Rapid Evaluation Framework for the CMIP7 Assessment Fast Track

Forrest M. Hoffman[1], Birgit Hassler[2], Ranjini Swaminathan[3], Jared Lewis[4], Bouwe Andela[5], Nathaniel Collier[1], Dóra Hegedűs[6,7], Jiwoo Lee[8], Charlotte Pascoe[6], Mika Pflüger[4], Martina Stockhause[9], Paul Ullrich[8], Min Xu[1], Lisa Bock[2], Felicity Chun[10], Bettina K. Gier[11,2], Douglas I. Kelley[12], Axel Lauer[2], Julien Lenhardt[13], Manuel Schlund[2], Mohanan G. Sreeush[14], Katja Weigel[11,2], Ed Blockley[15], Rebecca Beadling[16], Romain Beucher[10], Demiso D. Dugassa[17], Valerio Lembo[18], Jianhua Lu[19], Swen Brands[20], Jerry Tjiputra[21], Elizaveta Malinina[22], Brian Medeiros[23], Enrico Scoccimarro[24], Jeremy Walton[15], Phil Kershaw[6], André Lanfer Marquez[25], Malcolm J. Roberts[15], Eleanor O'Rourke[7], Beth Dingley[7], Briony Turner[7], Helene Hewitt[15], and John P. Dunne[26]

[1]Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831-6301, United States of America
[2]Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany
[3]Department of Meteorology and National Centre for Earth Observation, University of Reading, Reading RG6 6UR, United Kingdom
[4]Climate Resource, Fitzroy, Victoria, Australia
[5]Netherlands eScience Center, Amsterdam, The Netherlands
[6]RAL Space, Science and Technology Facilities Council, Rutherford Appleton Laboratory, Harwell Campus, Didcot OX11 OFD, United Kingdom
[7]CMIP International Project Office, European Space Agency, Harwell Campus, Didcot OX11 OFD, United Kingdom
[8]Lawrence Livermore National Laboratory, Livermore, California 94550, United States of America
[9]Instituto de Física de Cantabria (IFCA), CSIC-Universidad de Cantabria, Spain
[10]Australia's Climate Simulator (ACCESS-NRI), Australian National University, Canberra, Australia
[11]University of Bremen, Institute of Environmental Physics (IUP), Bremen, Germany
[12]United Kingdom Centre for Ecology & Hydrology, Wallingford, Oxfordshire OX10 8BB, United Kingdom
[13]Swedish Meteorological and Hydrological Institute, Norrköping, Sweden
[14]Alfred-Wegener-Institut, Helmholtz Zentrum für Polar- und Meeresforschung, Bremerhaven, Germany
[15]Met Office, FitzRoy Road, Exeter EX1 3PB, United Kingdom
[16]Department of Earth and Environmental Science, Temple University, Philadelphia, Pennsylvania 19122, United States of America
[17]Water Technology Institute, Faculty of Hydraulic and Water Resources Engineering, Arba Minch University, Arba Minch, Ethiopia
[18]Institute of Atmospheric Sciences and Climate, National Research Council of Italy, I-00133 Rome, Italy
[19]School of Atmospheric Sciences, Sun Yat-sen University & Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Zhuhai, China
[20]Instituto de Física de Cantabria (CSIC-UC), Santander, Spain
[21]NORCE Norwegian Research Centre, Bjerknes Centre for Climate Research, Bergen, Norway
[22]Environment and Climate Change Canada, Victoria, British Columbia, Canada
[23]NSF National Center for Atmospheric Research, Boulder, Colorado 80307, United States of America
[24]CMCC Foundation - Euro-Mediterranean Center on Climate Change, 73100 Lecce LE, Italy
[25]National Institute for Space Research (INPE), São José dos Campos, Brazil
[26]NOAA Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey 08540, United States of America

Correspondence: Forrest M. Hoffman (hoffmanfm@ornl.gov), Birgit Hassler (Birgit.Hassler@dlr.de), and Ranjini Swaminathan (r.swaminathan@reading.ac.uk)

Abstract. As Earth system models (ESMs) grow in complexity and in volume of output data, there is an increasing need for rapid, comprehensive evaluation of their scientific performance. The upcoming Assessment Fast Track for the Seventh Phase of the Coupled Model Intercomparison Project (CMIP7) will require expeditious response for model analyses designed to inform and drive integrated Earth system assessments. To meet this challenge, the Rapid Evaluation Framework (REF), a community-driven platform for benchmarking and performance assessment of ESMs, was designed and developed. The initial implementation of the REF, constructed to meet the near-term needs of the CMIP7 Assessment Fast Track, builds upon four disparate community evaluation and benchmarking tools that are coupled together using the Coordinated Model Evaluation Capabilities (CMEC) framework. The REF runs within a containerized workflow for portability and reproducibility and is aimed at generating and organizing diagnostics covering a variety of model variables. The REF leverages well documented observational datasets to provide assessments of model fidelity across a collection of diagnostics. All diagnostics were identified and selected with community involvement and consultation. Operational integration with the Earth System Grid Federation (ESGF) will permit automated execution of the REF for selected diagnostics as soon as model output data are published on ESGF by the originating modeling centers. The REF is designed to be portable across a range of current computational platforms to facilitate use by modeling centers for assessing the evolution of model versions or gauging the relative performance of CMIP simulations before being published on ESGF. When integrated into production simulation workflows, results from the REF provide immediate quantitative feedback that allows model developers and scientists to quickly identify model biases and performance issues. After the REF is released to the community, its subsequent development and support will be prioritized by an international consortium of scientists and engineers, enabling a broader impact across Earth science disciplines. For instance, the REF will facilitate improvements to models and will enhance confidence in model projections through process-based selection of models based on their performance with respect to observations. Production of reproducible diagnostics and community-based assessments are key features of the REF. Furthermore, providing interoperability with existing evaluation packages assures that contributions from previous community efforts will be available for use in future model intercomparison projects.

1 Introduction

Earth system models (ESMs) are the primary tools for the scientific community to study interactions at a wide range of scales, from sub-daily to millennial, between the atmosphere, land, ocean, cryosphere, and biosphere, and how the Earth system responds to human-induced and natural forcings (e.g., Mauritsen et al., 2019; Séférian et al., 2019; Yukimoto et al., 2019;

Boucher et al., 2020; Danabasoglu et al., 2020; Senior et al., 2020; Döscher et al., 2022). Currently, a new generation of ESMs is being finalized and new simulations will be generated for the Seventh Phase of the Coupled Model Intercomparison Project (CMIP7) and the precursory CMIP7 Assessment Fast Track. As the number of models, ensemble sizes, complexity and output requirements continue to grow, there is an urgent need to objectively evaluate the fidelity of these models and exploit the wealth of information they provide in order to efficiently advance our understanding of the Earth system and to inform climate mitigation and adaptation policies. This includes, specifically, identification of model uncertainties or systematic biases that may prevent us from objectively constraining model-derived projections of future climate change. Addressing this need requires developing efficient evaluation methods capable of making use of the growing archive of model output and reducing the time required to translate the output into meaningful scientific insight. The rapid growth of ESM data, driven by model complexity and computational advances, creates both opportunities and challenges. Effective, reproducible, accurate, and unbiased data processing is crucial for translating model outputs into actionable insights for climate policy (IPCC, 2023).

The CMIP Model Benchmarking Task Team (MB-TT) was created to address this challenge in preparation for CMIP7 and the contribution deadline for the Seventh Assessment Report of the Intergovernmental Panel on Climate Change (IPCC-AR7). Following the first phase of the MB-TT, which reviewed ESM evaluation and benchmarking approaches and identified the collection and collation of existing community benchmarking software packages (Hassler et al., 2026a), and outlined best practices for the use of observational datasets for model evaluation (Beadling et al., 2026); the MB-TT has now expanded its efforts toward developing a community-designed Rapid Evaluation Framework (REF) for routine and rapid benchmarking of CMIP simulations. The conceptual design of the REF was developed at an MB-TT workshop in May 2024 and approved by the CMIP Panel in July 2024, with development work commencing in October 2024. The initial design was strongly motivated by the ideas and vision developed for CMIP6 (Eyring et al., 2019); the goal of the REF for CMIP7 is to deliver a complete end-to-end system that will provide a high-level, systematic and rapid performance assessment of CMIP models, initially targeting the model experiments contributing to the CMIP7 Assessment Fast Track, which will support the IPCC-AR7 (Dunne et al., 2025). The vision of the REF is to be a community-owned evaluation framework, leveraging existing community-built model evaluation packages and incorporating an application programming interface (API) that will execute modules for generation of diagnostics and the metrics that underlie them.

Rather than directly ranking models, the REF is primarily concerned with providing objective measures of model performance to allow the wider community to make informed decisions about models that are most appropriate for their specific needs. This requires a standard set of diagnostics and performance metrics to facilitate the comparison of key variables simulated by models with standardized observational reference datasets, and assessment of whether fundamental processes in the Earth system are adequately represented in the models. Once expanded for community use beyond the initial Assessment Fast Track version, the REF will have a wider array of applications and users, including other modeling communities and scientific domains, as well as organizations utilizing CMIP models for conducting feedback analysis, impact assessments, or financial planning. To facilitate understanding of the descriptions of the REF, key terms used throughout the manuscript are defined here. This terminology may not be used consistently across all Earth science disciplines.

– **Reference Dataset** – A reference dataset is a collection of observationally constrained or model data used as a standard within a model evaluation diagnostic. Examples may include *in situ* measurements, extrapolated data (from statistical or AI/ML methods), remote sensing data, reanalysis products, or any other dataset that is meant to represent a best estimate of a geophysical quantity or a physical, chemical, biological, or ecological state or process.

– **Model Variable** – A model variable is any quantitative representation or characterization of a physical, chemical, biological, or ecological state or process that changes during execution of the model. Variables are used to represent mass, energy, velocity, momentum, flux rates, and other parameters within models. Model variables may or may not represent observable quantities, and they may be inferred, estimated, or calculated from other related variables or observables.

– **Diagnostic** – A diagnostic is a comparison of a model variable or some combination of model variables with a reference dataset or an intercomparison across models of a model variable or some combination of model variables (Hassler et al., 2026a). A diagnostic may also represent an evaluation of a relationship between multiple model variables and/or multiple reference datasets (i.e., Relationship Diagnostics). Diagnostics have sometimes been called "confrontations" since the objective is to confront models with best-available observations or with best-available model or model ensemble outputs. A diagnostic consists of one or more model performance metrics.

– **Diagnostic Collection** – A diagnostic collection is a grouping of one or more diagnostics for a given model variable, phenomena, or theme being evaluated.

– **Metric** – A metric is a single statistical evaluation contained within a diagnostic. A diagnostic may consist of more than one metric. Examples include bias, root mean squared error (RMSE), spatial or temporal correlations (Taylor, 2001), Earth Mover's Distance, Hellinger Distance (Hellinger, 1909), phase/timing of the seasonal cycle, amplitude of the seasonal cycle, inter-annual variability (Giorgi and Francisco, 2000). Not all metrics are useful for all variables or should be used with every observationally constrained dataset. Each metric may be evaluated to produce a metric scalar.

– **Metric Scalar** – A metric scalar is the numerical output resulting from the calculation of a performance metric (e.g., the calculated bias).

– **Score** – A score is a scalar value (0.0–1.0) transformed from a metric scalar or produced by aggregating multiple metric scalars or multiple scores together.

– **Verification** – Verification is the process of assessing model consistency in terms of correct implementation of the represented processes as expected from the model and simulation experimental design. Sometimes, this is accomplished 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 Baker et al. (2015) and the Time Step Consistency Test in Wan et al. (2017)).

– **Validation** – Validation is the process of determining the degree to which a model accurately represents processes in the real world, particularly for the intended uses of the model. Validation can include a broad range of aspects from ensuring correct units and the sign of the data produced, to the interactions between model components or variables and process representations.

– **Fidelity** – The fidelity of a model is a quantitative assessment of the degree to which model output corresponds to the reference data in aggregate, resulting from a validation exercise. One approach for deriving a fidelity metric is to aggregate relevant scores.

2 Conceptual Design of the REF

2.1 Overview of the REF

The REF was designed to be a community-owned evaluation framework that leverages existing open source, community-built model evaluation and benchmarking packages that are integrated together through a standard application programming interface (API) that will execute modules for generation of diagnostics and the metrics that underlie them. By incorporating existing tools and metrics with publicly available reference datasets, the REF standardizes a community workflow for CMIP activities, reduces duplication of efforts for evaluating models, and, through deployment on the Earth System Grid Federation (ESGF), provides rapid feedback to the research community about relative model performance across a wide range of model components and variables. Moreover, the REF is expected to offer a starting point for the research community to expand model evaluation and benchmarking capabilities and applications within their own institutions or modeling centers.

The high-level workflow for the REF as it runs on ESGF nodes, as shown in Fig. 1, is triggered by the publication of new simulations requiring evaluation. After the model output has undergone a quality control check to assure the metadata are correct, data storage and compute resources are allocated for a new execution of the diagnostics generation process. As agreed upon by the CMIP Panel, model output that does not conform to the required Controlled Vocabulary (CV) and metadata standards will not be evaluated by the REF. Next, an optimized directed acyclic graph of tasks is produced, and processes are initiated to calculate model evaluation metrics and construct diagnostics. The outputs of diagnostics are then staged on public websites for sharing with the research community. The initial implementation of the REF was created to evaluate Assessment Fast Track simulations (Dunne et al., 2025), using five to six diagnostics across five Earth system realms with the expectation that additional diagnostics would be added in the future.

Modeling centers usually evaluate their models during development with the overall performance documented and published once the models are finalized and key simulations are completed. However, this approach has three main limitations. First, the process is slow, and evaluation results often come too late to inform assessment reports or even later for stakeholders who need timely information. Second, these evaluation results are hard to access; they are scattered across different types of papers (technical or peer-reviewed, open access or not, etc.) across the modeling centers and often only partially included in IPCC reports, making them difficult to locate and piece together. Third, no consistent approach is applied across centers; each group

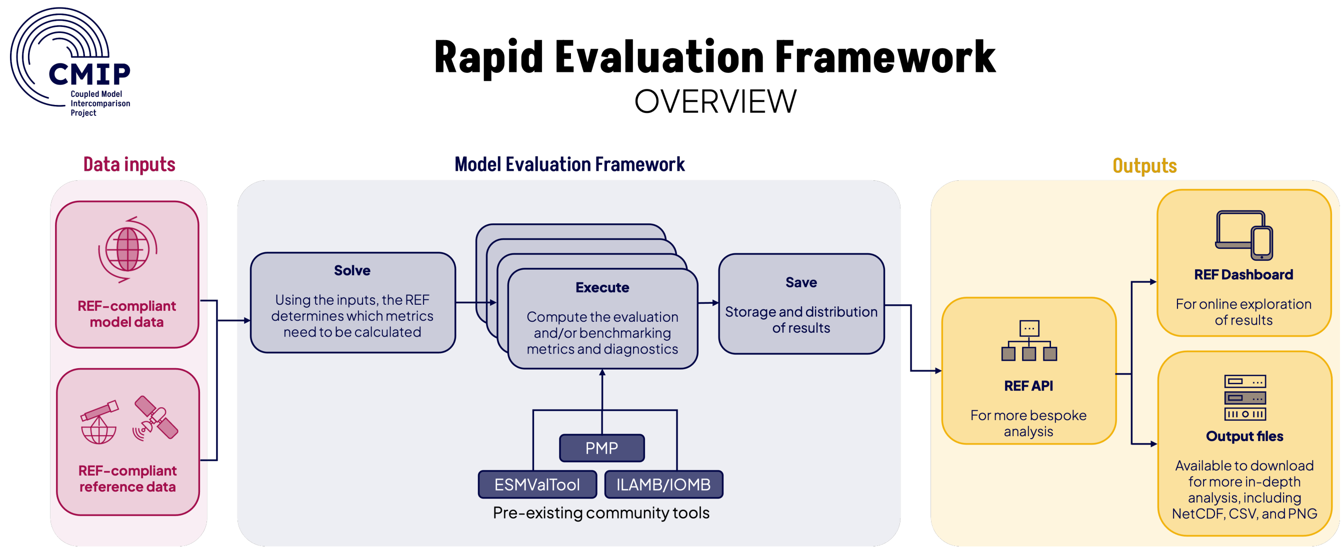


Figure 1. The high-level workflow for the Rapid Evaluation Framework (REF) run on Earth System Grid Federation (ESGF) nodes, shown here, combines quality-assured model output with a collection of observational reference data to initiate execution of relevant diagnostics and generation of tabular and graphical representations of a variety of metrics. The results are then published via an API for more bespoke analysis and on an online dashboard for community use. Obs is an abbreviation for observational, QA/QC is an acronym for Quality Assurance / Quality Control, and DAG is an acronym for Directed Acyclic Graph. Figure available at https://doi.org/10.5281/zenodo.20053946 (Hoffman et al., 2026).

runs its own evaluations using sometimes different methods, which makes comparisons between models difficult. The REF aims to address these issues by bringing together core evaluation frameworks and their diagnostic output in one place. This workflow is envisioned to help experts quickly assess, understand, and synthesize the performance of a new generation of ESMs for Assessment Fast Track simulations. The REF also makes it easier for stakeholders to access the information they need to support regional analysis for adaptation and mitigation, and it supports best practices for evaluation during model development. Running the REF before model output is submitted to ESGF offers modeling centers an additional option to systematically assess their models and to make targeted improvements during the development phase.

The REF aims to integrate well-documented reference datasets for comparison with model output. These reference datasets will be collected in obs4MIPs (Waliser et al., 2020), a project created to distribute data that support evaluation of ESMs via ESGF (nodes listed at https://esgf.github.io/nodes.html). Obs4MIPs compiles a range of observationally constrained data, formatted according to CMIP conventions to ensure compatibility with model output. The published obs4MIPs datasets include variables such as temperature, precipitation, and sea level pressure, spanning various spatial and temporal ranges and scales across Earth system realms (Ocean & Sea Ice, Land & Land Ice, Atmosphere, Earth System, and Impacts & Adaptation), providing a comprehensive basis for model evaluation (Waliser et al., 2020). The formatting of obs4MIPs data adheres to the Climate and Forecast (CF) conventions (Eaton et al., 2024) and the CMIP metadata standards (Eyring et al., 2016). These conventions ensure that data are consistently structured, with uniform variable naming, units, and metadata, facilitating easy comparison across model outputs.

