persistent across three CMIP generations. *Geophysical Research Letters*, *51*(8), e2023GL106639. https://doi.org/10.1029/2023GL106639

Douglass, D. H., Christy, J. R., Pearson, B. D., & Singer, S. F. (2008). A comparison of tropical temperature trends with model predictions. *International Journal of Climatology*, *28*(13), 1693–1701. https://doi.org/10.1002/joc.1651

Duffy, M. L., Medeiros, B., Gettelman, A., & Eidhammer, T. (2024). Perturbing parameters to understand cloud contributions to climate change. *Journal of Climate*, *37*(1), 213–227. https://doi.org/10.1175/JCLI-D-23-0250.1

Dunne, J. P., Hewitt, H. T., Arblaster, J., Bonou, F., Boucher, O., Cavazos, T., et al. (2025). An evolving Coupled Model Intercomparison Project phase 7 (CMIP7) and Fast Track in support of future climate assessment. *Geoscientific Model Development*, *18*(19), 6671–6700. https://doi.org/10.5194/gmd-18-6671-2025

Durack, P. J., Taylor, K. E., Gleckler, P. J., Meehl, G. A., Lawrence, B. N., Covey, C., et al. (2025). The Coupled Model Intercomparison Project (CMIP): Reviewing project history, evolution, infrastructure and implementation. *EGUsphere*, *2025*, 1–74. https://doi.org/10.5194/egusphere-2024-3729

Ebtehaj, I., & Bonakdari, H. (2023). A comprehensive comparison of the fifth and sixth phases of the coupled model intercomparison project based on the Canadian Earth system models in spatio-temporal variability of long-term flood susceptibility using remote sensing and flood frequency analysis. *Journal of Hydrology*, *617*, 128851. https://doi.org/10.1016/j.jhydrol.2022.128851

Edwards, P. N. (2011). History of climate modeling. *WIREs Climate Change*, *2*(1), 128–139. https://doi.org/10.1002/wcc.95

Elsaesser, G. S., van Lier-Walqui, M., Yang, Q., Kelley, M., Ackerman, A. S., Fridlind, A. M., et al. (2025). Using machine learning to generate a GISS ModelE Calibrated Physics Ensemble (CPE). *Journal of Advances in Modeling Earth Systems*, *17*(4), e2024MS004713. https://doi.org/10.1029/2024MS004713

Eman, K., Chung, E., & Ayugi, B. O. (2024). Investigating the skills of HighResMIP in capturing historical and future mean precipitation shifts over Pakistan. *International Journal of Climatology*, *44*(11), joc.8558–3911. https://doi.org/10.1002/joc.8558

ESGF Federated Nodes. (2025). ESGF MetaGrid - Version: V1.5.2. Retrieved from https://esgf-metagrid.cloud.dkrz.de/search/cmip6-dkrz/

ESMValTool Development Team. (2025). ESMValTool 2.14.0.dev36+gb86071480 documentation — docs.esmvaltool.org. Retrieved from https://docs.esmvaltool.org/en/latest/

Eyring, V., Bock, L., Lauer, A., Righi, M., Schlund, M., Andela, B., et al. (2020). Earth System Model Evaluation Tool (ESMValTool) v2.0 – An extended set of large-scale diagnostics for quasi-operational and comprehensive evaluation of Earth system models in CMIP. *Geoscientific Model Development*, *13*(7), 3383–3438. https://doi.org/10.5194/gmd-13-3383-2020

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

Eyring, V., Cox, P. M., Flato, G. M., Gleckler, P. J., Abramowitz, G., Caldwell, P., et al. (2019). Taking climate model evaluation to the next level. *Nature Climate Change*, *9*(2), 102–110. https://doi.org/10.1038/s41558-018-0355-y

Eyring, V., Gentine, P., Camps-Valls, G., Lawrence, D. M., & Reichstein, M. (2024). AI-empowered next-generation multiscale climate modelling for mitigation and adaptation. *Nature Geoscience*, *17*(10), 963–971. https://doi.org/10.1038/s41561-024-01527-w

Eyring, V., Gillett, N. P., Achuta Rao, K. M., Barimalala, R., Barreiro Parrillo, M., Bellouin, N., et al. (2021). Human influence on the climate system. In *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment report of the Intergovernmental Panel on Climate Change* (pp. 423–552). Cambridge University Press. https://doi.org/10.1017/9781009157896.005

Eyring, V., Gleckler, P. J., Heinze, C., Stouffer, R. J., Taylor, K. E., Balaji, V., et al. (2016). Towards improved and more routine Earth system model evaluation in CMIP. *Earth System Dynamics*, *7*(4), 813–830. https://doi.org/10.5194/esd-7-813-2016

Eyring, V., Righi, M., Lauer, A., Evaldsson, M., Wenzel, S., Jones, C., et al. (2016). ESMValTool (v1.0) – A community diagnostic and performance metrics tool for routine evaluation of Earth system models in CMIP. *Geoscientific Model Development*, *9*(5), 1747–1802. https://doi.org/10.5194/gmd-9-1747-2016

Fasullo, J. T. (2020). Evaluating simulated climate patterns from the CMIP archives using satellite and reanalysis datasets using the Climate Model Assessment Tool (CMATv1). *Geoscientific Model Development*, *13*(8), 3627–3642. https://doi.org/10.5194/gmd-13-3627-2020

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W., et al. (2013). Evaluation of climate models. In *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press. Retrieved from https://www.ipcc.ch/site/assets/uploads/2018/02/WG1AR5_Chapter09_FINAL.pdf

Fox-Kemper, B., DeRepentigny, P., Treguier, A. M., Stepanek, C., O'Rourke, E., Mackallah, C., et al. (2025). CMIP7 data request: Ocean and sea ice priorities and opportunities. *EGUsphere*, *2025*, 1–58. https://doi.org/10.5194/egusphere-2025-3083

Freedman, F. R., Pitts, K. L., & Bridger, A. F. (2014). Evaluation of CMIP climate model hydrological output for the Mississippi River Basin using GRACE satellite observations. *Journal of Hydrology*, *519*, 3566–3577. https://doi.org/10.1016/j.jhydrol.2014.10.036

Fu, W., Moore, J. K., Primeau, F., Collier, N., Ogunro, O. O., Hoffman, F. M., & Randerson, J. T. (2022). Evaluation of ocean biogeochemistry and carbon cycling in CMIP Earth system models with the International Ocean Model Benchmarking (IOMB) software system. *Journal of Geophysical Research: Oceans*, *127*(10), e2022JC018965. https://doi.org/10.1029/2022JC018965

Fujiwara, M., Wright, J. S., Manney, G. L., Gray, L. J., Anstey, J., Birner, T., et al. (2017). Introduction to the SPARC Reanalysis Intercomparison Project (S-RIp) and overview of the reanalysis systems. *Atmospheric Chemistry and Physics*, *17*(2), 1417–1452. https://doi.org/10.5194/acp-17-1417-2017

Galytska, E., Weigel, K., Handorf, D., Jaiser, R., Köhler, R., Runge, J., & Eyring, V. (2023). Evaluating causal Arctic-midlatitude teleconnections in CMIP6. *Journal of Geophysical Research: Atmospheres*, *128*(17), e2022JD037978. https://doi.org/10.1029/2022JD037978

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., et al. (2017). The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *Journal of Climate*, *30*(14), 5419–5454. https://doi.org/10.1175/JCLI-D-16-0758.1

Gent, P. R. (2011). The Gent–McWilliams parameterization: 20/20 hindsight. *Ocean Modelling*, *39*(1–2), 2–9. https://doi.org/10.1016/j.ocemod.2010.08.002

Gent, P. R., & McWilliams, J. C. (1990). Isopycnal mixing in ocean circulation models. *Journal of Physical Oceanography*, *20*(1), 150–155. https://doi.org/10.1175/1520-0485(1990)020⟨0150:IMIOCM⟩2.0.CO;2