While obs4MIPs serves an important role in distributing observational datasets, the REF is flexible and designed to use additional reference data available in the required format. This adaptability allows for the incorporation of new reference datasets as they become available, or when specific needs arise, ensuring that the framework can accommodate a broad range of observational data. The REF leverages the attributes provided by datasets to automate the diagnostic generation process. This automation requires standardization and coherent attributes across models, variables and experiments in order to correctly combine different datasets without manual selection by humans. Reference data used by the REF may span different time periods, a key point that analysts need to consider when interpreting the results of the available diagnostics. Moreover, community activities like CMIP should strive to ensure that reference data span the full contemporary time period of simulation experiments intended to reproduce recent historical trends and variability. The REF does not aim to reformat reference or model data, and REF users must ensure the data they wish to use follow the CF conventions and meet the CMIP metadata standards to be considered REF-compliant data (Figure 1). Quality checks (Hassler et al., 2026b) are required to assure the data used for evaluation and benchmarking diagnostics are consistent and reliable, enabling meaningful model-data comparisons (Hawkins and Sutton, 2016). These quality checks do, in some cases, go beyond the checks required for data publication to ESGF.

2.2 Community Engagement

Key stakeholders for the co-development of the REF were identified as modeling centers involved in the Assessment Fast Track, reference dataset providers, evaluation tool and package developers, diagnostic developers and climate scientists, including IPCC authors, seeking to analyze the Assessment Fast Track output. The MB-TT initiated work on the REF based on the results of the CMIP6 Community Survey (O'Rourke, 2025), conducted by the World Climate Research Programme (WCRP) from January to March 2022. Detailed analysis of the survey results by the Task Team (Lee, 2024) revealed calls for open and transparent model benchmarking and evaluation and the ability to run CMIP evaluation tools alongside ESGF, as part of the data publication process, with provision of evaluation results display.

Members of Fresh Eyes on CMIP worked with the CMIP-International Program Office and the MB-TT to develop a survey that was circulated to modeling center science and technical leads, as well as more widely in the Fresh Eyes on CMIP and CMIP community mailing lists and via social media. The survey was open May 7–17, 2024, and generated 152 unique responses, and the responses were analyzed by members of the Fresh Eyes on CMIP and reported at a MB-TT workshop. These survey results informed design and development of the initial REF proposal. Suggestions included considering multiple software tools, strategies to ensure accessibility, tools for quality control, and transparency and provenance for reference datasets and model output. Additional suggestions included acceptance of data not processed by the Computer Model Output Rewriter (CMOR; Taylor et al., 2004), potentially with AI-assisted reformatting, improved accessibility through use of ESGF and cloud-based computing to enhance data inclusivity and flexibility, as well as errata tracking. A detailed analysis of the survey results is available in the survey report by (Wang et al., 2025).

A separate community survey was subsequently disseminated to modeling centers in June 2024 that included requests for input on evaluation tools used, interest in use of a benchmarking framework within their own computing environments, and willingness to submit preliminary model output for quality assurance and control checks (O'Rourke, 2024). This survey informed refinement of the framework structure and implementation plan. In September 2024, a survey was disseminated to modelers, modeling centers, and reference dataset and data infrastructure providers, requesting input regarding the specific diagnostics that should be prioritized for inclusion in the REF, receiving 53 responses. A co-development session with members of the Earth observation community was held at the European Space Agency (ESA) Climate Change Initiative (CCI) and Climate Modelling User Group (CMUG) integration meeting in October 2024, which resulted in the inclusion of ozone-related metrics in existing diagnostics where appropriate and two additional cloud diagnostics. The outcomes of the survey, in combination with suggestions from the MB-TT, resulted in a set of diagnostics, shown in Section 4 and published separately (CMIP Model Benchmarking Task Team, 2024). The REF was thereby launched on November 4, 2024, at a community-engagement event that saw wide participation. Successive months were dedicated to implementation of the diagnostics from four existing evaluation and benchmarking packages within the REF architecture. The packages selected for providing diagnostics for the Assessment Fast Track REF were ESMValTool, PMP, ILAMB, and IOMB, which are described below in Section 4. During this period, engagement efforts focused on technical implementation, involving members of the community responsible for provision of supporting infrastructure and quality control, including the WGCM Infrastructure Panel (WIP), obs4MIPs Steering Panel, the ESGF-WIP Quality Assurance and Control Working Group, the CMIP Data Request and CVs Task Teams, as well as the IPCC Task Group on Data Support for Climate Change Assessments (TG-Data), on their citation and provenance requirements for the REF, see Section 4.6.

A preliminary prototype of the working suite was stress-tested for operational usage in time for the REF Hackathon, which was held at the Met Office, in Exeter, UK, from March 10–13, 2025, and included dedicated drop-in sessions for modelers and reference dataset providers. This prototype was able to ingest a sample of CMIP6 model output and obs4MIPs data, running a small subset of ILAMB, PMP, and ESMValTool diagnostics. In May 2025, modeling center science and technical personnel, as well as previous attendees of the REF launch, were invited to attend the launch of the beta version of the REF. Approximately 107 participants registered for the demo and to attend follow-up feedback sessions in June 2025 regarding the beta release. This beta testing period was instrumental for building an easy-to-use tutorial and documentation, solving potential issues with REF usage on HPC machines, and receiving feedback from users, with the aim of delivering a public release of the REF for actual usage with Assessment Fast Track outputs in October 2025.

3 Target Applications

Considering the REF scope, stakeholder consultation and initial implementation plan for the Assessment Fast Track simulations, three primary potential applications for the REF were identified: provide information about scientific performance of ESMs to stakeholders and policymakers, focus on key sensitivity indicators, and enhance equality in model data access. The primary application of the REF is to provide stakeholders from the data analysis and impacts assessment communities with timely information about the scientific performance of ESMs with respect to observationally constrained (reference) datasets across all Earth science realms (Atmosphere, Ocean and sea ice, Land and land ice, Earth system, Impacts and adaptation). Reference data are available primarily but not exclusively from the contemporary observational period, approximately over the last 50 years. Secondarily, the REF produces diagnostics focused on key sensitivity indicators that typically do not have corresponding observational constraints, such as equilibrium climate sensitivity (ECS) and transient climate response (TCR). Access to all of this diagnostic information assists stakeholders in selecting simulation results for further analysis, downscaling studies, impacts analyses, or other research. Moreover, the results of the REF can increase equality in climate data access for community members who lack adequate computer resources or Internet access. In some cases, diagnostics from the REF include graphs, charts, or figures that can be used in research publications or assessment reports, offering analysts more time to focus on specific research questions that may use information from the REF.

A key design goal of the REF is that it can be used by modeling centers, research institutions, and individual scientists to enable validation of ESM output prior to publication of simulations on ESGF, intercomparison of model results, and general purpose analysis. Running the REF prior to data submission provides the opportunity for data providers to gauge the performance and sensitivity of their simulations in a standard fashion at any time. Modeling centers typically use their own collections of diagnostics, often developed in-house, for routine model assessment, while other centers or institutions have adopted community-developed evaluation tools or employ a combination of community tools and in-house diagnostics. The REF offers a general purpose framework for model evaluation and is equally useful for tracking the scientific performance of different versions of the same model. The REF diagnostics also provide a convenient means for determining if model changes during development yield improvements. Thus, modeling centers may find that running the REF as a part of their workflow for repeated simulations provides a practical way to track evolving performance of their model. It may also be necessary to reformat the relevant model output variables for each simulation, making them CF-compliant and ensuring CMIP naming standards and metadata are provided, within the workflow for the REF to be able to run correctly. Moreover, analysts from different research communities may want to add functionality to the initial REF implementation to enable use of other observational datasets, additional diagnostics, and other metrics as discussed in Section 5.5.1.

Furthermore, the REF could serve as an early warning system for modeling centers to identify inconsistencies in the model variables in a more complex way than was done for the prior CMIP activities (e.g., Taylor et al., 2004) and thus reduce ESGF data traffic and storage of erroneous data, limiting data use by the wider community. The REF can also be used for a variety of model intercomparison activities outside of the scope of CMIP, and project leaders can ask simulation contributors to run the REF and provide standard diagnostic results instead of sharing large volumes of model output. Additionally, the REF enables individual researchers interested in Earth system science to explore model output and apply well documented reference data to better understand model capabilities and gaps in process representation. The REF provides a standard framework for integrating additional diagnostics for use by research institutions or local analysts and scientists. New diagnostics integrated into the REF can be easily shared amongst modeling centers and researchers, and they become key candidate additions to future public releases of the community version of the REF.

4 System Description of the CMIP7 Assessment Fast Track REF

4.1 Included Evaluation and Benchmarking Packages and the Coupling Strategy

The open source evaluation and benchmarking packages described below—ESMValTool, PMP, and ILAMB & IOMB—were chosen for inclusion in the first version of the REF because they are open source, were relatively mature packages at the time, and combined, they permit comprehensive model evaluation across all the atmosphere, ocean and sea ice, land and land ice, Earth system, and impacts and adaptation realms. Subsets of diagnostics from each package were selected based on input from the MB-TT and the community. For the initial version of the REF, diagnostics were restricted to analyze only monthly mean output from the expected new simulations. In this section, the packages are briefly described, although a more in-depth description of them can be found in Hassler et al. (2026a), and the Coordinated Model Evaluation Capabilities (CMEC) standards are also described, since they offer a strategy for integrating these disparate evaluation packages.

4.1.1 ESMValTool

The Earth System Model Evaluation Tool (ESMValTool) is an open source community-developed software package aimed at computing many different diagnostics and metrics for the evaluation and benchmarking of ESMs (Righi et al., 2020; Eyring et al., 2020; Lauer et al., 2020; Weigel et al., 2021; Schlund et al., 2023; Lauer et al., 2025; Schlund et al., 2025; Andela et al., 2025). Many of the diagnostics and metrics that have been officially released in ESMValTool have been systematically used for the production of multi-model intercomparisons embedded in several chapters of Working Group I (WGI) of the IPCC Sixth Assessment Report (AR6) (IPCC, 2023). ESMValTool strongly advocates traceability and reproducibility; therefore, all diagnostic results are provided with metadata documenting the provenance of the model output and reference data, the software packages used, and the calculated metrics and diagnostics. The software package also deals with the adjustment of model or observational datasets that are not strictly compliant with the CF conventions. Although the core capabilities of ESMValTool are fully Python-based, diagnostics can be based on other open source languages such as NCL (NCAR Command Language) or R. For all contributions, ESMValTool implements rigorous technical and scientific reviews before new code can be included in the official release, requiring Python Enhancement Proposal (PEP) 8 standards, and testing with pre-commit and Codacy for maturity of the code and standardization.

4.1.2 PMP

The PCMDI Metrics Package (PMP) is an open source Python software package developed for objective and rapid assessments and benchmarking of ESMs, against the most up-to-date observational datasets (Lee et al., 2025). The PMP has played an important role in the systematic evaluation of many simulations across CMIP generations, with a strong emphasis on physical climate metrics, particularly atmospheric means and variability. Among its diverse suite of metrics, a subset of metrics that are calculated based on the monthly time series of model variables were chosen for the first implementation in the REF (Table 1). This subset of metrics includes the annual cycle of different atmospheric variables (Gleckler et al., 2008), El Niño

Southern Oscillation (ENSO), CLIVAR (Climate and Ocean: Variability, Predictability and Change) metrics (Planton et al., 2021), extra-tropical modes of variability (Lee et al., 2021), and global monsoon metrics (Wang et al., 2011). By offering a database of pre-computed statistics for CMIP6 models, the PMP streamlines the comparison process, making it easier for modeling centers to evaluate their results against established benchmarks.

4.1.3 ILAMB and IOMB

The International Land Model Benchmarking (ILAMB) and International Ocean Model Benchmarking (IOMB) are open source Python software packages that share a large portion of the same codebase. They were developed to provide systematic assessment of land and ocean model performance, primarily for terrestrial and marine biogeochemistry, through comparison with reference datasets (Collier et al., 2018; Luo and Hoffman, 2022; Fu et al., 2022). Diagnostics and metrics within ILAMB and IOMB were developed with engagement of land and ocean modelers and with the *in situ* and remote sensing observational communities (Luo et al., 2012; Hoffman et al., 2017). Both packages were used to evaluate and intercompare historical simulations from CMIP5 and CMIP6 (IPCC, 2023, Chapter 5, Figure 5.7), as well as serving important roles in informing the development of land and ocean components for DOE's Energy Exascale Earth System Model (E3SM; Burrows et al., 2020; Zhu et al., 2019; Yang et al., 2019) and the Community Earth System Model (Lawrence et al., 2019). ILAMB and IOMB both offer a variety of statistical metrics, including bias, RMSE, timing/phase of the seasonal cycle, spatial correlation, and interannual variability. Scores from these metrics are aggregated to provide high-level scores for each model-dataset pairing. Functional relationship metrics within ILAMB and IOMB evaluate the degree to which model variable-to-variable relationships correspond to those of observational data. For example, the relationship between gross primary production (GPP) and mean annual temperature in the model should correspond well with the same relationship extracted from observational data. ILAMB and IOMB produce hierarchical webpages designed to offer users and analysts the ability to view many metrics at once for a given model-dataset pair, as well as to intercompare graphical representations of metrics across all models at once.

4.1.4 CMEC

Coordinated Model Evaluation Capabilities (CMEC) is an effort to bring together a diverse set of analysis packages that have been developed to facilitate the systematic evaluation of ESMs (Ordonez et al., 2025). CMEC provides the strategy for coupling multiple community benchmarking packages in the REF. CMEC includes capabilities supported by multiple agencies, and capabilities that have been contributed by community-based experts and international agencies. With widespread and rapid growth in the number of available diagnostic and model evaluation tools, a lack of standards within the evaluation community have meant that running even a single evaluation tool can require extensive user intervention. Given the significant commonality in how these evaluation tools operate, interoperability is a natural goal achievable through robust and light-weight standards. The three goals of the CMEC project include: (1) to develop robust and light-weight standards for the operation of evaluation packages and their output; (2) to develop accompanying tools for installation of evaluation packages, coordinated execution of evaluation packages, and obtaining data products necessary for operation of these tools; and (3) to build connections across groups, research centers, and individual investigators performing model evaluation. The CMEC standards were adopted for integrating the output of diagnostics produced by the model benchmarking packages described above.

4.2 Diagnostics

At the heart of the REF are the diagnostics that were selected in an iterative process (see Section 2.2) that can be calculated with each new simulation that is presented to the REF. The diagnostics are grouped according to the five different realms that were identified for the CMIP data request and each diagnostic included in this first version of the REF is calculated by only one software package. Table 1 provides an overview of all the diagnostics that were selected for the first version of the REF, based on the diagnostics table published on Zenodo (CMIP Model Benchmarking Task Team, 2024), their realm, the software package with which the diagnostic is calculated, and the reference datasets used in the comparison. A more detailed table of the diagnostics is presented in Appendix B. For some of the diagnostics, different methodologies adopting different software packages across the diagnostic providers were discovered (e.g., double Inter-tropical Convergence Zone (ITCZ) biases); in this case, the principle of minimal computational resources and least number of required variables was adopted in order to choose the appropriate tool to be responsible for the diagnostic in the REF.

During the REF development phase, it became clear that two of the identified diagnostics were not available from the software packages implemented in the Assessment Fast Track REF. These diagnostics, with their unique IDs, 5.1 (High amplitude Rossby waves) and 5.2 (Internal variability or ensemble spread for individual models), were then removed from the list of those to be integrated in the initial REF version, following consultation and agreement from the Impacts & Adaptation CMIP Data Request Author team leads and co-chairs of the Vulnerability, Impacts, Adaptation and Climate Services (VIACS) Advisory Board. One additional diagnostic was removed from the diagnostic list, ID 1.5 (Ocean heat content (OHC)), since also for this diagnostic no code for its calculation was available from the software packages. The original diagnostics ID 3.4 (Evaporation minus precipitation $(E-P)$) was moved to a different realm (now ID 2.8 (Precipitation minus evaporation $(P-E)$)), and diagnostic ID 3.5 (Double inter-tropical convergence zone (ITCZ)) was added to another existing diagnostics (ID 1.3 (El Niño Southern Oscillation (ENSO) diagnostics (lifecycle, seasonality, amplitude, teleconnections))) to better reflect the actual calculated metrics.

A more detailed description of each diagnostic and the rationale and methodology for its implementation is presented in Appendix C.

4.3 Data Request Opportunity

For CMIP7, the data request to modeling centers is based on different "opportunities" that the community could submit to the CMIP IPO, following a community-wide call, each focusing on a specific scientific topic. Each opportunity contains a concise description of its scientific goals and a list of all model variables that are requested from simulation output to achieve these goals. The different data request opportunities are grouped according to five realms (Ocean & Sea Ice, Land & Land Ice, Atmosphere, Earth System, and Impacts & Adaptation), which are the same as those that were used as a basis for selecting REF diagnostics (see Section 4.2).

Table 1. Based on community recommendations, an initial set of diagnostics was selected, published at https://zenodo.org/records/14284375, for incorporation into the initial version of the REF for evaluating relevant CMIP7 Assessment Fast Track simulations. Diagnostics 1.5, 5.1 and 5.2 are listed in strikethrough style because they were not implemented for the initial version of the REF. Diagnostic 3.4 was replaced by 2.8 for the Land & Land Ice realm, and 3.5 is available as part of 1.3 and not separately.

ID	Diagnostic	Package(s)	Reference Dataset(s)
Ocean & Sea Ice Realm			
1.1	Antarctic annual mean, Arctic September rate of sea ice area (SIA) loss per degree warming (dSIA / dGMST)	ESMValTool	OSI SAF/CCI, HadCRUT5
1.2	Atlantic meridional overturning circulation (AMOC)	IOMB	RAPID-2023 v1a
1.3	El Niño Southern Oscillation (ENSO) diagnostics (lifecycle, seasonality, amplitude, teleconnections)	PMP, ESMValTool	TropFlux v1.0, GPCP v2.3, GPCP v3.3, HadISST v1.1
1.4	Sea surface temperature (SST) bias, Sea surface salinity (SSS) bias	IOMB	WOA23 (climatology)
~~1.5~~	~~Ocean heat content (OHC)~~		
1.6	Antarctic & Arctic sea ice area seasonal cycle	ESMValTool	OSI-450-nh, OSI-450-sh
Land & Land Ice Realm			
2.1	Soil carbon	ILAMB	HWSD v2.0
2.2	Gross primary production (GPP)	ILAMB	WECANN v1.0, Fluxnet2015, GPCP v2.3, CRU TS v4.02
2.3	Runoff	ILAMB	LORA v1.0
2.4	Surface soil moisture	ILAMB	WangMao
2.5	Net ecosystem carbon balance	ILAMB	Hoffman v1.0
2.6	Leaf area index (LAI)	ILAMB	NOAA-NCEI-LAI-AVHRR v5.0
2.7	Snow cover	ILAMB	CCI-CryoClim-FSC v1.0
2.8	Precipitation minus evaporation $(P - E)$	ILAMB	GLEAM v3.3a, GPCP v2.3
Atmosphere Realm			
3.1	Annual cycle and seasonal mean of multiple variables	PMP, ESMValTool	ERA5, SAGE-CCI-OMPS v8.0, C3S-GTO-ECV v9.0, GPCP v3.3, CERES-EBAF v4.2
3.2	Radiative and heat fluxes at the surface and top of atmosphere (TOA)	PMP	CERES-EBAF v4.2
3.3	Climate variability modes (e.g., ENSO, Madden-Julian Oscillation (MJO), Extratropical modes of variability, monsoon)	PMP	NOAA-20CR v2.0, HadISST v1.1
~~3.4~~	~~Evaporation minus precipitation $(E - P)$~~		
~~3.5~~	~~Double inter-tropical convergence zone (ITCZ)~~		
3.6	Cloud radiative effects	ESMValTool	CERES-EBAF v4.2.1
3.7	Scatterplots of two cloud-relevant variables (for specific regions of the globe and specific cloud regimes)	ESMValTool	CLARA-AVHRR-A3, GPCP v3.3, CERES-EBAF v4.2.1, CALIPSO-ICECLOUD v2.0, ERA5
Earth System Realm			
4.1	Equilibrium climate sensitivity (ECS)	ESMValTool	N/A
4.2	Transient climate response (TCR)	ESMValTool	N/A
4.3	Transient climate response to cumulative emissions of carbon dioxide (TCRE)	ESMValTool	N/A
4.4	Zero emissions commitment (ZEC)	ESMValTool	N/A
4.5	Historical changes in climate variables (time series, trends)	ESMValTool	HadCRUT5, GPCP v3.3, ERA5
Impacts & Adaptation Realm			
~~5.1~~	~~High amplitude Rossby waves~~		
~~5.2~~	~~Internal variability or ensemble spread for individual models (precipitation and surface temperature)~~		
5.3	Evaluation of key climate variables at global warming levels	ESMValTool	N/A
5.4	Climate drivers for fire (fire burnt area, fire weather and fuel continuity)	ILAMB, ESMValTool	GFED v5.0, Weights

An opportunity regarding the REF was submitted to the Data Request Task Team's open call to ensure that the model variables required by the REF for generating the REF diagnostics were clearly identified and that modeling centers wishing to use the REF with their simulations would have a checklist of model variables needed for producing the REF diagnostics. While some REF diagnostics use only a single model variable from the historical simulation, others require two or more model variables from multiple different simulations, as described in the definitions in Section 1. Since the REF diagnostics span all five data request realms, a variable group request was submitted for each realm. The variable groups combined form the "REF CMIP7 data request" opportunity. The opportunity was published as part of the v1.2.1 release of the data request, consisting of five variable groups, containing 80 variables in total. Many of the variables are from the recently developed list of baseline climate variables for Earth system modeling (Juckes et al., 2025), which is a subset of variables reflecting the most frequently used elements of CMIP6. To facilitate finding information about the REF opportunity in the different data request documentation papers, it was decided that each paper would contain a very brief description of the opportunity and refer to the documentation paper about the atmosphere realm, where a more detailed REF opportunity description was added (Dingley et al., 2026).

4.4 Observations

Observations play a key role in the REF and several diagnostics require observed quantities for different climate variables across domains for model evaluation. For each diagnostic, we have identified at least one reference dataset for inclusion in the REF (see complete list in Appendix A). To ensure compliance with the Findability, Accessibility, Interoperability, and Reusability (FAIR) data standards (Wilkinson et al., 2016), all observational datasets included in the REF are required to have a fully open access license (e.g., CC-BY-4.0, CC0, OGL) and follow the CF conventions (Davis et al., 2024) to ensure technical alignment with CMIP standard output. This is enabled by making the datasets available for downloading on ESGF servers (Cinquini et al., 2014) as part of the obs4MIPs project (Gleckler et al., 2011; Teixeira et al., 2025; Ferraro et al., 2015; Waliser et al., 2020). For any datasets not meeting the open access criteria, the REF Delivery Team obtained a relaxation of the license constraint by formally requesting and receiving agreement from individual dataset providers. Observational datasets for the REF were then processed in compliance with the obs4MIPs Data Specifications 2.5 (ODS2.5) (Gleckler et al., 2024) with approval from the obs4MIPs Steering Panel (OSP). In facilitating observational data ingestion for the REF, two additional criteria not fully addressed by the ODS2.5 specifications were identified, for which the REF delivery team has drafted guidelines. These criteria relate to the treatment of uncertainty information in the data and for the formulation of complex citations (see discussion in Sections 4.5 and 4.6). In an effort to recognize the need for observational record lengths to be as closely aligned with CMIP simulations as possible, we specify that observational datasets should be available at least until the end of 2016, ensuring that the data is less than five years out of date with historical simulations from CMIP7, that run until the end of 2021. We consider the availability of data towards the end of the historical simulation period to be important for the calculation of certain statistical properties such as trends used in evaluation. Where the observational datasets currently do not meet this requirement, we have highlighted such entries in Table A1, so that the data sets may be extended or replaced as appropriate in the future.

Two pathways were identified for observational dataset providers to submit their datasets for future inclusion in the REF. The first step in both cases is to submit a dataset proposal to obs4MIPs for approval by the OSP, which can either be done directly by the dataset provider or via a third party with permission from the dataset provider. Once the dataset has been approved by the OSP, the registered content, including the dataset name, version, data provider details and release date should be submitted to the Program for Climate Model Diagnosis and Intercomparison (PCMDI) obs4MIPs CMOR Tables repository (https://github.com/PCMDI/obs4MIPs-cmor-tables). This process provides the dataset with a unique `source_id`, following CMIP conventions. Datasets can then be prepared for compliance in one of two ways:

- using the CMOR software as advised by the OSP to prepare their dataset, following instructions and examples on the PCMDI obs4MIPs CMOR GitHub repository (the "CMOR pathway").
- using software packages such as ESMValTool to generate CMOR-like datasets that have additional scripts ensuring CMOR and obs4MIPs compliance (the "CMOR-like pathway").

The CMOR pathway intrinsically provides a form of validation for datasets before publication through the use of the CMOR software. As the CMOR-like pathway may not provide the same level of validation, the REF Delivery Team has developed a validation script that the prepared CMOR-like datasets must pass before publication. The prepared datasets are then published to ESGF via one of the two Assessment Fast Track REF nodes, at the Center for Environmental Data Analysis (CEDA, United Kingdom) or at Oak Ridge National Laboratory (ORNL, United States of America).

For certain diagnostics (1.1 and 5.4 in Table 1), the reference datasets listed in Table 1 are pre-processed to produce one or more static files with information needed for the diagnostic. In such instances, these files are stored internally within the corresponding diagnostic package, and the REF includes acknowledgments and references for the input datasets used to produce the files. For such exceptions, the input datasets themselves are not required to be published on ESGF.

4.5 REF Requirements in Addition to obs4MIPs Compliance – Treatment of Uncertainties

Currently, most diagnostics within the REF do not incorporate uncertainty information from reference datasets. Where available for a reference dataset used in the diagnostics, the standard error of the mean or upper and lower bounds around the mean are used. Future plans for the REF include using multiple reference datasets to account for observational uncertainty as seen commonly in the literature. The unavailability of comprehensive and accurately characterized uncertainty information with the observational data was identified as a key barrier to incorporating this information by diagnostic developers. Feedback from observational data providers indicated that any associated uncertainty and instructions on how to correctly apply such information is best provided by the data providers themselves. This prompted the REF Delivery Team to develop guidance for including uncertainty information in observational datasets; which was necessary to allow ingestion of uncertainty information by the REF. The initial proposal, originating from community engagement at the REF Hackathon with observational dataset providers and metrics package developers, was refined following consultation with the OSP and the CMIP-CVs Task Team. For the CMOR pathway, the following requirements are outlined:

– All additional uncertainty information (see Table 2 for currently accepted information on uncertainty) is provided in a separate file. Initially, between one and three additional files will be used, depending on the extent of uncertainty information provided.

– The netCDF file containing the main geophysical variable has the global attribute `has_aux_unc`, and is set to `TRUE` when additional files containing uncertainty information are provided. It should be set to `FALSE` when no uncertainty information is provided. If this attribute does not exist, the REF will assume there is no uncertainty information provided to ensure back-compatibility with datasets already published through obs4MIPs.

– If `has_aux_unc` is set to `TRUE`, the netCDF file containing the main geophysical variable must have the global attribute `aux_uncertainty_id`. This contains all additional `variable_id`s for the uncertainty information provided, in the form of a string with spaces as delimiters.

– The CMIP global attribute, `variable_id` (as defined in (Taylor et al., 2026)), in the file containing the uncertainty information corresponds to variable name + accepted suffix (no separator between body and suffix). For accepted suffixes, see Table 2.

– Variable name within the netCDF file corresponds to the `variable_id`.

– A technical note including detailed explanation of additional uncertainty information, is provided for each dataset following guidance in (Hegedűs et al., 2025).

This distinction between how uncertainty information may be included in CMOR and CMOR-like datasets is due to the current capabilities of the CMOR software. Currently, CMOR does not accommodate the addition of ancillary variables in its output, and the required software update for inclusion is not feasible within the Assessment Fast Track REF timeline. For the CMOR-like pathway, there is no requirement to add the uncertainty information in separate files. Instead, additional uncertainty fields may be provided following CF conventions, by adding them as ancillary variables in a netCDF-CF file. The additional global attributes `has_aux_unc` and `aux_uncertainty_id` are still required, and the `variable_id` should be constructed as described above only using the suffixes from Table 2. This guidance was developed as a basic framework to enable the community to utilize a wider range of uncertainty information during the development of REF diagnostics.

4.6 REF Support of IPCC-AR7 Related to the Enhanced Traceability of its Results – Complex Citation

The REF Delivery Team consulted with the IPCC Task Group on Data Support for Climate Change Assessments (TG-Data) on their citation and provenance requirements for the REF. The IPCC TG-Data aims to enhance the traceability of key results presented in the AR7 (AR7; Stockhause et al., 2024) cycle using the new Complex Citation standard (Agarwal et al., 2025) for documentation and standardized provenance records for gathering the required information. This simple but flexible Complex Citation approach allows for the traceability of data product generation and the citation of multiple datasets or data subsets in a single referenced object called Complex Citation Object (CCO). TG-Data aims to include CCO references in every figure

Table 2. Suffixes for the variable_id were proposed to describe the uncertainty information included with the reference datasets.

Suffix	Long Name	Description
nobs	Number of observations	The number of discrete observations or measurements from which a data value has been derived.
ustd	Standard uncertainty	The per-datum standard uncertainty is a combination of independent, structured and common effects and is equal to the positive square root of the sum of the components. The standard uncertainty is provided at one standard deviation. Replacement for “standard error” due to updated terminology.
stderr	Standard error	Legacy term used for backward compatibility. Definition is identical to that of the standard uncertainty.
uind	Uncertainty due to independent effects	The per-datum component of uncertainty that is associated with uncorrelated effects, provided at one standard deviation.
ustr	Uncertainty due to structured effects	The per-datum component of uncertainty that is structured and correlated over a defined space/time scale, provided at one standard deviation. This correlation length scale in space and time must be provided in the variable metadata.
ucom	Uncertainty due to common effects	The per-datum component of uncertainty that is common to all observations of a given type (specific instrument, resolution, variable type) at one standard deviation.
lbnd	Lower bound of asymmetrical uncertainty	Alternative to stderr or ustd for observations with asymmetrical uncertainty distribution, lower bound of the uncertainty interval.
ubnd	Upper bound of asymmetrical uncertainty information	Alternative to stderr or ustd for observations with asymmetrical uncertainty distribution, upper bound of the uncertainty interval.

caption of the AR7. Prerequisite to this is the provision of detailed information on input data usage in the form of persistent identifiers (PID) for each file and each citable entity.

The REF has identified the need for the reference datasets to be published on ESGF with their Handle IDs and data collections with their DOIs. ODS2.5 already requires each file to have a global attribute called `tracking_id` used as PIDs on ESGF, a unique identifier with prefixes specified for each ESGF project. In addition, reference datasets used by the REF also require the inclusion of the DOI as a global attribute to ensure compatibility with CCO. The REF leads the task of defining a provenance template for authors of the IPCC AR7 as guidance on providing the CCO-related information.

Through the CMIP7 data citation, the CCO captures provenance information at the granularity of a model's contribution to an experiment. The list of the file handles of the specific data that were used from within the CMIP7 citation resource provides traceability.

4.7 Technical Implementation

The REF workflow consists of four steps:

- **Ingestion:** The user registers the source datasets that can be used (reference data and model output). The associated metadata is extracted and added to a local database along with the respective file paths. Only ingested files are used in any execution calculations.

- **Solve:** For each diagnostic, the possible executions that would be required are determined using the data requirements of a metric and the catalog of datasets that have been ingested. A hash of the datasets required for each execution is stored in the database and subsequently used to determine if a new execution is required or has already been performed.

- **Execute:** The executions that require running are then executed. The REF supports four key methods for executing diagnostics: Local-Serial, Local-Parallel, Celery and via an HPC Scheduler such as Slurm or PBS. The ESGF deployment uses the Celery-based executor, which runs diagnostics out-of-process and in parallel. Once the execution is complete, the CMEC-based outputs are parsed and any scalar value or diagnostic figures are added to the database for later display and use.

- **Visualize:** The results are then made available via a REST API (Representational State Transfer Application Programming Interface) and Typescript-based frontend. Python-based tooling may be developed in future releases to interact with the results, but is not currently implemented. The Frontend allows users to see the diagnostics that have been executed and the corresponding results. This includes the ability to track which datasets were used for a given execution, and the ability to download generated figures, datasets and log output. In some instances, the REF may also provide additional summary statistics across all executions for a diagnostic such as box and whisker plots.

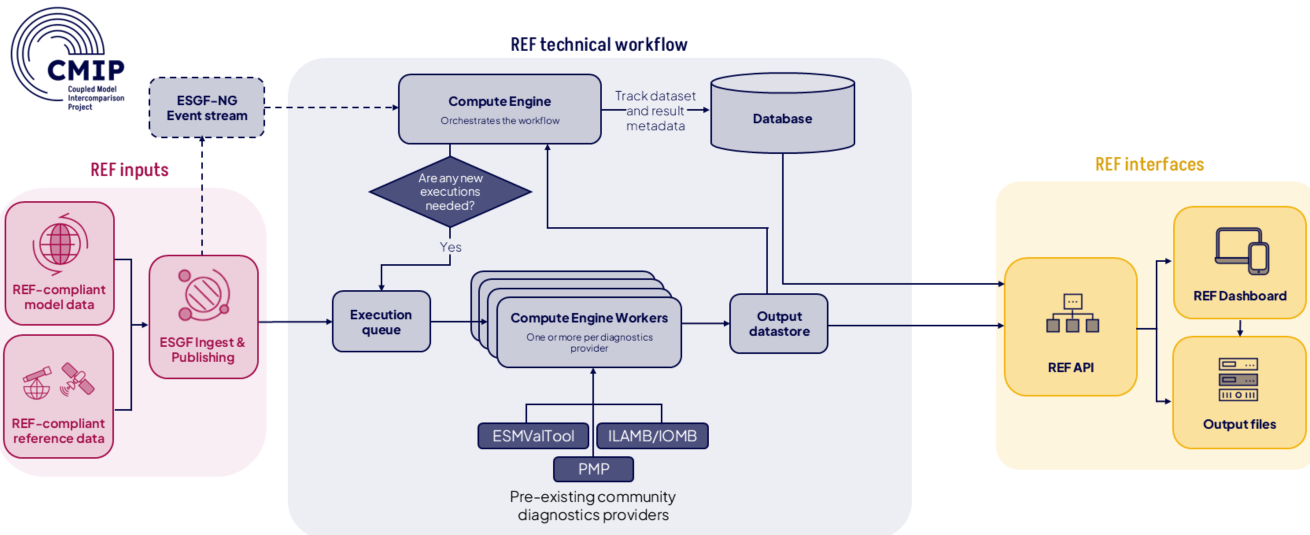


Figure 2. This detailed workflow diagram identifies key services provided by the REF and their relationships with external services and off-the-shelf components for the ESGF deployment. Figure available at https://doi.org/10.5281/zenodo.20054631 (Lewis et al., 2025).

4.8 ESGF Deployment

Figure 2 shows how the key services for the REF are integrated within the ESGF deployment. At the core of this system is the Compute Engine, which orchestrates the workflow by managing data ingestion and processing tasks. The Compute Engine manages a database (PostgreSQL or SQLite) of all the datasets, executions and results that have been performed. The Compute Engine will ingest new data when they become available either via consuming newly replicated events from the ESGF-Next Generation Event Stream, a Kafka-based service, or by periodically checking for newly available datasets. A "solve" is then be performed to determine if any new executions are required. For the ESGF deployment, each of the services (blue boxes in Fig. 2) are deployed as a Kubernetes service.

Data processing is handled by the Compute Engine Worker, which pulls executions from a queue and then executes the appropriate diagnostic. Each diagnostic provider is deployed as a separate service that can be scaled independently according to demand. This might be particularly important if one of the execution services requires more computational resources than the others. These workers require access to shared read-only mount containing input data (CMIP and observation datasets) and a mount for the Scratch Datastore where results are stored.

The Compute Engine watches for the success/failure of the executions on the workers. A subset of results are copied to the Output Datastore by the Compute Engine. During this process, metadata about the results are extracted into the database, including metric values, figures and data files. The use of a central database of results allows for efficient querying in the API across executions. This is critical to enable the ability to make dynamic data requests in the API. Users can access the REF through two primary interfaces: by directly querying the Data API, and the Portal, which consumes the Data API and serves as the landing page for information about the REF and provides synthesized results.

The REF can be deployed without the need for Kubernetes or docker depending on the constraints of the deployment target. Non-ESGF deployments will look very similar except that the ingestion and solving is performed via the Command Line Interface (CLI). The CLI is used to manually interact with the REF and its database, and is how users interact with the REF in a small deployment.

4.9 Prerequisites for Use

The REF operates on CF-compliant netCDF model output and reference datasets aligned with CMIP CV. When run within ESGF deployments, the REF is not be required to evaluate simulations that fail to meet required data-quality standards; if simulations cannot be processed by the REF, these simulations will be flagged or excluded from the public dashboard. The REF automatically accesses published model output and reference data; however, the REF also supports the analysis of datasets that are not published on ESGF, which allows modeling centers and individual researchers to perform their own model evaluations. A deliberate decision was made to enable the evaluation of datasets and model output that do not conform with the CMIP CV, enabling assessment of model versions not intended to be published on ESGF. In such offline cases, the local datasets must be available to all of the execution services, and ideally should have undergone a quality control (QC) process to assure that the metadata are correct and the data represent the desire output in the correct units.

The REF is designed to run on a variety of hardware platforms, from desktop computers to high performance computing (HPC) environments, and can be run either within virtual Conda environments or within Docker containers. Depending on the use case, either may be used; however, Docker containers are recommended for a production deployment, as they can be scaled independently. The instructions for installation and getting started are available via the REF documentation site at https://climate-ref.readthedocs.io/en/latest/ (last access: 12.06.2026) or the GitHub repository at https://github.com/Climate-REF/climate-ref (last access: 12.06.2026) (Lewis et al., 2026).