Gibson, P. B., Perkins-Kirkpatrick, S. E., Uotila, P., Pepler, A. S., & Alexander, L. V. (2017). On the use of self-organizing maps for studying climate extremes. *Journal of Geophysical Research: Atmospheres*, *122*(7), 3891–3903. https://doi.org/10.1002/2016JD026256

Gier, B. K., Buchwitz, M., Reuter, M., Cox, P. M., Friedlingstein, P., & Eyring, V. (2020). Spatially resolved evaluation of Earth system models with satellite column-averaged CO_2. *Biogeosciences*, *17*(23), 6115–6144. https://doi.org/10.5194/bg-17-6115-2020

Gier, B. K., Schlund, M., Friedlingstein, P., Jones, C. D., Jones, C., Zaehle, S., & Eyring, V. (2024). Representation of the terrestrial carbon cycle in CMIP6. *EGUsphere*, *2024*, 1–63. https://doi.org/10.5194/egusphere-2024-277

Gillett, N. P., Shiogama, H., Funke, B., Hegerl, G., Knutti, R., Matthes, K., et al. (2016). The Detection and Attribution Model Intercomparison Project (DAMIP v1.0) contribution to CMIP6. *Geoscientific Model Development*, *9*(10), 3685–3697. https://doi.org/10.5194/gmd-9-3685-2016

Gleckler, P., Doutriaux, C., Durack, P., Taylor, K., Zhang, Y., Williams, D., et al. (2016). A more powerful reality test for climate models. *Eos*, *97*. https://doi.org/10.1029/2016EO051663

Gleckler, P., Taylor, K., & Doutriaux, C. (2008). Performance metrics for climate models. *Journal of Geophysical Research*, *113*(D6), 2007JD008972. https://doi.org/10.1029/2007JD008972

Grewe, V., Moussiopoulos, N., Builtjes, P., Borrego, C., Isaksen, I. S. A., & Volz-Thomas, A. (2012). The ACCENT-protocol: A framework for benchmarking and model evaluation. *Geoscientific Model Development*, *5*(3), 611–618. https://doi.org/10.5194/gmd-5-611-2012

Guarino, M.-V., Sime, L. C., Schroeder, D., Lister, G. M. S., & Hatcher, R. (2020). Machine dependence and reproducibility for coupled climate simulations: The HadGEM3-GC3.1 CMIP Preindustrial simulation. *Geoscientific Model Development*, *13*(1), 139–154. https://doi.org/10.5194/gmd-13-139-2020

Hassler, B. (2025). Examples of metrics and diagnostics for different evaluation and benchmarking approaches [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.17671935

Hassler, B., & Bock, L. (2025). Pattern correlation plot generated by ESMValTool for selected CMIP5 and CMIP6 atmospheric variables [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.17534823

Hassler, B., Dingley, B., & Team, M. B. T. (2025). Definition of the terms model verification, process validation, evaluation and benchmarking for use in the climate model context [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.17649933

Hausfather, Z., Drake, H. F., Abbott, T., & Schmidt, G. A. (2020). Evaluating the performance of past climate model projections. *Geophysical Research Letters*, *47*(1), e2019GL085378. https://doi.org/10.1029/2019GL085378

Hawkins, E., & Sutton, R. (2009). The potential to narrow uncertainty in regional climate predictions. *Bulletin of the American Meteorological Society*, *90*(8), 1095–1108. https://doi.org/10.1175/2009BAMS2607.1

Hegerl, G. C., Black, E., Allan, R. P., Ingram, W. J., Polson, D., Trenberth, K. E., et al. (2015). Challenges in quantifying changes in the global water cycle. *Bulletin of the American Meteorological Society*, *96*(7), 1097–1115. https://doi.org/10.1175/BAMS-D-13-00212.1

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Hewitt, H. T., Flato, G., O'Rourke, E., Dunne, J. P., Adloff, F., Arblaster, J. M., et al. (2025). Towards provision of regularly updated climate data from the Coupled Model Intercomparison Project. *PloS Climate*, *4*(10), e0000708. https://doi.org/10.1371/journal.pclm.0000708

Hirsch, A. L., Ridder, N. N., Perkins-Kirkpatrick, S. E., & Ukkola, A. (2021). CMIP6 multimodel evaluation of present-day heatwave attributes. *Geophysical Research Letters*, *48*(22), e2021GL095161. https://doi.org/10.1029/2021GL095161

Hoffman, F. M., Collier, N., Xu, M., Koven, C. D., Mu, M., Lawrence, D., et al. (2025). Performance scores from a comparison of selected CMIP5 and CMIP6 land models across a variety of variables using the International Land Model Benchmarking (ILAMB) package [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.17596526

Hoffman, F. M., Hassler, B., Swaminathan, R., Lewis, J., Andela, B., Collier, N., et al. (2025). Rapid evaluation framework for the CMIP7 assessment fast track. *EGUsphere*, *2025*, 1–57. https://doi.org/10.5194/egusphere-2025-2685

Hoffman, F. M., Koven, C. D., Keppel-Aleks, G., Lawrence, D. M., Riley, W. J., Randerson, J. T., et al. (2017). *International Land Model Benchmarking (ILAMB) 2016 workshop report*. (Technical Report No. DOE/SC-0186). U.S. Department of Energy, Office of Science. https://doi.org/10.2172/1330803

Holland, M. M., Hannay, C., Fasullo, J., Jahn, A., Kay, J. E., Mills, M., et al. (2024). New model ensemble reveals how forcing uncertainty and model structure alter climate simulated across CMIP generations of the Community Earth System Model. *Geoscientific Model Development*, *17*(4), 1585–1602. https://doi.org/10.5194/gmd-17-1585-2024

Holt, L. A., Lott, F., Garcia, R. R., Kiladis, G. N., Cheng, Y., Anstey, J. A., et al. (2022). An evaluation of tropical waves and wave forcing of the QBO in the QBOi models. *Quarterly Journal of the Royal Meteorological Society*, *148*(744), 1541–1567. https://doi.org/10.1002/qj.3827

Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J.-C., Balaji, V., Duan, Q., et al. (2017). The art and science of climate model tuning. *Bulletin of the American Meteorological Society*, *98*(3), 589–602. https://doi.org/10.1175/BAMS-D-15-00135.1

Hsu, C.-W., Yin, J., Griffies, S. M., & Dussin, R. (2021). A mechanistic analysis of tropical Pacific dynamic sea level in GFDL-OM4 under OMIP-I and OMIP-II forcings. *Geoscientific Model Development*, *14*(5), 2471–2502. https://doi.org/10.5194/gmd-14-2471-2021

Huang, S., Wang, B., & Wen, Z. (2020). Dramatic weakening of the tropical easterly jet projected by CMIP6 models. *Journal of Climate*, *33*(19), 8439–8455. https://doi.org/10.1175/JCLI-D-19-1002.1

ILAMB 2.6. (2025). ilamb.org. Retrieved from https://www.ilamb.org/CMIP5v6/historical/

Infanti, J. M., Kirtman, B. P., Aumen, N. G., Stamm, J., & Polsky, C. (2020). Aligning climate models with stakeholder needs: Advances in communicating future rainfall uncertainties for South Florida decision makers. *Earth and Space Science*, *7*(7), e2019EA000725. https://doi.org/10.1029/2019EA000725

Inness, A., Ades, M., Agustí-Panareda, A., Barré, J., Benedictow, A., Blechschmidt, A.-M., et al. (2019). The CAMS reanalysis of atmospheric composition. *Atmospheric Chemistry and Physics*, *19*(6), 3515–3556. https://doi.org/10.5194/acp-19-3515-2019

IPCC-WG1. (2023). IPCC-WG1. Retrieved from https://github.com/IPCC-WG1

Jean-Michel, L., Eric, G., Romain, B.-B., Gilles, G., Angélique, M., Marie, D., et al. (2021). The Copernicus global 1/12° oceanic and sea ice GLORYS12 reanalysis. *Frontiers in Earth Science*, *9*, 698876. https://doi.org/10.3389/feart.2021.698876

Jiang, J. H., Su, H., Wu, L., Zhai, C., & Schiro, K. A. (2021). Improvements in cloud and water vapor simulations over the tropical oceans in CMIP6 compared to CMIP5. *Earth and Space Science*, *8*(5), e2020EA001520. https://doi.org/10.1029/2020EA001520

John, A., Douville, H., Ribes, A., & Yiou, P. (2022). Quantifying CMIP6 model uncertainties in extreme precipitation projections. *Weather and Climate Extremes*, *36*, 100435. https://doi.org/10.1016/j.wace.2022.100435

Jones, C. D., & Friedlingstein, P. (2020). Quantifying process-level uncertainty contributions to TCRE and carbon budgets for meeting Paris Agreement climate targets. *Environmental Research Letters*, *15*(7), 074019. https://doi.org/10.1088/1748-9326/ab858a

Jönsson, B. F., Follett, C. L., Bien, J., Dutkiewicz, S., Hyun, S., Kulk, G., et al. (2023). Using probability density functions to evaluate models (PDFEM, v1.0) to compare a biogeochemical model with satellite-derived chlorophyll. *Geoscientific Model Development*, *16*(16), 4639–4657. https://doi.org/10.5194/gmd-16-4639-2023

Kadow, C., Illing, S., Lucio-Eceiza, E. E., Bergemann, M., Ramadoss, M., Sommer, P. S., et al. (2021). Introduction to Freva – A free evaluation system framework for Earth system modeling. *Journal of Open Research Software*, *9*(1), 13. https://doi.org/10.5334/jors.253

Kang, I.-S., Jin, K., Wang, B., Lau, K.-M., Shukla, J., Krishnamurthy, V., et al. (2002). Intercomparison of the climatological variations of Asian summer monsoon precipitation simulated by 10 GCMs. *Climate Dynamics*, *19*, 383–395. https://doi.org/10.1007/s00382-002-0245-9

Kaps, A., Lauer, A., Camps-Valls, G., Gentine, P., Gomez-Chova, L., & Eyring, V. (2023). Machine-learned cloud classes from satellite data for process-oriented climate model evaluation. *IEEE Transactions on Geoscience and Remote Sensing*, *61*, 1–15. https://doi.org/10.1109/TGRS.2023.3237008

Karmouche, S., Galytska, E., Runge, J., Meehl, G. A., Phillips, A. S., Weigel, K., & Eyring, V. (2023). Regime-oriented causal model evaluation of Atlantic–Pacific teleconnections in CMIP6. *Earth System Dynamics*, *14*(2), 309–344. https://doi.org/10.5194/esd-14-309-2023

Keen, A., Blockley, E., Bailey, D. A., Boldingh Debernard, J., Bushuk, M., Delhaye, S., et al. (2021). An inter-comparison of the mass budget of the Arctic sea ice in CMIP6 models. *The Cryosphere*, *15*(2), 951–982. https://doi.org/10.5194/tc-15-951-2021

Kim, Y.-H., Min, S.-K., Zhang, X., Sillmann, J., & Sandstad, M. (2020). Evaluation of the CMIP6 multi-model ensemble for climate extreme indices. *Weather and Climate Extremes*, *29*, 100269. https://doi.org/10.1016/j.wace.2020.100269

Klingaman, N. P., Martin, G. M., & Moise, A. (2017). ASoP (v1.0): A set of methods for analyzing scales of precipitation in general circulation models. *Geoscientific Model Development*, *10*(1), 57–83. https://doi.org/10.5194/gmd-10-57-2017

Knutti, R. (2010). The end of model democracy? An editorial comment. *Climatic Change*, *102*(3–4), 395–404. https://doi.org/10.1007/s10584-010-9800-2

Knutti, R., & Hegerl, G. C. (2008). The equilibrium sensitivity of the Earth's temperature to radiation changes. *Nature Geoscience*, *1*(11), 735–743. https://doi.org/10.1038/ngeo337

Knutti, R., & Rugenstein, M. A. (2015). Feedbacks, climate sensitivity and the limits of linear models. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *373*(2054), 20150146. https://doi.org/10.1098/rsta.2015.0146

Knutti, R., Sedláček, J., Sanderson, B. M., Lorenz, R., Fischer, E. M., & Eyring, V. (2017). A climate model projection weighting scheme accounting for performance and interdependence. *Geophysical Research Letters*, *44*(4), 1909–1918. https://doi.org/10.1002/2016GL072012

Kobayashi, S., Ota, Y., Harada, Y., Ebita, A., Moriya, M., Onoda, H., et al. (2015). The JRA-55 reanalysis: General specifications and basic characteristics. *Journal of the Meteorological Society of Japan. Ser. II*, *93*(1), 5–48. https://doi.org/10.2151/jmsj.2015-001

Kovilakam, M., Thomason, L. W., Ernest, N., Rieger, L., Bourassa, A., & Millán, L. (2020). The global space-based stratospheric aerosol climatology (version 2.0): 1979–2018. *Earth System Science Data*, *12*(4), 2607–2634. https://doi.org/10.5194/essd-12-2607-2020

Lambert, S. J., & Boer, G. J. (2001). CMIP1 evaluation and intercomparison of coupled climate models. *Climate Dynamics*, *17*(2), 83–106. https://doi.org/10.1007/PL00013736

Lauer, A., Bock, L., Hassler, B., Jöckel, P., Ruhe, L., & Schlund, M. (2025). Monitoring and benchmarking Earth system model simulations with ESMValTool v2.12.0. *Geoscientific Model Development*, *18*(4), 1169–1188. https://doi.org/10.5194/gmd-18-1169-2025

Lauer, A., Bock, L., Hassler, B., Schröder, M., & Stengel, M. (2023). Cloud climatologies from global climate models—A comparison of CMIP5 and CMIP6 models with satellite data. *Journal of Climate*, *36*(2), 281–311. https://doi.org/10.1175/JCLI-D-22-0181.1

Lauer, A., Eyring, V., Bellprat, O., Bock, L., Gier, B. K., Hunter, A., et al. (2020). Earth System Model Evaluation Tool (ESMValTool) v2.0 – Diagnostics for emergent constraints and future projections from Earth system models in CMIP. *Geoscientific Model Development*, *13*(9), 4205–4228. https://doi.org/10.5194/gmd-13-4205-2020

Lauer, A., Eyring, V., Righi, M., Buchwitz, M., Defourny, P., Evaldsson, M., et al. (2017). Benchmarking CMIP5 models with a subset of ESA CCI Phase 2 data using the ESMValTool. *Remote Sensing of Environment*, *203*, 9–39. https://doi.org/10.1016/j.rse.2017.01.007

Lauer, A., & Hamilton, K. (2013). Simulating clouds with global climate models: A comparison of CMIP5 results with CMIP3 and satellite data. *Journal of Climate*, *26*(11), 3823–3845. https://doi.org/10.1175/JCLI-D-12-00451.1

Lauritzen, P. H., Kevlahan, N. K., Toniazzo, T., Eldred, C., Dubos, T., Gassmann, A., et al. (2022). Reconciling and improving formulations for thermodynamics and conservation principles in Earth System Models (ESMs). *Journal of Advances in Modeling Earth Systems*, *14*(9), e2022MS003117. https://doi.org/10.1029/2022MS003117

Lee, H., Goodman, A., McGibbney, L., Waliser, D. E., Kim, J., Loikith, P. C., et al. (2018). Regional Climate Model Evaluation System powered by Apache Open Climate Workbench v1.3.0: An enabling tool for facilitating regional climate studies. *Geoscientific Model Development*, *11*(11), 4435–4449. https://doi.org/10.5194/gmd-11-4435-2018

Lee, J. (2025). Portrait plot generated by the PCMDI Metrics Package, for seasonal climatology of multiple CMIP5 and CMIP6 variables [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.17689565

Lee, J., Chang, K., Gleckler, P., & Ullrich, P. (2025). PCMDI metrics package [Software]. *GitHub*. Retrieved from https://github.com/PCMDI/pcmdi_metrics