5 Discussion

5.1 Community Co-development of a Framework

The REF Delivery Team was composed of an international team that brought together three different evaluation package providers, each of which possessed their own workflows, assumptions, and conventions. An independent delivery team leader was chosen to coordinate the development and architect the end-to-end workflow. Significant time was invested in identifying commonalities among the providers with the goal of minimizing the amount of additional development needed to integrate all the selected diagnostics into the REF. These tools were never designed to be interoperable and have long-standing communities that use the packages outside of the REF, so a requirement for minimal changes to the underlying packages was also mandated.

Developing the appropriate abstractions and common language within the REF depended upon ongoing discussions and often required the prototyping of multiple potential solutions.

The REF system can be quite complex since it is composed of multiple services. A key requirement for the framework was to make it easy for additional packages to be integrated into the REF, both now and in the future. Care was taken to ensure that the amount of context that must be understood for a package developer to contribute diagnostics was kept to a minimum. The decision for the structure of the framework has two impacts, it provides a separation of concerns and minimizes the complexity required simply to contribute diagnostics, as well as providing the scope to support the different assumptions made by different providers. The ability to quickly refactor and modify the interfaces between the packages was critical. The use of a monolithic repository, where all of the REF-related packages were in a common GitHub repository with issue tracking and a Kanban board, was crucial for development and testing. Additionally this structure was important to enforce the need for type hints throughout the code base and providing high test coverage. The REF is openly developed and available on GitHub at https://github.com/Climate-REF/climate-ref under the Apache License 2.0 open source license. The version of the REF contemporaneous with the submission of this manuscript is archived on Zenodo at https://doi.org/10.5281/zenodo.15103441 (Lewis et al., 2025).

The REF Hackathon, hosted by the UK Met Office in Exeter in March 2025, was convened to bring together the REF Delivery Team and to solicit early feedback from potential users. The Hackathon successfully accelerated code development and reference data conversion, obtained feedback and facilitated discussions about approaches for computing diagnostics and supporting traceability for CCO, and tested an alpha version on the UK Met Office HPC facilities. The Hackathon was held relatively early in the project development lifecycle and significant time was spent workshopping the conceptual framework of the REF. After the Hackathon, the research community continued to offer ideas for additional capabilities that could be enabled by the REF framework. A CMIP6-ready version of the REF was subsequently launched at the ESM2025 Final General Assembly on October 7, 2025. It was then further refined based on user feedback and the CMIP7-read version was launched at the CMIP 2026 Community Workshop held in March 2026 in Kyoto, Japan.

5.2 Key Reflections from the CMIP Panel

The CMIP Panel initiated the MB-TT, with a call for members in November 2022. The intention of the Task Team was to put together a group of experts who would work together to provide a systematic, open, and rapid performance assessment of the expected large number of models participating in CMIP7, providing a set of informative diagnostics and performance metrics. The Task Team originally focused on assessing evaluation approaches and software packages. However, it then became apparent that there was a larger opportunity here to fully integrate evaluation tools into the CMIP publication workflow with diagnostic outputs published alongside model output on ESGF and results displayed on an easily accessible website.

The REF became an extremely attractive option to the CMIP panel as it enables 1) consistent assessment of CMIP output by modeling centers, 2) support for author teams in the context of the IPCC and other national climate assessments and 3) input for model selection for downstream applications. The modeling centers are essential to the CMIP endeavor, so having a tool that supports their activities is the most important factor. Many of the CMIP panel members have contributed to IPCC assessments or national climate assessments, so it was clear that there was an opportunity to support such reports if model diagnostics were easily available to authors (reducing the workload of both authors and chapter scientists). Wider community discussions facilitated by the CMIP panel have also revealed enthusiasm from users of CMIP output for regional downscaling and impacts applications. The REF allows to have the diagnostics quickly available to support the choice of a limited selection of climate model output to serve their particular application.

The CMIP Panel view the REF as a potential game changer for users of CMIP data. CMIP Panel co-chair Helene Hewitt reflected, during opening remarks at the March 2025 Hackathon on-boarding session, that "as someone that was a Coordinating Lead Author of the IPCC AR6 WGI, I saw how much work our chapter scientists did in producing all the metrics and plots, this builds on the then-CMIP6 Panel vision, we are so happy that the Task Team has taken this on. We hope that being able to integrate these tools into the workflow and ESGF publication process, on a website, will make it much easier for users to know what they are looking at. Going forward we want to streamline this chain, supporting model selection and downstream use of CMIP." Following the launch of the REF in March at the CMIP 2026 Community Workshop, held in Kyoto, Japan, co-chair of the CMIP Panel John Dunne said, "[T]his release marks an important milestone advancing global accessibility and usability of CMIP data."

5.3 Suggested Uses of the REF

The REF will produce diagnostics spanning a variety of metrics and figures, as well as calculate scalar scores as a gauge of correspondence between model output and reference datasets. The scores are not meant to discriminate "good models" from "bad models." Instead, the REF is intended to assist analysts in quickly identifying relative differences among models or model versions that researchers must then interpret by viewing the plots and maps that underlie the individual diagnostics. REF users must consider that only a limited number of diagnostics can be produced and evaluated and that reference datasets bring with them their own (usually unquantified) uncertainties. While it is hypothetically possible that modelers could "tune" or "overtune" their models to score well on a small set of specific diagnostics, this approach to improving model performance in the REF is of limited practical utility since the reference data are not consistent with each other and the relatively large number of diagnostics and metrics makes such an optimization impossible for a physics-constrained model. We emphasize that while the REF provides a suite of valuable diagnostics of model performance, it cannot replace dedicated analyses for each diagnostic that investigate in detail the mechanisms behind different model behaviors. Model improvements must come from such mechanistic understanding, and the REF may be useful in highlighting where to conduct such a detailed investigation.

Most multi-model assessments indicate that each model has strengths and weaknesses in different areas. For example, one model may perform well with regard to the distribution of precipitation but may not exhibit the desired distribution of sea surface temperatures, while other models may show the opposite behavior. Similarly, some models may capture biogeochemical responses on land well but have a weak representation of the hydrological cycle, while other models may capture runoff well but fail to capture the seasonal cycle of terrestrial productivity. The REF results are best used to identify which subset of models may be best for individual studies based on their performance in key science areas or spatial regions of interest. REF results may help inform the selection of models to include in downscaling studies or the choice of multi-model weights when optimizing for a particular metric. Of course, analysts must consider a wide range of factors when making such selections, including performance on a variety of metrics for relevant model variables (across spatial regions and through time) that may go beyond those incorporated into the initial version of the REF. All of these suggested uses require the researcher to "drill down" into the detailed results produced in the diagnostics, looking beyond performance scores that are merely a high-level indicator of large-scale average correspondence of model output with a reference dataset.

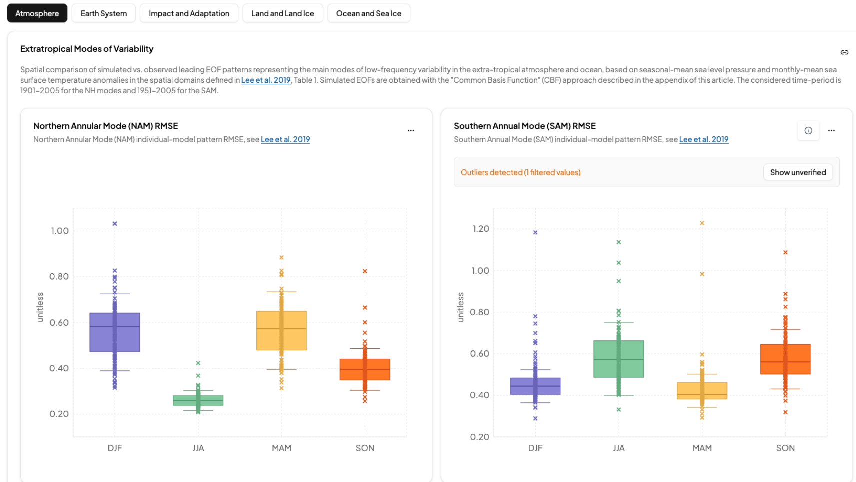
5.4 Example Applications of the REF

To demonstrate the utility of the REF, some example use cases, based on CMIP6 simulations, are described below for the atmosphere, land, and ocean components of models. These few simple examples are intended to highlight potential applications of the REF, not to report any of the features of models named in the descriptions below.

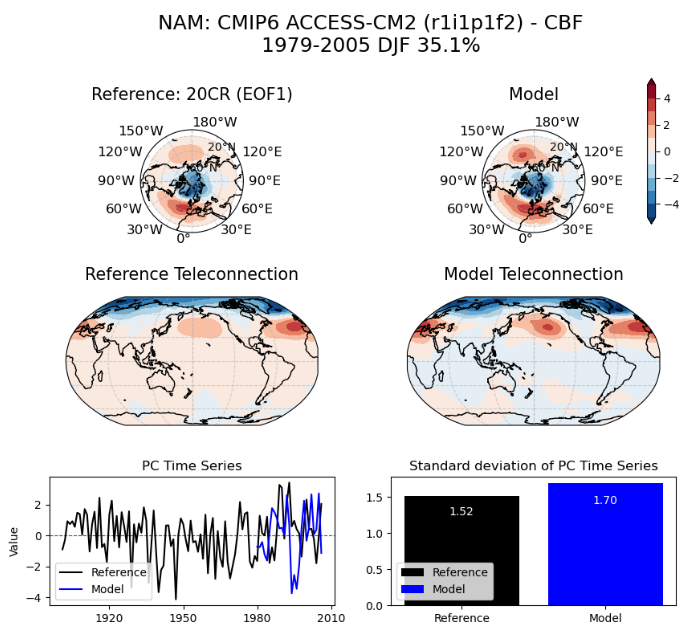
5.4.1 Atmosphere

Figure 3a shows an example metric for a subset of *extratropical modes of variability* in the atmosphere diagnosed through the REF: the Northern and Southern Annular Modes (NAM and SAM) describing the strength of the polar vortices and associated westerly winds in the mid-to-high latitudes in both hemispheres (Gong and Wang, 1999; Thompson and Wallace, 2000). In this example, the NOAA-CIRES 20th Century Reanalysis (Compo et al., 2011) (20CR) is used as observational reference dataset and the historical runs from a large number of different CMIP6 models are evaluated, including multiple runs for the same GCM. The metric shown in the boxplots is the spatial RMSE between the simulated vs. observed leading Empirical Orthogonal Function (EOF1), representing the spatial pattern of the mode, obtained with the *Common Basis Function* method described in (Lee et al., 2019) (see upper panel in Figure 3b). Markers in the Figure 3a indicate the RMSE of individual historical runs and boxes the inter-quartile range (IQR) of the errors of all considered runs. Whiskers are placed at a distance of $1.5 \times IQR$ below and above the 25th and 75th percentiles, respectively. Simulations beyond this range are considered outliers.

The results for the NAM and a *specific* historical simulation (the r1i1p1f2 run of the ACCESS-CM2 model) are illustrated in Figure 3b. The simulated vs. observed EOF1 is mapped in the top row of Figure 3b. The bottom row shows the modeled vs. observed principal component (PC) time series containing the weights to which the pattern is present at a given point in time (here: individual seasons), thus describing the sign and amplitude of the NAM over time. Also shown is a comparison of the modeled vs. observed standard deviation of the PC time series. It is here important to note that a comparison between models and observations only makes sense for *climatological* statistics, such as the temporal standard deviation shown here, but not for year-to-year comparisons, if the *historical* experiment is considered for evaluation. In the middle row, the slope obtained from regressing the grid-box scale sea-level pressure anomalies onto the PC is mapped, indicating the global teleconnection pattern associated with the NAM. A detailed description of the metric and methodology can be found in Lee et al. (2019) and Lee et al. (2025). Results indicate that the GCMs systematically perform poorer during the winter, spring, and autumn seasons, during which the Arctic polar vortex is active, than during the summer season, when the vortex is missing. Similar results are obtained for the SH, where the polar vortex is present throughout the entire calendar year. The NAM teleconnection pattern and PC standard deviation are well reproduced by the specific model run analyzed here.



(a) Spatial RMSE of the simulated vs. observed leading Empirical Orthogonal Function (EOF1) describing the spatial pattern of the Northern and Southern Annular Modes for a large number of historical GCM runs obtained from CMIP6



(b) Simulated vs. observed spatial pattern of the Northern Annular Mode (upper row), the mode's temporal variations in sign and amplitude as described by the principal component time series (PC) and the respective standard deviation (lower row), as well as the global SLP teleconnection pattern associated with the mode (center row); see text for more details

Figure 3. Evaluation metrics for the extratropical modes of variability in the atmosphere. Figures available at https://zenodo.org/records/20162570 (Lee, 2026).

5.4.2 Land

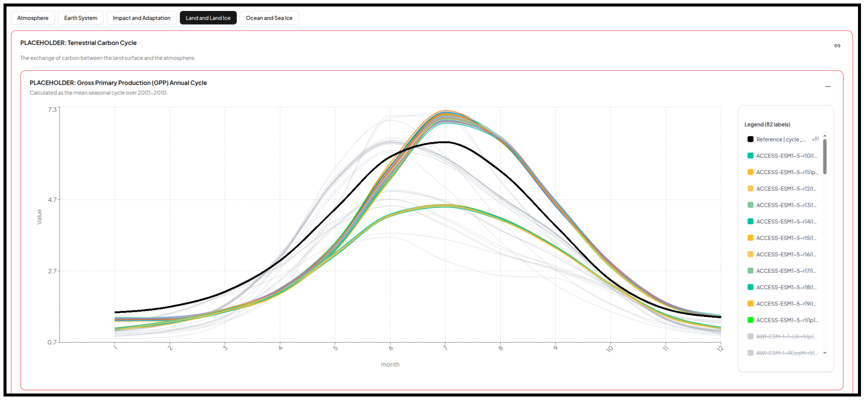
The ability of ESMs to represent plant productivity and its seasonal variation and trend is fundamental to simulating global biogeochemical cycles, the future trajectory of atmospheric carbon dioxide, and hence, the trend in future temperatures. The REF Data Explorer offers a flexible and interactive interface for visualizing the annual cycle of gross primary production (GPP), as shown in Figure 4a. Two models, CanESM5-1 and ACCESS-ESM1-5, were selected arbitrarily as examples to illustrate the diagnostic functionality of the REF. We can see that the CanESM5-1 model ensembles project a higher peak of global GPP than indicated by the reference data, and that peak occurs later in the season than the reference data suggests. Whereas, the ACCESS-ESM1-5 model exhibits timing resembling the reference data, the amplitude of the seasonal cycle is significantly lower than the reference data indicate.

Using diagnostics provided by the ILAMB package, the REF can evaluate model performance across pre-defined or user-defined regions. For example, as shown in Figure 4b, the seasonal timing of the variation in GPP from the ACCESS-ESM1-5 model is approximately correct globally, although the amplitude is weak in the model. However, for the tropical region, the integrated GPP from the model is biased high, and the phase of the seasonal variation is more poorly represented. These graphical outputs help analysts and data users better understand the strengths and weaknesses of models and their implications on future projections.

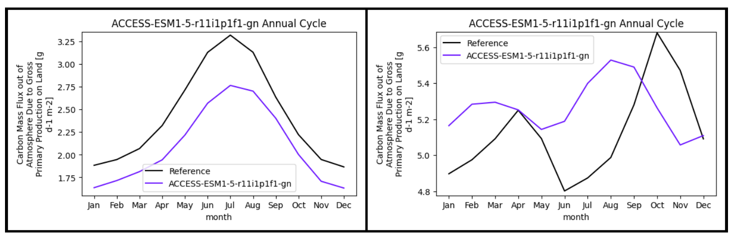
Contour maps in the REF offer a visual representation of the geographic distribution of biases, as shown in Figure 4c. Here, the month of maximum GPP in the reference data and the spatial distribution of the annual mean GPP are shown in the first column. The second column shows the same two quantities as represented by the ACCESS-ESM1-5 model, and the third column shows maps of the differences between the model and the reference data. The upper map in the third column indicates that the seasonal cycle of GPP is better represented by the model in the Northern Hemisphere than in the Southern Hemisphere, and that biases in timing of maximum GPP may lead or lag the observed timing, often by four months or more. The lower map in the third column shows that the strongest annual mean biases from the model occur in the Southern Hemisphere and that tropical biases are primarily positive and often exceed 3 $\mathrm{g\,m^{-2}\,day^{-1}}$.

5.4.3 Ocean

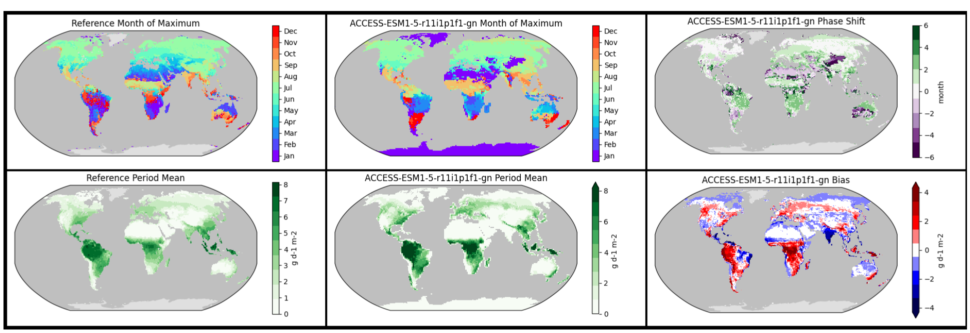
Figure 5 shows the Atlantic meridional overturning circulation (AMOC) as described in C1. In (a) we see multi-model statistics of the AMOC period mean in a box plot to summarize multi-model scalar results for comparison across the ensemble. The red colored ‘x’ mark shows the specific model when the mouse is hovered over any of the individual model values shown as purple ‘x’. In (b), the time series of the annual area averaged time series of the ocean meridional overturning mass streamfunction is plotted for a single model and the reference data set. In (c) a Taylor diagram is used to compare a single model with the reference dataset.



(a) Within the REF Data Explorer for Land and Land Ice, the gross primary production (GPP) annual cycle can be explored, and here, the CanESM5-1 model ensembles (above the black Reference line) and ACCESS-ESM1-5 model ensembles (below the black Reference line) are highlighted in color.



(b) Within the REF Diagnostics for Land and Land Ice, the GPP annual cycles for global land (left) and tropical land only (right) regions for a single ensemble member from the ACCESS-ESM1-5 model are displayed.



(c) Within the REF Diagnostics for Land and Land Ice, the month of maximum GPP (i.e., the phase of the annual cycle) is shown spatially (top row) for the Reference (first column), the ACCESS-ESM1-5 model ensemble (second column), and the difference between the model ensemble and Reference (third column). Similarly, the spatial distribution of the annual integrated GPP from the ACCESS-ESM1-5 model ensemble (second column) is compared to the Reference (first column) and the resulting bias (third column) in the bottom row.

Figure 4. Shown here are example graphics produced in the REF diagnostic for gross primary production (GPP) (Diagnostic ID 2.2 in Tables 1 and B1) that uses the WECANN v1.0 reference dataset (see Table A1). Figures available at https://doi.org/10.5281/zenodo.20162791 (Hoffman, 2026).

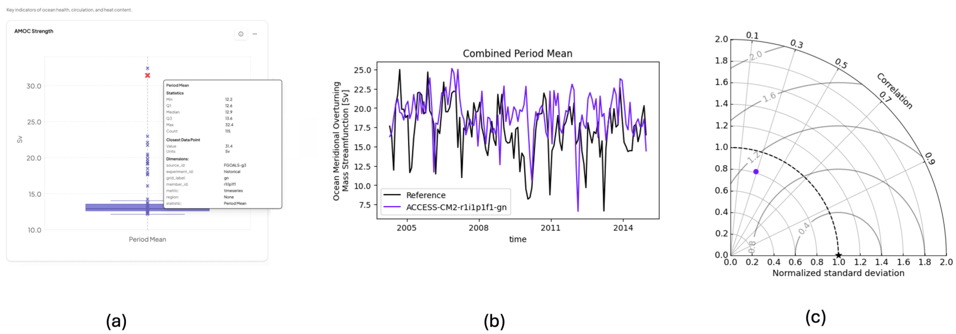


Figure 5. This is an example of the ways in which the REF interactive dashboard allows users to explore results, visualize metrics, and generate plots. Figure available at https://zenodo.org/records/20055565 (Beadling and Swaminathan, 2026).

5.5 The Future of the REF

5.5.1 Potential New Features and Capabilities

A significant advancement for the REF will be its adaptation for daily or sub-daily GCM output in order to capture synoptic and even mesoscale phenomena such as extra-tropical blocking events (Davini and D'Andrea, 2016; Dorrington et al., 2022; Dolores-Tesillos et al., 2025; Bacer et al., 2022), storm tracks (Priestley et al., 2020), weather types (Brands, 2022; Brands et al., 2023), tropical cyclones (Roberts et al., 2020) and other phenomena. These smaller-scale phenomena are known to drive extreme events, which have direct implications for societal impacts. Model evaluation across model generations in general involve dimensionality reduction through empirical orthogonal function (EOF) analysis (Fasullo et al., 2020; Hannachi et al., 2023), k-means clustering (Hoffman et al., 2005) and/or weather regime detection (Grams et al., 2017), for which a number of efficient diagnostics tools have already been developed, e.g., Mid-latitude Evaluation System (MiLES; Davini, 2019) or GCM evaluation with Lamb Weather Types (Brands, 2025).

The REF will likely be utilized by ongoing regional WCRP initiatives such as the Coordinated Regional Downscaling Experiment (CORDEX; Sobolowski et al., 2025), the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP; Warszawski et al., 2014) and the Ice Sheet Model Intercomparison Project (ISMIP; Nowicki et al., 2016), and significant potential exists to connect with other initiatives of this kind, e.g., Atmospheric River Tracking Method Intercomparison Project (ARTMIP; Rutz et al., 2019), etc. Through such collaborations, the REF can serve to foster synergies that will be mutually beneficial across research initiatives. A new collaboration effort, called the CORDEX Collection of Regional-Scale Climate Processes and Metrics for Climate Model Evaluation, will develop high-resolution diagnostics that could be integrated into the REF. As the need for higher spatial and temporal resolution grows, increasing temporal resolution is equally important for understanding extreme weather events, such as the development of tropical cyclones or heatwaves. These high-frequency variations, such as diurnal cycles or sub-daily phenomena, are key to better predicting impacts on society, and they are likely to be useful additions to the REF in the future.

Internal (or unforced) climate variability is commonly sampled by assessing the output of multiple historical (or scenario) runs of the same GCM, each initialized from different dates (branch points) of the corresponding pre-industrial control run. In these equally probable surrogates of the real climate system, the main drivers of natural variability, such as ENSO, as well as its associated teleconnections, evolve freely through time. This "initial conditions uncertainty" (Stainforth et al., 2007) produces random noise, which compared to the predictable signal exerted by external forcing agents (e.g., greenhouse gases and aerosols), is particularly large for atmospheric circulation variables such as sea-level pressure and are not directly affected by global warming (Deser et al., 2012, 2020). Consequently, GCM performance estimates for these variables are expected to be likewise affected by internal variability, especially if they are based on short time periods. However, this kind of error uncertainty has seldom been evaluated in past model performance assessments, particularly outside the atmosphere. Considering internal variability in future versions of the REF is a priority, as it will enable assignment of uncertainty ranges to model performance estimates, thereby making them more robust. This step will also improve understanding of the natural drivers of regional climates and provide better insights into how climate extremes evolve under different scenarios.

Another dimension along which the REF is expected to grow is the consideration of other components of the Earth system beyond the physical processes included in the coupled model configurations contributing to CMIP, such as atmospheric chemistry and terrestrial and marine biogeochemistry, which aim to provide a more accurate representation of the global carbon cycle (Séférian et al., 2020) and its feedback on the Earth system (Arora et al., 2020). As the community increasingly focuses on carbon emissions-driven experiments in CMIP7, the REF must be extended to quantify and reconcile carbon cycle biases (Hoffman et al., 2014) and to include the spatial and temporal evaluation of atmospheric CO_2 and CH_4 variability (Keppel-Aleks et al., 2013). This could help improve understanding of the relationships between carbon emissions and Earth system feedbacks, particularly in high-resolution models. Additionally, systematic evaluation of observed emerging signals of subsurface ocean acidification and deoxygenation, alongside warming, is desirable to benchmark transient changes (Tjiputra et al., 2023). Including these aspects will allow the REF to support a more comprehensive evaluation of oceanic changes that have substantial long-term impacts on marine ecosystems and the global carbon cycle.

The REF will be continuously exercised and updated through the CMIP7 process, and that will undoubtedly reveal opportunities for improvements. Along with such optimizations and incremental improvements, the future evolution of the REF will emerge as it is increasingly used. Many potential extensions of the REF are already envisioned, as illustrated by the above discussion. The expectation is that additional opportunities will be identified and new features will be desired as the next generation of CMIP models are scrutinized. In designing the REF, such evolution has been anticipated, and the design of the software is modular and flexible. This should facilitate community contributions to the REF. Although all the realms of ESMs are already represented by the REF, we anticipate new diagnostic packages may need to be incorporated into the REF. Packages that focus on high-frequency variability (e.g., diurnal cycle, extreme events at the sub-daily time scale) or specific phenomena (e.g., Madden-Julian Oscillation (MJO), ENSO) would be natural extensions of the current capabilities of the REF (e.g. the NOAA Model Diagnostics Task Force (MDTF) diagnostic package). Similarly, observation-based reference datasets will need to be updated as new observations are collected or new products are made available. In all of these future developments, a key aspect for the vitality of the REF will be community engagement. Community contributions and comments are welcome and necessary, as the REF is intended to be an open, community-driven project.

5.5.2 REF Governance

The MB-TT and co-sponsors of the Assessment Fast Track version of the REF agreed, based on significant community interest, that continued development of the REF should be coordinated and prioritized by an international consortium of community evaluation package developers, modelers, reference data providers, and Earth system scientists under WCRP. Such a scientific steering panel (SSP) would engage with the community to identify high priority diagnostics and reference datasets for the full set of CMIP7 simulations and other participating WCRP activities; coordinate with ESGF on quality assurance, platform upgrades, and performance enhancements; initiate and coordinate a technical development and support community; and ensure the open extensibility and portability of subsequent REF developments to support use of the evolving REF framework by modeling centers and individual researchers beyond CMIP activities. The SSP will be established to carry out these duties and to identify and propose to WCRP preferred long-term governance arrangements. An open global recruitment process will be conducted to select SSP members. Contributors to the sustainment and future engineering of the REF will likely acquire their own funding or support for collaborative maintenance and development, which will be coordinated through the REF-SSP.

6 Conclusions

The Rapid Evaluation Framework (REF) is an open source Python-based toolkit (compatible with versions ≥ 3.11) under active development, designed to automate and manage evaluations of Earth system model (ESM) output. The REF's primary objective is to enable near-real-time assessment of ESM output through comparison with well documented reference datasets, updating outputs dynamically as new simulation results are published. Functionally analogous to a Continuous Integration/Continuous Deployment (CI/CD) pipeline in software development, the REF streamlines continuous evaluation workflows for Earth system science. A beta version of the REF, targeted for use and testing by modeling centers and interested researchers, was released to the public on May 27, 2025.

Proposed by the MB-TT, the REF is initially deployed to support the Assessment Fast Track simulation campaign, providing diagnostics collaboratively selected by the MB-TT and the broader science community. While initially tailored for the Assessment Fast Track, the framework is intentionally agnostic to variable types and analytical metrics, ensuring adaptability for diverse Earth science applications beyond its initial scope. This extensibility underscores its potential long-term research utility. Key technical features include integration with CI/CD systems, availability on PyPI from version v0.5.0, comprehensive documentation, and community-driven development under an Apache License 2.0 open source license. Emphasizing collaboration, the REF invites contributions from researchers and developers, fostering a shared ecosystem for advancing model-data evaluation tools. The project's design prioritizes scalability and flexibility, aiming to serve as a foundational resource for real-time ESM benchmarking in both current and future research contexts. Interest in the REF across the research community is strong, and future governance of REF development will be coordinated and prioritized by a scientific steering panel that is currently being formed. A wide variety of new diagnostics are already proposed for integration in subsequent developments of the REF.

Code and data availability. The REF is openly developed and available on GitHub at https://github.com/Climate-REF/climate-ref under the Apache License 2.0 open source license. The version of the REF contemporaneous with the submission of this manuscript is v0.6.1 and is archived along with all versions on Zenodo at https://doi.org/10.5281/zenodo.15103441 (Lewis et al., 2025). Reference data used by the REF are available from the Earth System Grid Federation (ESGF) nodes listed at https://esgf.github.io/nodes.html. Original reference data are listed with appropriate citations in Table A1.

Appendix A: Reference Datasets Used by the REF

Table A1. Observationally constrained reference datasets shown here are used in the REF for comparison with model output. (*) indicates reference dataset ends more than five years before the end date of the CMIP7 historical simulations.

Reference Dataset	Citation	Start	End	Freq.	Diagnostic ID
C3S-GTO-ECV v9.0	Coldewey-Egbers et al. (2025, 2015)	1995-01-07	2022-04-30	monthly	3.1
CALIPSO-ICECLOUD v2.0	Winker (2024); Winker et al. (2024)	2007-01-01	2022-12-31	monthly	3.7
CCI-CryoClim-FSC v1.0	Solberg et al. (2023)	1978-11-09	2025-03-05	daily	2.7
CERES-EBAF v4.2	Loeb et al. (2018)	2000-03-15	2023-06-15	monthly	3.1, 3.2
CERES-EBAF v4.2.1	Loeb et al. (2018)	2000-03-15	2025-09-15	monthly	3.6, 3.7
CLARA-AVHRR-A3	Karlsson et al. (2023)	1979-01-01	2020-12-31	monthly	3.7
CRU TS v4.02	Harris et al. (2020)	1901-01-01	2017-12-31	monthly	2.2
ERA5	Hersbach et al. (2020)	1979-01-16	2022-12-16	monthly	3.1, 3.7, 4.5
Fluxnet2015*	Pastorello et al. (2020)	1991-01-01	2014-12-31	monthly	2.2
GFED v5.0	Chen et al. (2023a, b)	2001-01-01	2020-01-31	monthly	5.4
GLEAM v3.3a	Martens et al. (2017); Miralles et al. (2011)	1980-01-01	2019-12-31	monthly	2.8
GPCP v2.3	Adler et al. (2017, 2003, 2018)	1979-01-16	2019-07-16	monthly	1.3, 2.2, 2.8
GPCP v3.3	Adler et al. (2017, 2003, 2018)	1983-01-16	2024-09-16	monthly	1.3, 3.1, 3.7, 4.5
HadCRUT5	Morice et al. (2021)	1850-01-01	2025-02-28	monthly	1.1, 4.5
HadISST v1.1	Rayner et al. (2003)	1870-01-01	2025-01-31	monthly	1.3, 3.3
Hoffman v1.0*	Hoffman et al. (2014)	1850-01-08	2010-12-31	annual	2.5
HWSD v2.0	FAO and IIASA (2023)	2008-01-01	2022-12-31	monthly	2.1
LORA v1.0*	Hobeichi et al. (2019)	1980-01-01	2012-12-31	monthly	2.3
NOAA-20CR v2.0*	Slivinski et al. (2019)	1871-01-01	2012-12-31	monthly	3.3
NOAA-NCEI-LAI-AVHRR v5.0	Claverie et al. (2014, 2024)	1981-07-01	2024-12-31	monthly	2.6
OSI-450-nh	Lavergne et al. (2019)	1979-01-01	2020-12-31	monthly	1.1, 1.6
OSI-450-sh	Lavergne et al. (2019)	1979-01-01	2020-12-31	monthly	1.1, 1.6
RAPID-2023 v1a	Moat et al. (2025)	2004-04-02	2023-02-11	monthly	1.2
SAGE-CCI-OMPS v8.0	Copernicus Climate Data Store (2020); Sofieva et al. (2023); Coldewey-Egbers et al. (2025)	1984-10-01	2022-12-31	monthly	3.1
TropFlux v1.0	Kumar et al. (2012)	1979-01-16	2017-07-16	monthly	1.3
WangMao*	Wang and Mao (2021); Wang et al. (2021)	1900-01-01	2010-01-01	monthly	2.4
WECANN v1.0*	Alemohammad et al. (2017)	2007-01-01	2015-12-31	monthly	2.2
WOA23	Reagan et al. (2023)	2005-01-01	2022-12-31	monthly	1.4

Appendix B: Diagnostics Produced by the REF

Table B1. Based on community recommendations, an initial set of diagnostics was selected, published at https://zenodo.org/records/14284375, for incorporation into the initial version of the REF for evaluating relevant CMIP7 Assessment Fast Track simulations. Diagnostics 1.5, 3.4, 3.5, 5.1 and 5.2 are listed in strikethrough style because they were not implemented for the initial version of the REF. Diagnostic 2.8 was added in lieu of 3.4. The column representing CMIP7 physical variables and branded variables does not include fixed grid information (e.g., areacella, areacello) that may be required for execution of the diagnostic.

ID	Realm	Diagnostic Collection	Package(s)	CMIP7 Physical Variables and [Branded Variables]	DECK & AFT Experiments	Reference Dataset(s)
1.1	Ocean & Sea Ice	Antarctic annual mean, Arctic September rate of sea ice area (SIA) loss per degree warming (dSIA / dGMST)	ESMValTool	SImon.siconc [siconc_tavg-u-hxy-u], Amon.tas [tas_tavg-h2m-hxy-u]	historical, esm-hist	OSI-450-nh, OSI-450-sh, HadCRUT5
1.2	Ocean & Sea Ice	Atlantic meridional overturning circulation (AMOC)	IOMB	Omon.msftmz [msftm_tavg-ol-hyb-sea]	historical, esm-hist	RAPID-2023 v1a
1.3	Ocean & Sea Ice	El Niño Southern Oscillation (ENSO) diagnostics (lifecycle, seasonality, amplitude, teleconnections)	PMP, ESMValTool	Amon.pr [pr_tavg-u-hxy-u], Amon.tauu [tauu_tavg-u-hxy-u], Omon.tos [tos_tavg-u-hxy-sea]	historical, esm-hist	TropFlux v1.0, GPCP v2.3, GPCP v3.3, HadISST v1.1
1.4	Ocean & Sea Ice	Sea surface temperature (SST) bias, Sea surface salinity (SSS) bias	IOMB	Omon.sos [sos_tavg-u-hm-sea], Omon.tos [tos_tavg-u-hxy-sea]	historical, esm-hist	WOA23 (climatology)
~~1.5~~	~~Ocean & Sea Ice~~	~~Ocean heat content (OHC)~~				
1.6	Ocean & Sea Ice	Antarctic & Arctic sea ice area seasonal cycle	ESMValTool	SImon.siconc [siconc_tavg-u-hxy-u]	historical, esm-hist	OSI-450-nh, OSI-450-sh
2.1	Land & Land Ice	Soil carbon	ILAMB	Emon.cSoil [cSoil_tavg-u-hxy-lnd]	historical, esm-hist, land-hist	HWSD v2.0
2.2	Land & Land Ice	Gross primary production (GPP)	ILAMB	Lmon.gpp [gpp_tavg-u-hxy-lnd], Amon.pr [pr_tavg-u-hxy-u], Amon.tas [tas_tavg-h2m-hxy-u]	historical, esm-hist, land-hist	WECANN v1.0, Fluxnet2015, GPCP v2.3, CRU TS v4.02
2.3	Land & Land Ice	Runoff	ILAMB	Lmon.mrro [mrro_tavg-u-hxy-lnd]	historical, esm-hist, land-hist	LORA v1.0
2.4	Land & Land Ice	Surface soil moisture	ILAMB	Lmon.mrso [mrso_tavg-u-hxy-lnd]	historical, esm-hist, land-hist	WangMao

continued on next page

continued from previous page

ID	Realm	Diagnostic Collection	Package(s)	CMIP7 Physical Variables and [Branded Variables]	DECK & AFT Experiments	Reference Dataset(s)
2.5	Land & Land Ice	Net ecosystem carbon balance	ILAMB	netAtmosLandCO2Flux [netAtmosLandCO2Flux _tavg-u-hxy-lnd] or nbp [nbp_tavg-u-hxy-lnd]	historical, esm-hist, land-hist	Hoffman v1.0
2.6	Land & Land Ice	Leaf area index (LAI)	ILAMB	Lmon.lai [lai_tavg-u-hxy-lnd]	historical, esm-hist, land-hist	NOAA-NCEI-LAI-AVHRR v5.0
2.7	Land & Land Ice	Snow cover	ILAMB	Eday.snc [snc_tavg-u-hxy-lnd]	historical, esm-hist, land-hist	CCI-CryoClim-FSC v1.0
2.8	Land & Land Ice	Precipitation minus evaporation $(P - E)$	ILAMB	Amon.evpspbl [evspsbl_tavg-u-hxy-u], Amon.pr [pr_tavg-u-hxy-u]	historical, esm-hist, land-hist	GLEAM v3.3a, GPCP v2.3
3.1	Atmosphere	Annual cycle and seasonal mean of multiple variables	PMP, ESMValTool	Amon.o3 [o3_tavg-p19-hxy-air], Amon.psl [psl_tavg-u-hxy-u], Amon.pr [pr_tavg-u-hxy-u], Amon.rlds [rlds_tavg-u-hxy-u], Amon.rlus [rlus_tavg-u-hxy-u], Amon.rlut [rlut_tavg-u-hxy-u], Amon.rsds [rsds_tavg-u-hxy-u], Amon.rsus [rsus_tavg-u-hxy-u], Amon.rsut [rsut_tavg-u-hxy-u], Amon.ts [ts_tavg-u-hxy-u], Amon.ua [ua_tavg-p19-hxy-air], Amon.va [va_tavg-p19-hxy-air]	amip, historical, esm-hist	ERA5, SAGE-CCI-OMPS v8.0, C3S-GTO-ECV v9.0, GPCP v3.3, CERES-EBAF v4.2