Lee, J., Gleckler, P. J., Ahn, M.-S., Ordonez, A., Ullrich, P. A., Sperber, K. R., et al. (2024). Systematic and objective evaluation of Earth system models: PCMDI Metrics Package (PMP) version 3. *Geoscientific Model Development*, *17*(9), 3919–3948. https://doi.org/10.5194/gmd-17-3919-2024

Lee, J., Xue, Y., De Sales, F., Diallo, I., Marx, L., Ek, M., et al. (2019). Evaluation of multi-decadal UCLA-CFSv2 simulation and impact of interactive atmospheric-ocean feedback on global and regional variability. *Climate Dynamics*, *52*(5–6), 3683–3707. https://doi.org/10.1007/s00382-018-4351-8

Lee, J.-Y., Marotzke, J., Bala, G., Cao, L., Corti, S., Dunne, J., et al. (2021). Future global climate: Scenario-based projections and near-term information. In *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 553–672). Cambridge University Press. https://doi.org/10.1017/9781009157896.006

Lehner, F., Deser, C., Maher, N., Marotzke, J., Fischer, E. M., Brunner, L., et al. (2020). Partitioning climate projection uncertainty with multiple large ensembles and CMIP5/6. *Earth System Dynamics*, *11*(2), 491–508. https://doi.org/10.5194/esd-11-491-2020

Lembo, V., Hassler, B., Dingley, B., & Team, M. B. T. (2024). Different approaches that are most commonly used for the evaluation and benchmarking of climate models [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.13934006

Lembo, V., Lunkeit, F., & Lucarini, V. (2019). TheDiaTo (v1.0) – A new diagnostic tool for water, energy and entropy budgets in climate models. *Geoscientific Model Development*, *12*(8), 3805–3834. https://doi.org/10.5194/gmd-12-3805-2019

Lewis, J., Andela, B., Collier, N., Lee, J., Hegedus, D., Pflüger, M., & Xu, M. (2025). Rapid evaluation framework [Software]. *Zenodo*. https://doi.org/10.5281/zenodo.15103441

Li, D., Folini, D., & Wild, M. (2023). Assessment of top of atmosphere, atmospheric and surface energy budgets in CMIP6 models on regional scales. *Earth and Space Science*, *10*(4), e2022EA002758. https://doi.org/10.1029/2022EA002758

Li, J., Liu, Z., Yao, Z., & Wang, R. (2019). Comprehensive assessment of Coupled Model Intercomparison Project Phase 5 global climate models using observed temperature and precipitation over mainland Southeast Asia. *International Journal of Climatology*, *39*(10), 4139–4153. https://doi.org/10.1002/joc.6064

Li, J., Miao, C., Wei, W., Zhang, G., Hua, L., Chen, Y., & Wang, X. (2021). Evaluation of CMIP6 global climate models for simulating land surface energy and water fluxes during 1979–2014. *Journal of Advances in Modeling Earth Systems*, *13*(6), e2021MS002515. https://doi.org/10.1029/2021MS002515

Li, L.-L., Li, J., & Yu, R.-C. (2022). Evaluation of CMIP6 HighResMIP models in simulating precipitation over Central Asia. *Advances in Climate Change Research*, *13*(1), 1–13. https://doi.org/10.1016/j.accre.2021.09.009

Li, S., Huang, G., Li, X., Liu, J., & Fan, G. (2021). An assessment of the Antarctic sea ice mass budget simulation in CMIP6 historical experiment. *Frontiers in Earth Science*, *9*, 649743. https://doi.org/10.3389/feart.2021.649743

Li, Y., Tang, G., O'Rourke, E., Minallah, S., e Braga, M. M., Nowicki, S., et al. (2025). CMIP7 data request: Land and land ice priorities and opportunities. *EGUsphere*, *2025*, 1–48. https://doi.org/10.5194/egusphere-2025-3207

Lin, X., Massonnet, F., Fichefet, T., & Vancoppenolle, M. (2021). SITool (v1.0) – A new evaluation tool for large-scale sea ice simulations: Application to CMIP6 OMIP. *Geoscientific Model Development*, *14*(10), 6331–6354. https://doi.org/10.5194/gmd-14-6331-2021

Lu, J. (2024). Paradigm shifts of climate science for climate solutions. *Innovation*, *5*(4), 100628. https://doi.org/10.1016/j.xinn.2024.100628

Lukas, J., Gutmann, E., Harding, B., & Lehner, F. (2020). Chapter 11: Climate change-informed hydrology. In J. Lukas & E. Payton (Eds.), *Colorado River Basin Climate and Hydrology: State of the Science*, *Western Water Assessment* (pp. 384–449). University of Colorado.

Luo, Y. Q., Randerson, J. T., Abramowitz, G., Bacour, C., Blyth, E., Carvalhais, N., et al. (2012). A framework for benchmarking land models. *Biogeosciences*, *9*(10), 3857–3874. https://doi.org/10.5194/bg-9-3857-2012

Malinina, E., & Gillett, N. P. (2024). The 2021 heatwave was less rare in Western Canada than previously thought. *Weather and Climate Extremes*, *43*, 100642. https://doi.org/10.1016/j.wace.2024.100642

Manabe, S., & Bryan, K. (1969). Climate calculations with a combined ocean-atmosphere model. *Journal of the Atmospheric Sciences*, *26*(4), 786–789. https://doi.org/10.1175/1520-0469(1969)026⟨0786:CCWACO⟩2.0.CO;2

Martinez-Villalobos, C., Neelin, J. D., & Pendergrass, A. G. (2022). Metrics for evaluating CMIP6 representation of daily precipitation probability distributions. *Journal of Climate*, *35*(17), 5719–5743. https://doi.org/10.1175/JCLI-D-21-0617.1

Massonnet, F., Fichefet, T., Goosse, H., Bitz, C. M., Philippon-Berthier, G., Holland, M. M., & Barriat, P.-Y. (2012). Constraining projections of summer Arctic sea ice. *The Cryosphere*, *6*(6), 1383–1394. https://doi.org/10.5194/tc-6-1383-2012

Mauritsen, T., Stevens, B., Roeckner, E., Crueger, T., Esch, M., Giorgetta, M., et al. (2012). Tuning the climate of a global model. *Journal of Advances in Modeling Earth Systems*, *4*(3), 2012MS000154. https://doi.org/10.1029/2012MS000154

Mayer, M., Fasullo, J. T., Trenberth, K. E., & Haimberger, L. (2016). ENSO-driven energy budget perturbations in observations and CMIP models. *Climate Dynamics*, *47*(12), 4009–4029. https://doi.org/10.1007/s00382-016-3057-z

McAvaney, B. J., Covey, C., Joussaume, S., Kattsov, V., Kitoh, A., Ogana, W., et al. (2001). Model evaluation. In *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 471–523). Cambridge University Press.

McKitrick, R., McIntyre, S., & Herman, C. (2010). Panel and multivariate methods for tests of trend equivalence in climate data series. *Atmospheric Science Letters*, *11*(4), 270–277. https://doi.org/10.1002/asl.290

McPartland, M. Y., Lovato, T., Koven, C. D., Wilson, J. D., Turner, B., Petrik, C. M., et al. (2025). CMIP7 data request: Earth system priorities and opportunities. *EGUsphere*, *2025*, 1–61. https://doi.org/10.5194/egusphere-2025-3246

Meehl, G. A. (2023). The role of the IPCC in climate science. In *Oxford Research Encyclopedia of Climate Science*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190228620.013.933

Meehl, G. A., Boer, G. J., Covey, C., Latif, M., & Stouffer, R. J. (1997). Intercomparison makes for a better climate model. *Eos, Transactions American Geophysical Union*, *78*(41), 445–451. https://doi.org/10.1029/97EO00276

Meehl, G. A., Covey, C., Delworth, T., Latif, M., McAvaney, B., Mitchell, J. F. B., et al. (2023). The WCRP CMIP3 multimodel dataset: A new era in climate change research. *Bulletin of the American Meteorological Society*, *88*(9), 1383–1394. https://doi.org/10.1175/BAMS-88-9-1383

Meehl, G. A., Senior, C. A., Eyring, V., Flato, G., Lamarque, J.-F., Stouffer, R. J., et al. (2020). Context for interpreting equilibrium climate sensitivity and transient climate response from the CMIP6 Earth system models. *Science Advances*, *6*(26), eaba1981. https://doi.org/10.1126/sciadv.aba1981

Merchant, C. J., Paul, F., Popp, T., Ablain, M., Bontemps, S., Defourny, P., et al. (2017). Uncertainty information in climate data records from Earth observation. *Earth System Science Data*, *9*(2), 511–527. https://doi.org/10.5194/essd-9-511-2017

Milroy, D. J., Baker, A. H., Hammerling, D. M., & Jessup, E. R. (2018). Nine time steps: Ultra-fast statistical consistency testing of the Community Earth System Model (pyCECT v3.0). *Geoscientific Model Development*, *11*(2), 697–711. https://doi.org/10.5194/gmd-11-697-2018

Model Benchmarking Task Team. (2025). Model benchmarking and evaluation tools - Coupled Model Intercomparison Project. Retrieved from https://wcrp-cmip.org/tools/model-benchmarking-and-evaluation-tools/#tools_gallery

Morice, C. P., Kennedy, J. J., Rayner, N. A., Winn, J. P., Hogan, E., Killick, R. E., et al. (2021). An updated assessment of near-surface temperature change from 1850: The HadCRUT5 data set. *Journal of Geophysical Research: Atmospheres*, *126*(3), e2019JD032361. https://doi.org/10.1029/2019JD032361

Neelin, J. D., Krasting, J. P., Radhakrishnan, A., Liptak, J., Jackson, T., Ming, Y., et al. (2023). Process-oriented diagnostics: Principles, practice, community development, and common standards. *Bulletin of the American Meteorological Society*, *104*(8), E1452–E1468. https://doi.org/10.1175/BAMS-D-21-0268.1

Nicholls, Z. R., Meinshausen, M. A., Lewis, J., Rojas Corradi, M., Dorheim, K., Gasser, T., et al. (2021). Reduced Complexity Model Intercomparison Project Phase 2: Synthesising Earth system knowledge for probabilistic climate projections. *ESS Open Archive*. https://doi.org/10.1002/essoar.10504793.2

Nie, Y., Lin, X., Yang, Q., Liu, J., Chen, D., & Uotila, P. (2023). Differences Between the CMIP5 and CMIP6 Antarctic sea ice concentration budgets. *Geophysical Research Letters*, *50*(23), e2023GL105265. https://doi.org/10.1029/2023GL105265

Nigro, M. A., Cassano, J. J., & Seefeldt, M. W. (2011). A weather-pattern-based approach to evaluate the Antarctic Mesoscale Prediction System (AMPS) forecasts: Comparison to automatic weather station observations. *Weather and Forecasting*, *26*(2), 184–198. https://doi.org/10.1175/2010WAF2222444.1

Nijsse, F. J. M. M., Cox, P. M., & Williamson, M. S. (2020). Emergent constraints on transient climate response (TCR) and equilibrium climate sensitivity (ECS) from historical warming in CMIP5 and CMIP6 models. *Earth System Dynamics*, *11*(3), 737–750. https://doi.org/10.5194/esd-11-737-2020

Notz, D. (2015). How well must climate models agree with observations? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *373*(2052), 20140164. https://doi.org/10.1098/rsta.2014.0164

Notz, D., & Community, S. (2020). Arctic sea ice in CMIP6. *Geophysical Research Letters*, *47*(10), e2019GL086749. https://doi.org/10.1029/2019GL086749

Nowack, P., Runge, J., Eyring, V., & Haigh, J. D. (2020). Causal networks for climate model evaluation and constrained projections. *Nature Communications*, *11*(1), 1415. https://doi.org/10.1038/s41467-020-15195-y

O'Neill, B. C., Tebaldi, C., Van Vuuren, D. P., Eyring, V., Friedlingstein, P., Hurtt, G., et al. (2016). The Scenario Model Intercomparison Project (ScenarioMIP) for CMIP6. *Geoscientific Model Development*, *9*(9), 3461–3482. https://doi.org/10.5194/gmd-9-3461-2016

Orbe, C., Rind, D., Jonas, J., Nazarenko, L., Faluvegi, G., Murray, L. T., et al. (2020). GISS model E2.2: A climate model optimized for the middle atmosphere—2. Validation of large-scale transport and evaluation of climate response. *Journal of Geophysical Research: Atmospheres*, *125*(24), e2020JD033151. https://doi.org/10.1029/2020JD033151

Ordonez, A. C. (2023). CMEC driver [Computer Software]. *U.S. Department of Energy (DOE)*. https://doi.org/10.11578/dc.20230424.2

Paçal, A., Hassler, B., Weigel, K., Kurnaz, M. L., Wehner, M. F., & Eyring, V. (2023). Detecting extreme temperature events using Gaussian mixture models. *Journal of Geophysical Research: Atmospheres*, *128*(18), e2023JD038906. https://doi.org/10.1029/2023JD038906

Parding, K. M., Dobler, A., McSweeney, C. F., Landgren, O. A., Benestad, R., Erlandsen, H. B., et al. (2020). GCMeval – An interactive tool for evaluation and selection of climate model ensembles. *Climate Services*, *18*, 100167. https://doi.org/10.1016/j.cliser.2020.100167

Pathak, R., Dasari, H. P., Ashok, K., & Hoteit, I. (2023). Effects of multi-observations uncertainty and models similarity on climate change projections. *npj Climate and Atmospheric Science*, *6*(1), 144. https://doi.org/10.1038/s41612-023-00473-5

Pathak, R., Sahany, S., Mishra, S. K., & Dash, S. K. (2019). Precipitation biases in CMIP5 models over the South Asian Region. *Scientific Reports*, *9*(1), 9589. https://doi.org/10.1038/s41598-019-45907-4

PCMDI Metrics Package. (2025). PMP mean climate metrics — pcmdi.llnl.gov. Retrieved from https://pcmdi.llnl.gov/pmp-preliminary-results/interactive_plot/portrait_plot/mean_clim/cmip6/historical/v20240430/mean_clim_portrait_plot_4seasons_cmip6_historical_rms_xy_v20240430.html

Peña-Angulo, D., Vicente-Serrano, S. M., Domínguez-Castro, F., Murphy, C., Reig, F., Tramblay, Y., et al. (2020). Long-term precipitation in Southwestern Europe reveals no clear trend attributable to anthropogenic forcing. *Environmental Research Letters*, *15*(9), 094070. https://doi.org/10.1088/1748-9326/ab9c4f

Petrie, R., Denvil, S., Ames, S., Levavasseur, G., Fiore, S., Allen, C., et al. (2021). Coordinating an operational data distribution network for CMIP6 data. *Geoscientific Model Development*, *14*(1), 629–644. https://doi.org/10.5194/gmd-14-629-2021

Phillips, A. S., Deser, C., & Fasullo, J. (2014). Evaluating modes of variability in climate models. *Eos, Transactions American Geophysical Union*, *95*(49), 453–455. https://doi.org/10.1002/2014EO490002

Pierce, D. W., Barnett, T. P., Santer, B. D., & Gleckler, P. J. (2009). Selecting global climate models for regional climate change studies. *Proceedings of the National Academy of Sciences*, *106*(21), 8441–8446. https://doi.org/10.1073/pnas.0900094106

Pincus, R., Batstone, C. P., Hofmann, R. J. P., Taylor, K. E., & Glecker, P. J. (2008). Evaluating the present-day simulation of clouds, precipitation, and radiation in climate models. *Journal of Geophysical Research*, *113*(D14), 2007JD009334. https://doi.org/10.1029/2007JD009334