continued on next page

continued from previous page

ID	Realm	Diagnostic Collection	Package(s)	CMIP7 Physical Variables and [Branded Variables]	DECK & AFT Experiments	Reference Dataset(s)
3.2	Atmosphere	Radiative and heat fluxes at the surface and top of atmosphere (TOA)	PMP	Amon.rlds [rlds_tavg-u-hxy-u], Amon.rlus [rlus_tavg-u-hxy-u], Amon.rlut [rlut_tavg-u-hxy-u], Amon.rlutcs [rlutcs_tavg-u-hxy-u], Amon.rsds [rsds_tavg-u-hxy-u], Amon.rsdscs [rsdscs_tavg-u-hxy-u], Amon.rsdt [rsdt_tavg-u-hxy-u], Amon.rsut [rsut_tavg-u-hxy-u], Amon.rsutcs [rsutcs_tavg-u-hxy-u]	amip, historical, esm-hist	CERES-EBAF v4.2
3.3	Atmosphere	Climate variability modes (e.g., ENSO, Madden-Julian Oscillation (MJO), Extratropical modes of variability, monsoon)	PMP	Amon.psl [psl_tavg-u-hxy-u], Omon.tos [tos_tavg-u-hxy-sea]	amip, historical, esm-hist	NOAA-20CR v2.0, HadISST v1.1
~~3.4~~	~~Atmosphere~~	~~Evaporation minus precipitation ($E-P$)~~				
~~3.5~~	~~Atmosphere~~	~~Double inter-tropical convergence zone (ITCZ)~~				
3.6	Atmosphere	Cloud radiative effects	ESMValTool	Amon.rlut [rlut_tavg-u-hxy-u], Amon.rlutcs [rlutcs_tavg-u-hxy-u], Amon.rsut [rsut_tavg-u-hxy-u], Amon.rsutcs [rlutcs_tavg-u-hxy-u]	amip, historical, esm-hist	CERES-EBAF v4.2.1
3.7	Atmosphere	Scatterplots of two cloud-relevant variables (for specific regions of the globe and specific cloud regimes)	ESMValTool	Amon.cli [cli_tavg-al-hxy-u], Amon.clivi [clivi_tavg-u-hxy-u], Amon.clt [clt_tavg-u-hxy-u], Amon.clwvi [clwvi_tavg-u-hxy-u], Amon.pr [pr_tavg-u-hxy-u], Amon.rlut [rlut_tavg-u-hxy-u], Amon.rlutcs [rlutcs_tavg-u-hxy-u], Amon.rsut [rsut_tavg-u-hxy-u, Amon.rsutcs [rsutcs_tavg-u-hxy-u], Amon.ta [ta_tavg-p19-hxy-air]	amip, historical, esm-hist	CLARA-AVHRR-A3, GPCP v3.3, CERES-EBAF v4.2.1, CALIPSO-ICECLOUD v2.0, ERA5

continued on next page

continued from previous page

ID	Realm	Diagnostic Collection	Package(s)	CMIP7 Physical Variables and [Branded Variables]	DECK & AFT Experiments	Reference Dataset(s)
4.1	Earth System	Equilibrium climate sensitivity (ECS)	ESMValTool	Amon.rsdt [rsdt_tavg-u-hxy-u], Amon.rlut [rlut_tavg-u-hxy-u], Amon.rsut [rsut_tavg-u-hxy-u], Amon.tas [tas_tavg-h2m-hxy-u]	piControl, abrupt-4xCO2	N/A
4.2	Earth System	Transient climate response (TCR)	ESMValTool	Amon.tas [tas_tavg-h2m-hxy-u]	1pctCO2, piControl	N/A
4.3	Earth System	Transient climate response to cumulative emissions of carbon dioxide (TCRE)	ESMValTool	Amon.fco2antt [fco2antt_tavg-u-hxy-u], Amon.tas [tas_tavg-h2m-hxy-u]	esm-1pctCO2*, esm-piControl, esm-flat10	N/A
4.4	Earth System	Zero emissions commitment (ZEC)	ESMValTool	Amon.tas [tas_tavg-h2m-hxy-u]	esm-flat10, esm-flat10-zec, 1pctCO2*, esm-1pct-brch-1000Pg*	N/A
4.5	Earth System	Historical changes in climate variables (time series, trends)	ESMValTool	Amon.hus [hus_tavg-p19-hxy-u], Amon.pr [pr_tavg-u-hxy-u], Amon.psl [psl_tavg-u-hxy-u], Amon.tas [tas_tavg-h2m-hxy-u], Amon.ua [ua_tavg-p19-hxy-air]	amip, historical, esm-hist	HadCRUT5, GPCP v3.3, ERA5
~~5.1~~	~~Impacts & Adaptation~~	~~High amplitude Rossby waves~~				
~~5.2~~	~~Impacts & Adaptation~~	~~Internal variability or ensemble spread for individual models (precipitation and surface temperature)~~				
5.3	Impacts & Adaptation	Evaluation of key climate variables at global warming levels	ESMValTool	Amon.pr [pr_tavg-u-hxy-u], Amon.tas [tas_tavg-h2m-hxy-u]	amip, historical, esm-hist, scenarios	N/A

continued on next page

continued from previous page

ID	Realm	Diagnostic Collection	Package(s)	CMIP7 Physical Variables and [Branded Variables]	DECK & AFT Experiments	Reference Dataset(s)
5.4	Impacts & Adaptation	Climate drivers for fire (fire burnt area, fire weather and fuel continuity)	ILAMB, ESMValTool	Lmon.burntFractionAll [burntFractionAll_tavg-u-hxy-u], Lmon.cVeg [cVeg_tavg-u-hxy-lnd], Amon.hurs [hurs_tavg-h2m-hxy-u], Amon.pr [pr_tavg-u-hxy-u], Lmon.treeFrac [treeFrac_tavg-u-hxy-u], Emon.vegFrac [vegFrac_tavg-u-hxy-u], Amon.tas [tas_tavg-h2m-hxy-u]	historical, esm-hist, land-hist, scenarios	GFED v5.0, Weights†

*CMIP6 variable; not available/necessary for CMIP7 Assessment Fast Track

†Pre-computed weights are available at https://zenodo.org/records/14917245 (Kelley et al., 2025)

Appendix C: Rationale for Diagnostics Produced by the REF

This Appendix contains the list of diagnostics, v1.0 (CMIP Model Benchmarking Task Team, 2024), originally devised by the CMIP Model Benchmarking Task Team to be implemented in the CMIP7 Assessment Fast Track Rapid Evaluation Framework (REF).

C1 Ocean & Sea Ice Realm

1.1 Antarctic annual mean, Arctic September rate of sea ice area (SIA) loss per degree warming (dSIA / dGMST)

Sea ice responds strongly to climate forcing and warming. The rapid decline in Arctic summer sea ice areal extent is a highly visible indicator of climate change. Previous sea ice benchmarking assessments have highlighted the fact that CMIP models systematically underestimate the amount of Arctic sea ice area loss per degree of global warming. Very few CMIP models are able to simulate both a plausible sea ice loss and a plausible change in global mean temperature over the satellite period. This metric evaluates the rate of sea ice loss per degree of global warming, following the approach used for sea ice benchmarking within the Sea Ice Model Intercomparison Project analysis (Notz and SIMIP Community, 2020; Roach et al., 2020). The metric is calculated by regressing the time-series of sea ice area on global mean temperature on an annual basis. The annual mean sea ice area is used for the Antarctic, while the September-mean sea ice area is used for the Arctic.

1.2 Atlantic meridional overturning circulation (AMOC)

The Atlantic Meridional Overturning Circulation (AMOC) provides a key indicator of the strength of ocean circulation, which redistributes freshwater, heat and carbon across the Atlantic Basin (Le Bras et al., 2023). A weakening of AMOC, expected to occur as a result of ocean warming, will have global climate consequences. The strength of the AMOC at 26.5°N is commonly used for evaluation of model fidelity since it can be compared with the long-term RAPID-MOCHA (Rapid Climate Change - Meridional Overturning Circulation and Heatflux Array) observational dataset (Moat et al., 2025). RAPID datasets are widely used to validate the AMOC strength in models and are available from 1st April 2005 to 11th February 2023. The AMOC at 26.5°N is calculated as the maximum of the meridional overturning streamfunction, which is provided by CMIP models as the variable “msftmz”. The AMOC is a key component of the global ocean conveyor belt and plays an important role in transporting heat poleward and ocean biogeochemical tracers from the surface into the ocean interior.

1.3 El Niño Southern Oscillation (ENSO) diagnostics (lifecycle, seasonality, amplitude, teleconnections)

The El Niño Southern Oscillation (ENSO) is the primary mode of the global interannual climate variability, mainly reflected by the variations in surface wind stress and ocean temperature in the tropical Pacific Ocean. Through teleconnections, the ENSO affects seasonal temperature and precipitation in other parts of the globe (Chen and Wallace, 2015; Vaittinada Ayar et al., 2023). The ENSO variability can be calculated from both sea surface temperature and atmospheric pressure differences between different tropical Pacific areas. The Southern Oscillation Index (SOI) uses pressure differences between the Tahiti and

Darwin regions. The Oceanic Niño Index (ONI) summarizes SST anomalies in the Niño 3.4 region. Given its implications for regional climate variability, capturing the observed ENSO spatial and temporal characteristics would increase the fidelity and robustness in a model's climate projections.

1.4 Sea surface temperature (SST) bias, Sea surface salinity (SSS) bias

The SST and SSS distributions provide large scale patterns of surface ocean circulation as well as reflecting dynamical air-sea interactions and ocean-sea ice interactions in the polar regions. SST and SSS biases have a significant impact on the coupling of ESM's two majors components, the atmosphere and the ocean. Satellite data products and localized moored sensors are used to produce measurements that are incorporated into reference data to calculate SST and SSS biases in models.

1.6 Antarctic & Arctic sea ice area seasonal cycle

The sea ice area, calculated as the sum over the Northern (Arctic) and Southern (Antarctic) Hemisphere grid cell areas multiplied by the sea ice fraction within each cell, exhibits a distinct seasonal cycle. Arctic sea ice area typically has minimum values in September, while Antarctic sea ice area is lowest in February. The seasonal cycle is driven by the seasonal cycle of the insolation, sea ice processes, as well as the exchange with the atmosphere and ocean and can be seen as an overview metric for the general state of the sea ice in a model. Since sea ice has a much higher albedo than the ocean surface, sea ice area plays an important role in the surface energy and radiation budgets. In addition to the multi-year average seasonal cycle of Arctic and Antarctic sea ice area, the diagnostic also produces time series of the September (Arctic) and February (Antarctic) sea ice area.

C2 Land & Land Ice Realm

2.1 Soil carbon

Soil carbon is the organic matter and inorganic carbon in global soils. It is an important component of the global carbon cycle and affects soil moisture retention and saturation. Soils are the largest stores of carbon on Earth, and warming can impact the ability of soils to store and retain carbon, especially in the Arctic, where permanently frozen soil can release large quantities of carbon into the atmosphere if it thaws due to warming (Trumbore and Czimczik, 2008). Analyzing stored soil carbon helps track quantify the dynamics of the terrestrial carbon cycle within models and the movement of carbon through the Earth system.

2.2 Gross primary production (GPP)

Gross primary production is the process by which plants "fix" atmospheric or aqueous carbon dioxide through photosynthetic reduction into organic compounds, and it is affected by increases in atmospheric carbon dioxide (CO_2) levels and warming (Anav et al., 2015). A fraction of gross primary productivity supports plant respiration and the rest is stored as biomass in stems, leaves, roots, or other plant parts. Land use change, heat and drought stress due to anthropogenic warming, and rising atmospheric CO_2 will differentially influence gross primary production in ecosystems and alter the global carbon cycle. Thus, models must be evaluated to ensure they capture the observed responses to these changes.

2.3 Runoff

Surface water runoff plays an important role in the hydrological cycle by returning excess precipitation to the oceans and controlling how much water flows into water systems (Trenberth et al., 2007; Trenberth and Caron, 2001). Changes in atmospheric circulation and distributions of precipitation have a direct effect on changes in runoff from land. Models must be evaluated to ensure they exhibit the observed responses to precipitation and soil moisture processes that lead to runoff and transport of freshwater into rivers and oceans.

2.4 Surface soil moisture

Surface soil moisture is an important hydrological cycle variable governing interactions between the land surface and atmosphere, and with the oceans through surface water runoff. Soil moisture partitions incoming energy into latent and sensible heat fluxes and also controls how precipitation is partitioned into runoff or for evapotranspiration. It therefore is at the nexus of the carbon, energy, and water cycles and is key to understanding a broad range of processes from drought, floods to agricultural management. Models must be evaluated to ensure they exhibit observed responses in soil moisture to variation in precipitation, temperature, land use change, soil carbon content and biogeochemical cycles (Seneviratne et al., 2010).

2.5 Net ecosystem carbon balance

The net ecosystem carbon balance is a quantitative estimate of the overall terrestrial carbon uptake or loss, sometimes referred to as the land carbon sink (Keenan and Williams, 2018). The land sink represents the annual carbon uptake on land through gross primary production minus losses through ecosystem respiration, disturbance, wood and agricultural harvest, and land use change. Along with the global marine carbon sink, the global terrestrial carbon sink is important for sequestering anthropogenic carbon from the atmosphere and is influenced by climate change. Models must be evaluated to ensure they exhibit observed responses in the net ecosystem carbon balance since it directly affects the amount anthropogenic carbon retained in the atmosphere (Le Quéré et al., 2018).

2.6 Leaf area index (LAI)

Leaf area index (LAI) is a measure of the total area of leaves per unit ground area, and it is used to characterize the structure of plant canopies (Fang et al., 2019). Thus, it is an important variable to model mass and energy exchange through leaf surfaces between the biosphere and atmosphere. LAI is influenced by gross primary production and plant physiological allocation of carbon to leaves. Models must be evaluated to ensure they exhibit observed responses to seasonal and trend changes in LAI in response to variations in temperature, precipitation, soil moisture, nutrient availability, and atmospheric CO_2.

2.7 Snow cover

Globally, snow cover is a key determinant of the Earth surface albedo and is also known to affect the large-scale atmospheric circulation, particularly in the Northern Hemisphere (Barnett et al., 1989; Cohen and Entekhabi, 1999). It is also a key element of the Arctic Amplification phenomenon (Cohen et al., 2014) and directly impacts the terrestrial hydrological cycle by providing a meltwater source in mid-latitudes. On the regional scale, snow cover is a paramount climate driver in the mid-to-high latitudes of the Northern Hemisphere, where it determines the length of the winter season and associated significant changes in hydrology, soil properties, and vegetation activity. A persistent snow-cover essentially blocks interactions between the atmosphere and underlying land-surface (Bokhorst et al., 2016). Models must be evaluated to ensure they exhibit observed responses to seasonal and trend changes in precipitation, temperature, and other important driving mechanisms.

2.8 Precipitation minus evaporation ($P - E$)

The seasonal state of water balance on land offers insights into water availability, soil moisture, latent heat energy, and runoff. While these states and fluxes can be independently evaluated, the overall water balance on land is ultimately the input (precipitation) minus the output (evaporation) exchanges with the atmosphere. In this case, precipitation is the moisture that falls on land in any form: solid, liquid, or vapor. Here, evaporation includes evaporation of water from all surfaces, including soils, plants, impervious surfaces, and rivers and reservoirs; as well as transpiration from plants and sublimation from frozen water. Models need to be evaluated to determine, not only if the soil moisture and humidity are correct, but if they are correct because the model sufficiently captured the water balance of fluxes with the atmosphere.

C3 Atmosphere Realm

3.1 Annual cycle and seasonal mean of multiple variables

The annual cycle provides an integrative measure of skill at one of the fundamental forced time scales, yet ESMs often exhibit pathologic biases in the phase or amplitude of key quantities (e.g., Scaife et al., 2010; Hoffman et al., 2014; Miller et al., 2015). Evaluating the seasonal and annual variability in models helps to ensure that key processes are correctly represented in models.

3.2 Radiative and heat fluxes at the surface and top of atmosphere (TOA)

The radiative and heat fluxes at the surface and top of atmosphere (TOA) represent the fundamental flows of energy through the climate system. The TOA radiative budget indicates the relative disequilibrium of the climate system, and is the primary approach to quantifying the equilibrium climate sensitivity. Similarly, capturing the transient response at TOA is a basic measure of a model's ability to realistically represent the system response to forcing (Loeb et al., 2020). Fluxes at the surface quantify the energy exchange between the atmosphere and surface and are directly related to the redistribution of heat through the system and the ocean heat uptake (Mayer et al., 2024). Imbalances at the TOA in unforced experiments are a typical source of bias, inducing spurious drifts in model behavior (e.g., Mauritsen et al., 2012), inducing an erroneous interpretation of the thermodynamics of the climate system (Lucarini et al., 2017) and in particular of the relation between heat gradients and the general circulation of the coupled atmosphere-ocean system, as well as the patterns of the response of the system to an inhomogeneous forcing (Irving et al., 2019; Lembo et al., 2019). Calculations of radiative and heat fluxes at the surface and TOA are usually carried out by the radiative transfer modules in atmospheric models, and parametrization of latent and sensible heat fluxes at the surface through bulk formulas. The imbalances are obtained by combining these fluxes, whereas transports can either be implied through integration of fluxes (Trenberth and Caron, 2001), or explicitly retrieved through evaluation of the internal energy within the atmosphere (Moist Static Energy; MSE) or inside the oceans (ocean heat content; Cheng et al., 2024) and other subdomains.

3.3 Climate variability modes (e.g., NAM, NAO, NPGO, NPO, PDO, PNA, SAM)

The main modes of low-frequency variability in the atmosphere, such as the North Atlantic, Arctic and Antarctic Oscillations, the Pacific North American and Pacific South American patterns, are important drivers of climate variability on the hemispheric- to continental-scale because they determine the broad direction and strength of the atmospheric flow (Wallace and Gutzler, 1981). Since the flow determines the temperature and moisture characteristics of the air masses transported to a any specific region, these modes also explain a significant fraction of the regional-scale climate variability around the globe (Hurrell et al., 2001). Originating in the equatorial Pacific, the El Niño-Southern Oscillation (ENSO) is the most important (ocean-atmosphere) mode of climate variability and is associated with typical climate anomalies in many regions around the globe (Trenberth et al., 1998). Most of these modes are considered internal (or unforced) oscillations, meaning that their characteristics are largely robust to anthropogenic forcing (Deser et al., 2012). Due to their large scale, they are the primary diagnostics a global climate model should be able to reproduce, before one would proceed to evaluate model performance on smaller scales (Fernández-Granja et al., 2024).