Planton, Y. Y., Guilyardi, E., Wittenberg, A. T., Lee, J., Gleckler, P. J., Bayr, T., et al. (2021). Evaluating climate models with the CLIVAR 2020 ENSO metrics package. *Bulletin of the American Meteorological Society*, *102*(2), E193–E217. https://doi.org/10.1175/BAMS-D-19-0337.1

Plummer, S., Lecomte, P., & Doherty, M. (2017). The ESA Climate Change Initiative (CCI): A European contribution to the generation of the Global Climate Observing System. *Remote Sensing of Environment*, *203*, 2–8. https://doi.org/10.1016/j.rse.2017.07.014

Po-Chedley, S., Santer, B. D., Fueglistaler, S., Zelinka, M. D., Cameron-Smith, P. J., Painter, J. F., & Fu, Q. (2021). Natural variability contributes to model–satellite differences in tropical tropospheric warming. *Proceedings of the National Academy of Sciences*, *118*(13), e2020962118. https://doi.org/10.1073/pnas.2020962118

Potter, G. L., Carriere, L., Hertz, J., Bosilovich, M., Duffy, D., Lee, T., & Williams, D. N. (2018). Enabling reanalysis research using the Collaborative Reanalysis Technical Environment (CREATE). *Bulletin of the American Meteorological Society*, *99*(4), 677–687. https://doi.org/10.1175/BAMS-D-17-0174.1

Randerson, J. T., Hoffman, F. M., Thornton, P. E., Mahowald, N. M., Lindsay, K., Lee, Y.-H., et al. (2009). Systematic assessment of terrestrial biogeochemistry in coupled climate-carbon models. *Global Change Biology*, *15*(10), 2462–2484. https://doi.org/10.1111/j.1365-2486.2009.01912.x

Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., Rowell, D. P., et al. (2003). Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *Journal of Geophysical Research*, *108*(D14), 2002JD002670. https://doi.org/10.1029/2002JD002670

Reichler, T., & Kim, J. (2008). How well do coupled models simulate today's climate? *Bulletin of the American Meteorological Society*, *89*(3), 303–312. https://doi.org/10.1175/BAMS-89-3-303

Righi, M., Andela, B., Eyring, V., Lauer, A., Predoi, V., Schlund, M., et al. (2020). Earth System Model Evaluation Tool (ESMValTool) v2.0 – Technical overview. *Geoscientific Model Development*, *13*(3), 1179–1199. https://doi.org/10.5194/gmd-13-1179-2020

Rivera, J. A., & Arnould, G. (2020). Evaluation of the ability of CMIP6 models to simulate precipitation over Southwestern South America: Climatic features and long-term trends (1901–2014). *Atmospheric Research*, *241*, 104953. https://doi.org/10.1016/j.atmosres.2020.104953

Roberts, M. J., Camp, J., Seddon, J., Vidale, P. L., Hodges, K., Vanniere, B., et al. (2020). Impact of model resolution on tropical cyclone simulation using the HighResMIP–PRIMAVERA multimodel ensemble. *Journal of Climate*, *33*(7), 2557–2583. https://doi.org/10.1175/JCLI-D-19-0639.1

Rosenblum, E., & Eisenman, I. (2016). Faster Arctic sea ice retreat in CMIP5 than in CMIP3 due to volcanoes. *Journal of Climate*, *29*(24), 9179–9188. https://doi.org/10.1175/JCLI-D-16-0391.1

Ruane, A. C., Pascoe, C. L., Teichmann, C., Brayshaw, D. J., Buontempo, C., Diouf, I., et al. (2025). CMIP7 data request: Impacts and adaptation priorities and opportunities. *EGUsphere*, *2025*, 1–61. https://doi.org/10.5194/egusphere-2025-3408

Sanderson, B. M., Booth, B. B. B., Dunne, J., Eyring, V., Fisher, R. A., Friedlingstein, P., et al. (2024). The need for carbon-emissions-driven climate projections in CMIP7. *Geoscientific Model Development*, *17*(22), 8141–8172. https://doi.org/10.5194/gmd-17-8141-2024

Sanderson, B. M., Knutti, R., & Caldwell, P. (2015). A representative democracy to reduce interdependency in a multimodel ensemble. *Journal of Climate*, *28*(13), 5171–5194. https://doi.org/10.1175/JCLI-D-14-00362.1

Schär, C., Fuhrer, O., Arteaga, A., Ban, N., Charpilloz, C., Di Girolamo, S., et al. (2020). Kilometer-scale climate models: Prospects and challenges. *Bulletin of the American Meteorological Society*, *101*(5), E567–E587. https://doi.org/10.1175/BAMS-D-18-0167.1

Schlund, M., Andela, B., Benke, J., Comer, R., Hassler, B., Hogan, E., et al. (2025). Advanced climate model evaluation with ESMValTool v2.11.0 using parallel, out-of-core, and distributed computing. *Geoscientific Model Development Discussions*, *2025*, 1–21. https://doi.org/10.5194/gmd-2024-236

Schlund, M., Eyring, V., Camps-Valls, G., Friedlingstein, P., Gentine, P., & Reichstein, M. (2020). Constraining uncertainty in projected gross primary production with machine learning. *Journal of Geophysical Research: Biogeosciences*, *125*(11), e2019JG005619. https://doi.org/10.1029/2019JG005619

Schlund, M., Hassler, B., Lauer, A., Andela, B., Jöckel, P., Kazeroni, R., et al. (2023). Evaluation of native Earth system model output with ESMValTool v2.6.0. *Geoscientific Model Development*, *16*(1), 315–333. https://doi.org/10.5194/gmd-16-315-2023

Schlund, M., Lauer, A., Gentine, P., Sherwood, S. C., & Eyring, V. (2020). Emergent constraints on equilibrium climate sensitivity in CMIP5: Do they hold for CMIP6? *Earth System Dynamics*, *11*(4), 1233–1258. https://doi.org/10.5194/esd-11-1233-2020

Schmidt, G. A., Bader, D., Donner, L. J., Elsaesser, G. S., Golaz, J.-C., Hannay, C., et al. (2017). Practice and philosophy of climate model tuning across six US modeling centers. *Geoscientific Model Development*, *10*(9), 3207–3223. https://doi.org/10.5194/gmd-10-3207-2017

Schneider, S. H., & Dickinson, R. E. (1974). Climate modeling. *Reviews of Geophysics*, *12*(3), 447–493. https://doi.org/10.1029/RG012i003p00447

Schneider, T., Leung, L. R., & Wills, R. C. J. (2024). Opinion: Optimizing climate models with process knowledge, resolution, and artificial intelligence. *Atmospheric Chemistry and Physics*, *24*(12), 7041–7062. https://doi.org/10.5194/acp-24-7041-2024

Schwalm, C. R., Huntinzger, D. N., Michalak, A. M., Fisher, J. B., Kimball, J. S., Mueller, B., et al. (2013). Sensitivity of inferred climate model skill to evaluation decisions: A case study using CMIP5 evapotranspiration. *Environmental Research Letters*, *8*(2), 024028. https://doi.org/10.1088/1748-9326/8/2/024028

Schwanitz, V. J. (2013). Evaluating integrated assessment models of global climate change. *Environmental Modelling & Software*, *50*, 120–131. https://doi.org/10.1016/j.envsoft.2013.09.005

Scoccimarro, E., & Gualdi, S. (2020). Heavy daily precipitation events in the CMIP6 worst-case scenario: Projected twenty-first-century changes. *Journal of Climate*, *33*(17), 7631–7642. https://doi.org/10.1175/JCLI-D-19-0940.1

Scoccimarro, E., Gualdi, S., Bellucci, A., Zampieri, M., & Navarra, A. (2013). Heavy precipitation events in a warmer climate: Results from CMIP5 models. *Journal of Climate*, *26*(20), 7902–7911. https://doi.org/10.1175/JCLI-D-12-00850.1

Seiler, C., Melton, J. R., Arora, V. K., & Wang, L. (2021). CLASSIC v1.0: The open-source community successor to the Canadian Land Surface Scheme (CLASS) and the Canadian Terrestrial Ecosystem Model (CTEM) – Part 2: Global benchmarking. *Geoscientific Model Development*, *14*(5), 2371–2417. https://doi.org/10.5194/gmd-14-2371-2021