3.6 Cloud radiative effects

Clouds play an important role in climate by reflecting incoming solar radiation (shortwave) and by absorbing and emitting outgoing thermal radiation (longwave). These cloud radiative effects can be quantified by calculating the differences in TOA clear-sky and all-sky radiative fluxes. The diagnostic calculates multi-year annual average shortwave and longwave TOA cloud radiative effects that are then used to create maps and zonal means of the cooling and warming effects of clouds of the models in comparison to observations.

3.7 Scatterplots of two cloud-relevant variables (for specific regions of the globe and specific cloud regimes)

Despite their pivotal role in the radiation budget and the hydrological cycle, clouds have proven notoriously challenging to simulate with global climate models. The diagnostic investigates the relationship between modeled integral cloud properties such as total cloud cover, cloud water path (sum of cloud ice and cloud liquid) and cloud ice water path and climate-relevant quantities such as long- and shortwave TOA cloud radiative effects and precipitation that can then be compared to observations. In addition, the relation between three-dimensional cloud ice water content and three-dimensional air temperature is investigated and compared to reference data.

C4 Earth System Realm

4.1 Equilibrium climate sensitivity (ECS)

ECS is defined as the change in global mean near-surface air temperature that results from an instantaneous doubling of the atmospheric CO_2 concentration after the climate system has reached its new equilibrium. To avoid having to simulate thousands of model years until the system has reached equilibrium, ECS is usually approximated with the "effective climate sensitivity" following a method by Gregory et al. (2004), which estimates ECS as the x-intercept of a linear regression of the change in global mean TOA net radiation flux against the change in global mean near-surface air temperature (see Schlund et al. (2020) for details). Even though ECS is an idealized metric, it is still relevant for policymakers and scientists since it measures the equilibrium warming response of the climate system to CO_2 forcing.

4.2 Transient climate response (TCR)

TCR is defined as the global mean near-surface air temperature change at the time of CO_2 doubling in a simulation where the atmospheric CO_2 concentration is increased by 1% per year from the preindustrial level (Gregory et al., 2009). In practice, it is calculated by averaging the global mean near-surface air temperature change over 20-year period centered around the time of CO_2 doubling (year 70) in the 1pctCO2 increase simulation (see Meehl et al. (2020) for details). Similar to ECS, it is an idealized metric that describes the warming response of the climate system to CO_2 forcing, but unlike ECS, it does not assume radiative equilibrium of the system.

4.3 Transient climate response to cumulative emissions of carbon dioxide (TCRE)

Unlike ECS and TCR, TCRE describes the warming response of the Earth system to cumulative emissions of CO_2 instead of atmospheric CO_2 concentrations and thus takes carbon cycle feedbacks into account (Gregory et al., 2009). Following Sanderson et al. (2025), TCRE is estimated from an experiment with constant CO_2 emissions of 10 Pg C per year ("esm-flat10") as the global mean near-surface air temperature change observed after the emission of 1000 Pg C of CO_2 (i.e., at year 100) averaged over a 20-year period (i.e., years 90–110). TCRE is one of the most important policy-relevant climate change metrics since it directly links global warming and CO_2 emissions in a linear relation, which allows a straightforward estimation of carbon budgets that remain to reach specific warming targets.

4.4 Zero Emissions Commitment (ZEC)

The Zero Emissions Commitment (ZEC) quantifies the change in global mean temperature expected to occur after net carbon dioxide (CO_2) emissions cease (MacDougall et al., 2020). ZEC is therefore important to consider when estimating the remaining carbon budget. Calculation of ZEC requires dedicated simulations with anthropogenic carbon emissions set to zero, branching off a base simulation. For CMIP7 fast track, the base simulation will be "esm-flat10", while the dedicated simulation will be called "esm-flat10-zec" (Sanderson et al., 2025).

4.5 Historical changes in climate variables (time series, trends)

To assess the ability of climate models to look into the future, it is important to evaluate them on the changes, which have been observed in recent decades, where observations and reanalysis data provide a reference. In addition to the global mean, time series and trends of key climate variable like temperature, pressure, wind and humidity are compared regionally, using, e.g., the IPCC climate reference regions for subcontinental analysis of climate model data (Iturbide et al., 2020). A regional analysis is important to assure climatic consistency and the representation of regional features within global model data as well as climate model data with regional focus, e.g., from the Coordinated Regional climate Downscaling Experiment (CORDEX; Diez-Sierra et al., 2022; Sobolowski et al., 2025).

C5 Impacts & Adaptation Realm

5.3 Evaluation of key climate variables at global warming levels

Global warming level (GWL) exceedance years or the years at which the global mean surface temperature warms over specific values is an important marker of climate change. The exceedance years are calculated as the 21-year mean anomaly of the area averaged global mean surface temperature with respect to the pre-industrial control mean temperature as described in Swaminathan et al. (2022). Assessing global values of key climate variables such as temperature and precipitation at specific GWLs can tell us how different regions are affected by climate change and whether these changes scale linearly with increases in temperature thereby providing important information for adaptation and mitigation efforts. Evaluating climate at GWLs is also relevant for policy, for instance the UNFCCC Paris agreement's central aim is to keep warming below 2°C and if possible below 1.5°C.

5.4 Climate drivers for fire (fire burnt area, fire weather and fuel continuity)

Many climate models do not explicitly simulate fire processes. As a result, assessing fire risk and its impacts often relies on Fire Danger or Fire Weather Indices, which use meteorological information derived from climate models. These indices are traditionally based on daily or sub-daily data. This diagnostic constructs testable fire weather and fire fuel indices specifically designed for evaluating monthly climate model output. To achieve this, we use a Bayesian inference framework, which represents and optimizes different controls of burnt area, to generate fire weather and fire vegetation indicators from monthly observed or reanalysis meteorological and vegetation data. Assessment of meteorological and biological conditions that drive fire would aid the development of these models. Developing fire weather indices that directly target burnt area, rather than relying solely on traditional fire danger formulations, presents a more robust alternative for climate model assessment.

Author contributions. The Model Benchmarking Task Team co-leads conceptualized the structure of the paper and led the writing of the draft. Authors responsible for the Rapid Evaluation Framework conceptualization: FH and BH. Funding acquisition: FH, BH, BT, and EOR. Methodology: FH and BH. Reference dataset coordination: DH, RS. Data provenance and citation protocol and coding: M. Stockhause, DH, J. Lewis, and BA. Project management: BT, EOR, and ED. Writing and original draft preparation: All authors. Coding and implementation of the REF: J. Lewis, MP, BA, NC, J. Lee, and MX. Development and revision of diagnostics: NC, M. Sreeush, and SB. Writing, review, and editing: All authors.

Competing interests. At least one of the (co-)authors is a member of the editorial board of Geoscientific Model Development. The authors have no other competing interests to declare.

Acknowledgements. The work developing the concept of the Rapid Evaluation Framework was coordinated and led by Forrest Hoffman (ORNL) and Birgit Hassler (DLR), co-leads of the Model Benchmarking Task Team. The work overseeing delivery of the Assessment Fast Track Rapid Evaluation Framework is being coordinated by Forrest Hoffman (ORNL), Birgit Hassler (DLR) and Ranjini Swaminathan (NCEO, UoR) as co-leads of the Model Benchmarking Task Team. Birgit Hassler (DLR) led the final selection process of the diagnostics required for the Assessment Fast Track. The REF implementation team for the Assessment Fast Track REF is led by Jared Lewis (Climate Resource) and is supported by expertise from the following software packages and organizations: Climate Resource, eScience Center, ESMValTool, ILAMB, IOMB, PMP, and CMEC. We wish to acknowledge and thank all the diagnostic recipe and code developers within those packages. The two ESGF nodes committed to hosting, indexing and replication of the REF are ORNL and CEDA. CEDA is also archiving the Beta reference dataset collection. Quality assurance of reference datasets has been conducted through dataset proposal review by the following members of the obs4MIPs Steering Panel; Peter Gleckler, Simon Pinnock, Greg Elsaesser, Alison Waterfall, Birgit Hassler, and Kate Willett. We also acknowledge the ESGF architecture developers who have directly worked with the REF delivery team. This includes Sasha Ames, LLNL, Lee Liming, Globus, Steve Turoscy, Globus and Dave Poulter CEDA. The project administration is provided by the CMIP IPO, which is hosted by the European Space Agency, with staff provided on contract by HE Space Operations Ltd. All the figures have been produced by and/or commissioned by the CMIP International Project Office and are under a Creative Commons Attribution 4.0 International license. We gratefully acknowledge the valuable feedback provided during development of the REF by Maureen Wanzala (WCRP), ESMO Scientific Steering Group and International Project Office, WGCM Infrastructure Panel, CMIP Panel, obs4MIPs Steering Panel, pre-Alpha tester modeling centers; UK Met Office, ISAC-CNR, CCMC, and AS-RCEC as well as all the participants that participated in the stakeholder surveys, drop-ins and the hackathon. ESMValTool diagnostic development was performed using resources of the Deutsches Klimarechenzentrum (DKRZ) granted by its Scientific Steering Committee (WLA) under project no. bd0854.

Financial support. The work developing the Assessment Fast Track Rapid Evaluation Framework has been made possible by funding from the European Space Agency and the US Department of Energy. The work of Forrest M. Hoffman, Nathan Collier, and Min Xu was supported by the Reducing Uncertainties in Biogeochemical Interactions through Synthesis and Computation (RUBISCO) Science Focus Area, which is sponsored by the Biological and Environmental Research (BER) program in the US Department of Energy Office of Science. The Earth

System Grid Federation (ESGF) is an international consortium of individually funded data provider institutions; the ESGF2-US Project in the United States of America is sponsored by the Data Management Program in BER in the US Department of Energy Office of Science, and the ESGF activity in the United Kingdom is supported by the Centre for Environmental Data Analysis (CEDA), which is sponsored by the Science and Technology Facilities Council (STFC) and the National Environment Research Council (NERC). Oak Ridge National Laboratory (ORNL) is managed by UT-Battelle, LLC, for the US Department of Energy under Contract No. DE-AC05-00OR22725. The work of Birgit Hassler, Lisa Bock, and Manuel Schlund was supported by the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 101003536 (ESM2025 – Earth System Models for the Future). The work of Ranjini Swaminathan was supported by UKRI-NERC TerraFIRMA (NE/W004895/1) and the AI4PEX project (the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee under grant agreement number 10114295, 10103109, and 10093450). The work of Jiwoo Lee and Paul Ullrich was supported by the Program for Climate Model Diagnosis and Intercomparison (PCMDI), which is sponsored by the BER program. Lawrence Livermore National Laboratory (LLNL) is managed by Lawrence Livermore National Security, LLC (LLNS), for the US Department of Energy's National Nuclear Security Administration (NNSA) under contract number DE-AC52-07NA27344. The work of Lisa Bock and Axel Lauer (AL) was supported by the ESA Climate Change Initiative Climate Model User Group (ESA CCI CMUG) under contract 4000125156/18/I-NB. The work of Bettina K. Gier and Katja Weigel was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through the Gottfried Wilhelm Leibniz Prize awarded to Veronika Eyring (Reference number EY 22/2-1) and the project S1 ("Diagnosis and Metrics in Climate Models") of the Collaborative Research Centre TRR 181 "Energy Transfers in Atmosphere and Ocean" (Grant No. 274762653). Douglas I. Kelley was supported by the Natural Environment Research Council as part of the LTSM2 TerraFIRMA project (NE/W004895/1). Axel Lauer also received support by the ESA CCI Ozone project (Ozone_cci phase 3) under contract number 4000126562/19/I-NB. Julien Lenhardt's work was supported by European Union's Horizon 2020 research and innovation action programme under Grant Agreement No. 101003536 (ESM2025 – Earth System Models for the Future) and Grant Agreement No. 101137682 (AI4PEX – Artificial Intelligence and Machine Learning for Enhanced Representation of Processes and Extremes in Earth System Models). The work of Manuel Schlund was also supported by the BMBF under CAP7 project, Grant Agreement No. 01LP2401C. The work of Mohanan G. Sreeush was supported by the European Union's Horizon Europe research and innovation program under Grant 101083922 (OceanICU Improving Carbon Understanding) and Alfred Wegener Institute (AWI) PROCEED short term Research Grant 2025. The work reflects only the authors' views; the European Commission and their executive agency are not responsible for any use that may be made. Ed Blockley and Helene Hewitt were supported by the Met Office Hadley Centre Climate Programme funded by the Department for Science, Innovation and Technology. The work of Rebecca L. Beadling was funded by NSF Division of Polar Programs Grant NSF2319828 and NOAA award NA24OARX431C0057-T1-01. Valerio Lembo (VL) received funding from the Italian Ministry of Education, University and Research (MIUR) (project MEDHEX "Mediterranean Heat budget and EXtreme transports: From the heat budget to extreme events in the Mediterranean region", CUP B53C24007120006), and from the European Union's Horizon Europe research and innovation programme (OptimESM, Grant No. 101081193). The work of Jianhua Lu was supported by National Natural Science Foundation of China (Grant No. 42442507). S.B. was supported by the project ATLAS2 (PID2024-162703OB-I00) funded by MCIN/AEI/10.13039/501100011033 and by ERDF/EU. Jerry Tjiputra was supported by the Research Council of Norway funded projects INES2 (350390), NorESM4CMIP7 (352204), and NAVIGATE (352142). The work of Brian Medeiros was supported by the RGMA under Award Number DE-SC0022070 and National Science Foundation (NSF) IA 1947282; this work was also supported by the National Center for Atmospheric Research (NCAR), which is a major facility sponsored by the NSF under Cooperative Agreement No. 1852977. Enrico Scoccimarro was supported by the EU project Artemis under grant agreement No. 101225852. The work of Jeremy Walton was supported by the Natural Environment Research Council under the TerraFIRMA project (Grant reference NE/W004895/1) and the Met Office Hadley Centre Climate Programme funded by the

Department for Science, Innovation and Technology. Andre L. Marquez was supported by the project “Development of the Community Earth System Model – MONAN”, under the contract No. 01340.005344/2021-50 funded by Brazilian National Foundation for Science and Technology Development(FNDCT). Malcolm J. Roberts was supported by UK Research and Innovation (UKRI) under the UK government’s Horizon Europe funding guarantee (grant no. 10057890) as a part of EU EERIE.

References

Adler, R., Wang, J.-J., Sapiano, M., Huffman, G., Chiu, L., Xie, P.-P., Ferraro, R., Schneider, U., Becker, A., Bolvin, D., Nelkin, E., Gu, G., and NOAA CDR Program: Global Precipitation Climatology Project (GPCP) Climate Data Record (CDR), Version 2.3 (Monthly), https://doi.org/10.7289/V56971M6, 2017.

Adler, R. F., Huffman, G. J., Chang, A., Ferraro, R., Xie, P.-P., Janowiak, J., Rudolf, B., Schneider, U., Curtis, S., Bolvin, D., Gruber, A., Susskind, J., Arkin, P., and Nelkin, E.: The Version-2 Global Precipitation Climatology Project (GPCP) Monthly Precipitation Analysis (1979–Present), Journal of Hydrometeorology, 4, 1147–1167, https://doi.org/10.1175/1525-7541(2003)004<1147:TVGPCP>2.0.CO;2, 2003.

Adler, R. F., Sapiano, M. R. P., Huffman, G. J., Wang, J.-J., Gu, G., Bolvin, D., Chiu, L., Schneider, U., Becker, A., Nelkin, E., Xie, P., Ferraro, R., and Shin, D.-B.: The Global Precipitation Climatology Project (GPCP) Monthly Analysis (New Version 2.3) and a Review of 2017 Global Precipitation, Atmosphere, 9, 138, https://doi.org/10.3390/atmos9040138, 2018.

Agarwal, D., Ayliffe, J., Buck, J. J. H., Damerow, J., Parton, G., Stall, S., Stockhause, M., and Wyborn, L.: Complex Citation Working Group Recommendation, https://doi.org/10.15497/RDA00130, report produced by the Research Data Alliance (RDA), 2025.

Alemohammad, S. H., Fang, B., Konings, A. G., Aires, F., Green, J. K., Kolassa, J., Miralles, D., Prigent, C., and Gentine, P.: Water, Energy, and Carbon with Artificial Neural Networks (WECANN): A Statistically Based Estimate of Global Surface Turbulent Fluxes and Gross Primary Productivity Using Solar-induced Fluorescence, Biogeosciences, 14, 4101–4124, https://doi.org/10.5194/bg-14-4101-2017, 2017.

Anav, A., Friedlingstein, P., Beer, C., Ciais, P., Harper, A., Jones, C., Murray-Tortarolo, G., Papale, D., Parazoo, N. C., Peylin, P., Piao, S., Sitch, S., Viovy, N., Wiltshire, A., and Zhao, M.: Spatiotemporal Patterns of Terrestrial Gross Primary Production: A Review, Reviews of Geophysics, 53, 785–818, https://doi.org/https://doi.org/10.1002/2015RG000483, 2015.

Andela, B., Broetz, B., de Mora, L., Drost, N., Eyring, V., Koldunov, N., Lauer, A., Mueller, B., Predoi, V., Righi, M., Schlund, M., Vegas-Regidor, J., Zimmermann, K., Adeniyi, K., Castellani, G., Arnone, E., Bellprat, O., Berg, P., Billows, C., Blockley, E., Bock, L., Bodas-Salcedo, A., Caron, L.-P., Carvalhais, N., Cionni, I., Cortesi, N., Corti, S., Crezee, B., Davin, E. L., Davini, P., Deser, C., Diblen, F., Docquier, D., Dreyer, L., Ehbrecht, C., Earnshaw, P., Geddes, T., Gier, B., Gillett, E., Gonzalez-Reviriego, N., Goodman, P., Hagemann, S., Hall, S., Hardacre, C., von Hardenberg, J., Hassler, B., Heuer, H., Hogan, E., Hunter, A., Kadow, C., Kindermann, S., Koirala, S., Kuehbacher, B., Lledó, L., Lejeune, Q., Lembo, V., Little, B., Loosveldt-Tomas, S., Lorenz, R., Lovato, T., Lucarini, V., Malinina, E., Massonnet, F., Mohr, C. W., Amarjiit, P., Parsons, N., Pérez-Zanón, N., Phillips, A., Proft, M., Russell, J., Sandstad, M., Sellar, A., Senftleben, D., Serva, F., Sillmann, J., Stacke, T., Storkey, D., Swaminathan, R., Tomkins, K., Torralba, V., Weigel, K., Sarauer, E., Schulze, K., Roberts, C., Kalverla, P., Alidoost, S., Verhoeven, S., Vreede, B., Smeets, S., Soares Siqueira, A., Kazeroni, R., Potter, J., Winterstein, F., Beucher, R., Kraft, J., Ruhe, L., Bonnet, P., Munday, G., Chun, F., and Ellis, H.: ESMValTool, https://doi.org/10.5281/zenodo.18998400, 2025.