Sharma, B., Kumar, J., Collier, N., Ganguly, A. R., & Hoffman, F. M. (2022). Quantifying carbon cycle extremes and attributing their causes under climate and land use and land cover change from 1850 to 2300. *Journal of Geophysical Research: Biogeosciences*, *127*(6), e2021JG006738. https://doi.org/10.1029/2021JG006738

Simpson, I. R., Bacmeister, J., Neale, R. B., Hannay, C., Gettelman, A., Garcia, R. R., et al. (2020). An evaluation of the large-scale atmospheric circulation and its variability in CESM2 and other CMIP Models. *Journal of Geophysical Research: Atmospheres*, *125*(13), e2020JD032835. https://doi.org/10.1029/2020JD032835

Simpson, I. R., McKinnon, K. A., Davenport, F. V., Tingley, M., Lehner, F., Fahad, A. A., & Chen, D. (2021). Emergent constraints on the large-scale atmospheric circulation and regional hydroclimate: Do they still work in CMIP6 and how much can they actually constrain the future? *Journal of Climate*, *34*(15), 6355–6377. https://doi.org/10.1175/JCLI-D-21-0055.1

Simpson, I. R., Shaw, T. A., Ceppi, P., Clement, A. C., Fischer, E., Grise, K. M., et al. (2025). Confronting Earth System Model trends with observations. *Science Advances*, *11*(11), eadt8035. https://doi.org/10.1126/sciadv.adt8035

Smith, K., Barthel, A. M., Conlon, L. M., Van Roekel, L. P., Bartoletti, A., Golez, J.-C., et al. (2024). The DOE E3SM Version 2.1: Overview and assessment of the impacts of parameterized ocean submesoscales. *Geoscientific Model Development*, 1–38. https://doi.org/10.5194/gmd-2024-149

Song, X., Wang, D.-Y., Li, F., & Zeng, X.-D. (2021). Evaluating the performance of CMIP6 Earth system models in simulating global vegetation structure and distribution. *Advances in Climate Change Research*, *12*(4), 584–595. https://doi.org/10.1016/j.accre.2021.06.008

Spafford, L., & MacDougall, A. H. (2020). Quantifying the probability distribution function of the transient climate response to cumulative CO_2 emissions. *Environmental Research Letters*, *15*(3), 034044. https://doi.org/10.1088/1748-9326/ab6d7b

Srivastava, A., Grotjahn, R., & Ullrich, P. A. (2020). Evaluation of historical CMIP6 model simulations of extreme precipitation over contiguous US regions. *Weather and Climate Extremes*, *29*, 100268. https://doi.org/10.1016/j.wace.2020.100268

Stephens, G. L., Vane, D. G., Boain, R. J., Mace, G. G., Sassen, K., Wang, Z., et al. (2002). The CloudSat mission and the A-Train: A new dimension of space-based observations of clouds and precipitation. *Bulletin of the American Meteorological Society*, *83*(12), 1771–1790. https://doi.org/10.1175/BAMS-83-12-1771

Stevens, B., Giorgetta, M., Esch, M., Mauritsen, T., Crueger, T., Rast, S., et al. (2013). Atmospheric component of the MPI-M Earth System Model: ECHAM6. *Journal of Advances in Modeling Earth Systems*, *5*(2), 146–172. https://doi.org/10.1002/jame.20015

Stevenson, D. S., Zhao, A., Naik, V., O'Connor, F. M., Tilmes, S., Zeng, G., et al. (2020). Trends in global tropospheric hydroxyl radical and methane lifetime since 1850 from AerChemMIP. *Atmospheric Chemistry and Physics*, *20*(21), 12905–12920. https://doi.org/10.5194/acp-20-12905-2020

Storto, A., Alvera-Azcárate, A., Balmaseda, M. A., Barth, A., Chevallier, M., Counillon, F., et al. (2019). Ocean reanalyses: Recent advances and unsolved challenges. *Frontiers in Marine Science*, *6*. https://doi.org/10.3389/fmars.2019.00418

Stouffer, R. J., & Manabe, S. (2017). Assessing temperature pattern projections made in 1989. *Nature Climate Change*, *7*(3), 163–165. https://doi.org/10.1038/nclimate3224

Su, H., Jiang, J. H., Zhai, C., Shen, T. J., Neelin, J. D., Stephens, G. L., & Yung, Y. L. (2014). Weakening and strengthening structures in the Hadley Circulation change under global warming and implications for cloud response and climate sensitivity. *Journal of Geophysical Research: Atmospheres*, *119*(10), 5787–5805. https://doi.org/10.1002/2014JD021642

Sung, H. M., Kim, J., Shim, S., Seo, J.-B., Kwon, S.-H., Sun, M.-A., et al. (2021). Climate change projection in the twenty-first century simulated by NIMS-KMA CMIP6 model based on new GHGs concentration pathways. *Asia-Pacific Journal of Atmospheric Sciences*, *57*(4), 851–862. https://doi.org/10.1007/s13143-021-00225-6

Swaminathan, R., Dingley, B., & Team, M. B. T. (2025). Different community open-source evaluation and benchmarking tools with their respective characteristics [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.17629992

Swaminathan, R., Quaife, T., & Allan, R. (2024). A machine learning framework to evaluate vegetation modeling in Earth system models. *Journal of Advances in Modeling Earth Systems*, *16*(7), e2023MS004097. https://doi.org/10.1029/2023MS004097

Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research*, *106*(D7), 7183–7192. https://doi.org/10.1029/2000JD900719

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, *93*(4), 485–498. https://doi.org/10.1175/BAMS-D-11-00094.1

Teixeira, J., Waliser, D., Ferraro, R., Gleckler, P., Lee, T., & Potter, G. (2014). Satellite observations for CMIP5: The genesis of Obs4MIPs. *Bulletin of the American Meteorological Society*, *95*(9), 1329–1334. https://doi.org/10.1175/BAMS-D-12-00204.1

Tian, B., & Dong, X. (2020). The double-ITCZ bias in CMIP3, CMIP5, and CMIP6 models based on annual mean precipitation. *Geophysical Research Letters*, *47*(8), e2020GL087232. https://doi.org/10.1029/2020GL087232

Tokarska, K. B., Arora, V. K., Gillett, N. P., Lehner, F., Rogelj, J., Schleussner, C.-F., et al. (2020). Uncertainty in carbon budget estimates due to internal climate variability. *Environmental Research Letters*, *15*(10), 104064. https://doi.org/10.1088/1748-9326/abaf1b

Tselioudis, G., Rossow, W. B., Jakob, C., Remillard, J., Tropf, D., & Zhang, Y. (2021). Evaluation of clouds, radiation, and precipitation in CMIP6 models using global weather states derived from ISCCP-H cloud property data. *Journal of Climate*, *34*(17), 7311–7324. https://doi.org/10.1175/JCLI-D-21-0076.1

Tsujino, H., Urakawa, L. S., Griffies, S. M., Danabasoglu, G., Adcroft, A. J., Amaral, A. E., et al. (2020). Evaluation of global ocean–sea-ice model simulations based on the experimental protocols of the Ocean Model Intercomparison Project phase 2 (OMIP-2). *Geoscientific Model Development*, *13*(8), 3643–3708. https://doi.org/10.5194/gmd-13-3643-2020

Ullrich, P., Barnes, E., Collins, W., Dagon, K., Jablonowski, C., Lee, J., et al. (2025). Recommendations for comprehensive and independent evaluation of machine learning based Earth-System Models. (Accepted for publication to *Journal of Geophysical Research: Machine Learning and Computation* Oct 2024). *Journal of Geophysical Research: Machine Learning and Computation*. https://doi.org/10.48550/arXiv.2410.19882

Vicente-Serrano, S. M., García-Herrera, R., Peña-Angulo, D., Tomas-Burguera, M., Domínguez-Castro, F., Noguera, I., et al. (2022). Do CMIP models capture long-term observed annual precipitation trends? *Climate Dynamics*, *58*(9–10), 2825–2842. https://doi.org/10.1007/s00382-021-06034-x

Vissio, G., Lembo, V., Lucarini, V., & Ghil, M. (2020). Evaluating the performance of climate models based on Wasserstein distance. *Geophysical Research Letters*, *47*(21), e2020GL089385. https://doi.org/10.1029/2020GL089385

Von Clarmann, T., Degenstein, D. A., Livesey, N. J., Bender, S., Braverman, A., Butz, A., et al. (2020). Overview: Estimating and reporting uncertainties in remotely sensed atmospheric composition and temperature. *Atmospheric Measurement Techniques*, *13*(8), 4393–4436. https://doi.org/10.5194/amt-13-4393-2020

Wagener, T., Reinecke, R., & Pianosi, F. (2022). On the evaluation of climate change impact models. *WIREs Climate Change*, *13*(3), e772. https://doi.org/10.1002/wcc.772

Waliser, D., Gleckler, P. J., Ferraro, R., Taylor, K. E., Ames, S., Biard, J., et al. (2020). Observations for Model Intercomparison Project (Obs4MIPs): Status for CMIP6. *Geoscientific Model Development*, *13*(7), 2945–2958. https://doi.org/10.5194/gmd-13-2945-2020

Wan, H., Zhang, K., Rasch, P. J., Singh, B., Chen, X., & Edwards, J. (2017). A new and inexpensive non-bit-for-bit solution reproducibility test based on time step convergence (TSC1.0). *Geoscientific Model Development*, *10*(2), 537–552. https://doi.org/10.5194/gmd-10-537-2017

Wargan, K., Weir, B., Manney, G. L., Cohn, S. E., Knowland, K. E., Wales, P. A., & Livesey, N. J. (2023). M2-SCREAM: A stratospheric composition reanalysis of Aura MLS data with MERRA-2 transport. *Earth and Space Science*, *10*(2), e2022EA002632. https://doi.org/10.1029/2022EA002632

Watson-Parris, D., Rao, Y., Olivié, D., Seland, Î., Nowack, P., Camps-Valls, G., et al. (2022). ClimateBench v1.0: A benchmark for data-driven climate projections. *Journal of Advances in Modeling Earth Systems*, *14*(10), e2021MS002954. https://doi.org/10.1029/2021MS002954

Waugh, D. W., & Eyring, V. (2008). Quantitative performance metrics for stratospheric-resolving chemistry-climate models. *Atmospheric Chemistry and Physics*, *8*(18), 5699–5713. https://doi.org/10.5194/acp-8-5699-2008

WCRP CMIP. (2025a). Climate Model Benchmarking - Coupled Model Intercomparison Project. Retrieved from https://wcrp-cmip.org/cmip7-task-teams/model-benchmarking/

WCRP CMIP. (2025b). Model benchmarking and evaluation tools - Coupled Model Intercomparison Project. Retrieved from https://wcrp-cmip.org/tools/model-benchmarking-and-evaluation-tools/

Webb, M. J., Andrews, T., Bodas-Salcedo, A., Bony, S., Bretherton, C. S., Chadwick, R., et al. (2017). The Cloud Feedback Model Intercomparison Project (CFMIP) contribution to CMIP6. *Geoscientific Model Development*, *10*(1), 359–384. https://doi.org/10.5194/gmd-10-359-2017

Weigel, K., Bock, L., Gier, B. K., Lauer, A., Righi, M., Schlund, M., et al. (2021). Earth System Model Evaluation Tool (ESMValTool) v2.0 – Diagnostics for extreme events, regional and impact evaluation, and analysis of Earth system models in CMIP. *Geoscientific Model Development*, *14*(6), 3159–3184. https://doi.org/10.5194/gmd-14-3159-2021

Wild, M., Folini, D., Hakuba, M. Z., Schär, C., Seneviratne, S. I., Kato, S., et al. (2015). The energy balance over land and oceans: An assessment based on direct observations and CMIP5 climate models. *Climate Dynamics*, *44*(11–12), 3393–3429. https://doi.org/10.1007/s00382-014-2430-z

Williams, D. (2015). *2014 Earth System Grid Federation and Ultrascale Visualization Climate Data Analysis Tools Conference Report*. (Technical Report Nos. LLNL-TR–666753, 1182238). LLNL. https://doi.org/10.2172/1182238

Williams, R. G., Ceppi, P., & Katavouta, A. (2020). Controls of the transient climate response to emissions by physical feedbacks, heat uptake and carbon cycling. *Environmental Research Letters*, *15*(9), 0940c1. https://doi.org/10.1088/1748-9326/ab97c9

Wu, J., Shi, Y., & Xu, Y. (2020). Evaluation and projection of surface wind speed over China based on CMIP6 GCMs. *Journal of Geophysical Research: Atmospheres*, *125*(22), e2020JD033611. https://doi.org/10.1029/2020JD033611

Xie, P., Arkin, P. A., & Janowiak, J. E. (2007). CMAP: The CPC merged analysis of precipitation. In *Measuring precipitation from space: EURAINSAT and the future* (pp. 319–328). Springer.

Xin, X., Wu, T., Zhang, J., Yao, J., & Fang, Y. (2020). Comparison of CMIP6 and CMIP5 simulations of precipitation in China and the East Asian summer monsoon. *International Journal of Climatology*, *40*(15), 6423–6440. https://doi.org/10.1002/joc.6590

Yang, X., Wood, E. F., Sheffield, J., Ren, L., Zhang, M., & Wang, Y. (2018). Bias correction of historical and future simulations of precipitation and temperature for China from CMIP5 models. *Journal of Hydrometeorology*, *19*(3), 609–623. https://doi.org/10.1175/JHM-D-17-0180.1

Yihui, D., & Chan, J. C. (2005). The East Asian summer monsoon: An overview. *Meteorology and Atmospheric Physics*, *89*(1), 117–142. https://doi.org/10.1007/s00703-005-0125-z

Yu, T., Chen, W., Gong, H., Feng, J., & Chen, S. (2023). Comparisons between CMIP5 and CMIP6 models in simulations of the climatology and interannual variability of the East Asian summer monsoon. *Climate Dynamics*, *60*(7), 2183–2198. https://doi.org/10.1007/s00382-022-06408-9

Zechlau, S., Schlund, M., Cox, P. M., Friedlingstein, P., & Eyring, V. (2022). Do emergent constraints on carbon cycle feedbacks hold in CMIP6? *Journal of Geophysical Research: Biogeosciences*, *127*(12), e2022JG006985. https://doi.org/10.1029/2022JG006985

Zhang, C., Xie, S., Tao, C., Tang, S., Emmenegger, T., Neelin, J., et al. (2021). Evaluating climate models: The ARM data-oriented metrics and diagnostics toolkit. *Bulletin of the American Meteorological Society*, *102*(4), 347–350. https://doi.org/10.1175/BAMS-D-19-0282.A

Zhang, Q., Liu, B., Li, S., & Zhou, T. (2023). Understanding models' global sea surface temperature bias in mean state: From CMIP5 to CMIP6. *Geophysical Research Letters*, *50*(4), e2022GL100888. https://doi.org/10.1029/2022GL100888

Zhao, M., Golaz, J., Held, I. M., Guo, H., Balaji, V., Benson, R., et al. (2018). The GFDL global atmosphere and land model AM4.0/LM4.0: 1. Simulation characteristics with prescribed SSTs. *Journal of Advances in Modeling Earth Systems*, *10*(3), 691–734. https://doi.org/10.1002/2017MS001208

Zumwald, M., Knüsel, B., Baumberger, C., Hirsch Hadorn, G., Bresch, D. N., & Knutti, R. (2020). Understanding and assessing uncertainty of observational climate datasets for model evaluation using ensembles. *WIREs Climate Change*, *11*(5), e654. https://doi.org/10.1002/wcc.654