Arora, V. K., Katavouta, A., Williams, R. G., Jones, C. D., Brovkin, V., Friedlingstein, P., Schwinger, J., Bopp, L., Boucher, O., Cadule, P., Chamberlain, M. A., Christian, J. R., Delire, C., Fisher, R. A., Hajima, T., Ilyina, T., Joetzjer, E., Kawamiya, M., Koven, C. D., Krasting, J. P., Law, R. M., Lawrence, D. M., Lenton, A., Lindsay, K., Pongratz, J., Raddatz, T., Séférian, R., Tachiiri, K., Tjiputra, J. F., Wiltshire, A., Wu, T., and Ziehn, T.: Carbon–concentration and Carbon–climate Feedbacks in CMIP6 Models and Their Comparison to CMIP5 Models, Biogeosciences, 17, 4173–4222, https://doi.org/10.5194/bg-17-4173-2020, publisher: Copernicus GmbH, 2020.

Bacer, S., Jomaa, F., Beaumet, J., Gallée, H., Le Bouëdec, E., Ménégoz, M., and Staquet, C.: Impact of Climate Change on Wintertime European Atmospheric Blocking, Weather and Climate Dynamics, 3, 377–389, https://doi.org/10.5194/wcd-3-377-2022, 2022.

Baker, A. H., Hammerling, D. M., Levy, M. N., Xu, H., Dennis, J. M., Eaton, B. E., Edwards, J., Hannay, C., Mickelson, S. A., Neale, R. B., Nychka, D., Shollenberger, J., Tribbia, J., Vertenstein, M., and Williamson, D.: A new ensemble-based consistency test for the Community Earth System Model (pyCECT v1.0), Geoscientific Model Development, 8, 2829–2840, https://doi.org/10.5194/gmd-8-2829-2015, 2015.

Barnett, T. P., Dümenil, L., Schlese, U., Roeckner, E., and Latif, M.: The Effect of Eurasian Snow Cover on Regional and Global Climate Variations, Journal of the Atmospheric Sciences, 46, 661–686, https://doi.org/10.1175/1520-0469(1989)046<0661:TEOESC>2.0.CO;2, 1989.

Beadling, R. and Swaminathan, R.: AMOC example of REF interactive dashboard plots., https://doi.org/10.5281/zenodo.20055565, 2026.

Beadling, R. L., Swaminathan, R., Beucher, R., Blockley, E., Brands, S., Hassler, B., Hegedűs, D., Hoffman, F. M., Lee, J., Lewis, J., Lu, J., Malinina, E., Medeiros, B., Scoccimarro, E., Tjiputra, J., Turner, B., and Watson-Parris, D.: Observational Data for Next-Generation Climate Model Evaluation: Requirements, Considerations, and Best Practices, Bulletin of the American Meteorological Society, 107, E813 – E835, https://doi.org/10.1175/BAMS-D-25-0079.1, 2026.

Bokhorst, S., Pedersen, S. H., Brucker, L., Anisimov, O., Bjerke, J. W., Brown, R. D., Ehrich, D., Essery, R. L. H., Heilig, A., Ingvander, S., Johansson, C., Johansson, M., Jónsdóttir, I. S., Inga, N., Luojus, K., Macelloni, G., Mariash, H., McLennan, D., Rosqvist, G. N., Sato, A., Savela, H., Schneebeli, M., Sokolov, A., Sokratov, S. A., Terzago, S., Vikhamar-Schuler, D., Williamson, S., Qiu, Y., and Callaghan, T. V.: Changing Arctic Snow Cover: A Review of Recent Developments and Assessment of Future Needs for Observations, Modelling, and Impacts, Ambio, 45, 516–537, https://doi.org/10.1007/s13280-016-0770-0, 2016.

Boucher, O., Servonnat, J., Albright, A. L., Aumont, O., Balkanski, Y., Bastrikov, V., Bekki, S., Bonnet, R., Bony, S., Bopp, L., Braconnot, P., Brockmann, P., Cadule, P., Caubel, A., Cheruy, F., Codron, F., Cozic, A., Cugnet, D., D'Andrea, F., Davini, P., de Lavergne, C., Denvil, S., Deshayes, J., Devilliers, M., Ducharne, A., Dufresne, J.-L., Dupont, E., Éthé, C., Fairhead, L., Falletti, L., Flavoni, S., Foujols, M.-A., Gardoll, S., Gastineau, G., Ghattas, J., Grandpeix, J.-Y., Guenet, B., Guez, Lionel, E., Guilyardi, E., Guimberteau, M., Hauglustaine, D., Hourdin, F., Idelkadi, A., Joussaume, S., Kageyama, M., Khodri, M., Krinner, G., Lebas, N., Levavasseur, G., Lévy, C., Li, L., Lott, F., Lurton, T., Luyssaert, S., Madec, G., Madeleine, J.-B., Maignan, F., Marchand, M., Marti, O., Mellul, L., Meurdesoif, Y., Mignot, J., Musat, I., Ottlé, C., Peylin, P., Planton, Y., Polcher, J., Rio, C., Rochetin, N., Rousset, C., Sepulchre, P., Sima, A., Swingedouw, D., Thiéblemont, R., Traore, A. K., Vancoppenolle, M., Vial, J., Vialard, J., Viovy, N., and Vuichard, N.: Presentation and Evaluation of the IPSL-CM6A-LR Climate Model, Journal of Advances in Modeling Earth Systems, 12, e2019MS002 010, https://doi.org/10.1029/2019MS002010, 2020.

Brands, S.: A Circulation-based Performance Atlas of the CMIP5 and 6 Models for Regional Climate Studies in the Northern Hemisphere Mid-to-high Latitudes, Geoscientific Model Development, 15, 1375–1411, https://doi.org/10.5194/gmd-15-1375-2022, 2022.

Brands, S.: pyLamb - A Climate Model Verification Tool Based on Lamb Weather Types, https://doi.org/10.5281/zenodo.15346363, 2025.

Brands, S., Fernández-Granja, J. A., Bedia, J., Casanueva, A., and Fernández, J.: A Global Climate Model Performance Atlas for the Southern Hemisphere Extratropics Based on Regional Atmospheric Circulation Patterns, Geophysical Research Letters, 50, e2023GL103 531, https://doi.org/10.1029/2023GL103531, 2023.

Burrows, S. M., Maltrud, M., Yang, X., Zhu, Q., Jeffery, N., Shi, X., Ricciuto, D. M., Wang, S., Bisht, G., Tang, J., Wolfe, J., Harrop, B. E., Singh, B., Brent, L., Baldwin, S., Zhou, T., Cameron-Smith, P., Keen, N., Collier, N., Xu, M., Hunke, E. C., Elliott, S. M., Turner, A. K., Li, H.-Y., Wang, H., Golaz, J.-C., Bond-Lamberty, B., Hoffman, F. M., Riley, W. J., Thornton, P. E., Calvin, K., and Leung, L. R.: The DOE E3SM v1.1 Biogeochemistry Configuration: Description and Simulated Ecosystem-Climate Responses to Historical Changes in Forcing, Journal of Advances in Modeling Earth Systems, 12, e2019MS001 766, https://doi.org/10.1029/2019MS001766, 2020.

Chen, X. and Wallace, J. M.: ENSO-Like Variability: 1900–2013, Journal of Climate, 28, 9623–9641, https://doi.org/10.1175/JCLI-D-15-0322.1, 2015.

Chen, Y., Hall, J., Van Wees, D., Andela, N., Hantson, S., Giglio, L., Van Der Werf, G. R., Morton, D. C., and Randerson, J. T.: Global Fire Emissions Database (GFED5) Burned Area, https://doi.org/10.5281/ZENODO.7668424, data set, 2023a.

Chen, Y., Hall, J., van Wees, D., Andela, N., Hantson, S., Giglio, L., van der Werf, G. R., Morton, D. C., and Randerson, J. T.: Multi-decadal Trends and Variability in Burned Area from the Fifth Version of the Global Fire Emissions Database (GFED5), Earth System Science Data, 15, 5227–5259, https://doi.org/10.5194/essd-15-5227-2023, 2023b.

Cheng, L., Pan, Y., Tan, Z., Zheng, H., Zhu, Y., Wei, W., Du, J., Yuan, H., Li, G., Ye, H., Gouretski, V., Li, Y., Trenberth, K. E., Abraham, J., Jin, Y., Reseghetti, F., Lin, X., Zhang, B., Chen, G., Mann, M. E., and Zhu, J.: IAPv4 Ocean Temperature and Ocean Heat Content Gridded Dataset, Earth System Science Data, 16, 3517–3546, https://doi.org/10.5194/essd-16-3517-2024, 2024.

Cinquini, L., Crichton, D., Mattmann, C., Harney, J., Shipman, G., Wang, F., Ananthakrishnan, R., Miller, N., Denvil, S., Morgan, M., Pobre, Z., Bell, G. M., Doutriaux, C., Drach, R., Williams, D., Kershaw, P., Pascoe, S., Gonzalez, E., Fiore, S., and Schweitzer, R.: The Earth System Grid Federation: An Open Infrastructure for Access to Distributed Geospatial Data, Future Generation Computer Systems, 36, 400–417, https://doi.org/10.1016/j.future.2013.07.002, 2014.

Claverie, M., Vermote, E., and NOAA CDR Program: NOAA Climate Data Record (CDR) of Leaf Area Index (LAI) and Fraction of Absorbed Photosynthetically Active Radiation (FAPAR), Version 4 (Version Superseded), https://doi.org/10.7289/V5M043BX, published by NOAA National Centers for Environmental Information, 2014.

Claverie, M., Vermote, E., Justice, C., Csiszar, I., Myneni, R., Baret, F., Masuoka, E., Wolfe, R., Ray, J. P., and NOAA CDR Program: NOAA Climate Data Record (CDR) of VIIRS Leaf Area Index (LAI) and Fraction of Absorbed Photosynthetically Active Radiation (FAPAR), Version 1, https://doi.org/10.25921/9X3M-0E02, published by NOAA National Centers for Environmental Information, 2024.

CMIP Model Benchmarking Task Team: CMIP7 Assessment Fast Track Diagnostics List for the Rapid Evaluation Framework, https://doi.org/10.5281/zenodo.14284374, 2024.

Cohen, J. and Entekhabi, D.: Eurasian Snow Cover Variability and Northern Hemisphere Climate Predictability, Geophysical Research Letters, 26, 345–348, https://doi.org/10.1029/1998GL900321, 1999.

Cohen, J., Screen, J. A., Furtado, J. C., Barlow, M., Whittleston, D., Coumou, D., Francis, J., Dethloff, K., Entekhabi, D., Overland, J., and Jones, J.: Recent Arctic Amplification and Extreme Mid-latitude Weather, Nature Geoscience, 7, 627–637, https://doi.org/10.1038/ngeo2234, 2014.

Coldewey-Egbers, M., Loyola, D. G., Koukouli, M., Balis, D., Lambert, J.-C., Verhoelst, T., Granville, J., van Roozendael, M., Lerot, C., Spurr, R., Frith, S. M., and Zehner, C.: The GOME-type Total Ozone Essential Climate Variable (GTO-ECV) Data Record from the ESA Climate Change Initiative, Atmospheric Measurement Techniques, 8, 3923–3940, 10.5194/amt-8-3923-2015, 2015.

Coldewey-Egbers, M., Loyola, D. G., Latter, B., Siddans, R., Kerridge, B., Hubert, D., van Roozendael, M., and Eisinger, M.: The novel GOME-type Ozone Profile Essential Climate Variable (GOP-ECV) Data Record Covering the Past 26 Years, Atmospheric Measurement Techniques Discussions, 2025, 1–31, https://doi.org/10.5194/amt-2024-196, 2025.

Collier, N., Hoffman, F. M., Lawrence, D. M., Keppel-Aleks, G., Koven, C. D., Riley, W. J., Mu, M., and Randerson, J. T.: The International Land Model Benchmarking (ILAMB) System: Design, Theory, and Implementation, Journal of Advances in Modeling Earth Systems, 10, 2731–2754, https://doi.org/10.1029/2018MS001354, 2018.

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., Gleason, B. E., Vose, R. S., Rutledge, G., Bessemoulin, P., Brönnimann, S., Brunet, M., Crouthamel, R. I., Grant, A. N., Groisman, P. Y., Jones, P. D., Kruk, M. C., Kruger, A. C., Marshall, G. J., Maugeri, M., Mok, H. Y., Nordli, Ø., Ross, T. F., Trigo, R. M., Wang, X. L., Woodruff, S. D., and Worley, S. J.: The Twentieth

Century Reanalysis Project, Quarterly Journal of the Royal Meteorological Society, 137, 1–28, https://doi.org/10.1002/qj.776, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qj.776, 2011.

Copernicus Climate Data Store: Ozone Monthly Gridded Data from 1970 to Present Derived from Satellite Observations, https://doi.org/10.24381/CDS.4EBFE4EB, published by European Center for Medium-range Weather Forecast, 2020.

Danabasoglu, G., Lamarque, J.-F., Bacmeister, J., Bailey, D. A., DuVivier, A. K., Edwards, J., Emmons, L. K., Fasullo, J., Garcia, R., Gettelman, A., Hannay, C., Holland, M. M., Large, W. G., Lauritzen, P. H., Lawrence, D. M., Lenaerts, J. T. M., Lindsay, K., Lipscomb, W. H., Mills, M. J., Neale, R., Oleson, K. W., Otto-Bliesner, B., Phillips, A. S., Sacks, W., Tilmes, S., van Kampenhout, L., Vertenstein, M., Bertini, A., Dennis, J., Deser, C., Fischer, C., Fox-Kemper, B., Kay, J. E., Kinnison, D., Kushner, P. J., Larson, V. E., Long, M. C., Mickelson, S., Moore, J. K., Nienhouse, E., Polvani, L., Rasch, P. J., and Strand, W. G.: The Community Earth System Model Version 2 (CESM2), Journal of Advances in Modeling Earth Systems, 12, e2019MS001 916, https://doi.org/10.1029/2019MS001916, 2020.

Davini, P.: MiLES - Mid Latitude Evaluation System, https://doi.org/10.5281/zenodo.2578139, 2019.

Davini, P. and D'Andrea, F.: Northern Hemisphere Atmospheric Blocking Representation in Global Climate Models: Twenty Years of Improvements?, Journal of Climate, 29, 8823–8840, https://doi.org/10.1175/JCLI-D-16-0242.1, 2016.

Davis, E., Taylor, K. E., Adloff, F., Gregory, J., Lawrence, B., and Lee, D.: Supporting Open Science with the CF Metadata Conventions for NetCDF, https://doi.org/10.5281/zenodo.15015065, slides for the presentation given on 11 December 2024 at the AGU Annual Meeting to the session for the AGU Open Science Recognition Prize, 2024.

Deser, C., Phillips, A., Bourdette, V., and Teng, H.: Uncertainty in Climate Change Projections: The Role of Internal Variability, Climate Dynamics, 38, 527–546, https://doi.org/10.1007/s00382-010-0977-x, 2012.

Deser, C., Lehner, F., Rodgers, K. B., Ault, T., Delworth, T. L., DiNezio, P. N., Fiore, A., Frankignoul, C., Fyfe, J. C., Horton, D. E., Kay, J. E., Knutti, R., Lovenduski, N. S., Marotzke, J., McKinnon, K. A., Minobe, S., Randerson, J., Screen, J. A., Simpson, I. R., and Ting, M.: Insights from Earth System Model Initial-condition Large Ensembles and Future Prospects, Nature Climate Change, 10, 277–286, https://doi.org/10.1038/s41558-020-0731-2, 2020.

Diez-Sierra, J., Iturbide, M., Gutiérrez, J. M., Fernández, J., Milovac, J., Cofiño, A. S., Cimadevilla, E., Nikulin, G., Levavasseur, G., Kjellström, E., Bülow, K., Horányi, A., Brookshaw, A., García-Díez, M., Pérez, A., Baño-Medina, J., Ahrens, B., Alias, A., Ashfaq, M., Bukovsky, M., Buonomo, E., Caluwaerts, S., Chou, S. C., Christensen, O. B., Ciarlò, J. M., Coppola, E., Corre, L., Demory, M.-E., Djurdjevic, V., Evans, J. P., Fealy, R., Feldmann, H., Jacob, D., Jayanarayanan, S., Katzfey, J., Keuler, K., Kittel, C., Kurnaz, M. L., Laprise, R., Lionello, P., McGinnis, S., Mercogliano, P., Nabat, P., Önol, B., Ozturk, T., Panitz, H.-J., Paquin, D., Pieczka, I., Raffaele, F., Remedio, A. R., Scinocca, J., Sevault, F., Somot, S., Steger, C., Tangang, F., Teichmann, C., Termonia, P., Thatcher, M., Torma, C., Meijgaard, E. v., Vautard, R., Warrach-Sagi, K., Winger, K., and Zittis, G.: The Worldwide C3S CORDEX Grand Ensemble: A Major Contribution to Assess Regional Climate Change in the IPCC AR6 Atlas, Bulletin of the American Meteorological Society, 103, E2804–E2826, https://doi.org/10.1175/BAMS-D-22-0111.1, 2022.

Dingley, B., Anstey, J. A., Abalos, M., Abraham, C., Bergman, T., Bock, L., Fiddes, S., Hassler, B., Kramer, R. J., Luo, F., O'Connor, F. M., Šácha, P., Simpson, I. R., Wilcox, L. J., and Zelinka, M. D.: CMIP7 Data Request: atmosphere priorities and opportunities, Geoscientific Model Development, 19, 2945–2984, https://doi.org/10.5194/gmd-19-2945-2026, 2026.

Dolores-Tesillos, E., Martius, O., and Quinting, J.: On the Role of Moist and Dry Processes in Atmospheric Blocking Biases in the Euro-Atlantic Region in CMIP6, Weather and Climate Dynamics, 6, 471–487, https://doi.org/10.5194/wcd-6-471-2025, 2025.

Dorrington, J., Strommen, K., and Fabiano, F.: Quantifying Climate Model Representation of the Wintertime Euro-Atlantic Circulation Using Geopotential-Jet Regimes, Weather and Climate Dynamics, 3, 505–533, https://doi.org/10.5194/wcd-3-505-2022, 2022.

Dunne, J. P., Hewitt, H. T., Arblaster, J. M., Bonou, F., Boucher, O., Cavazos, T., Dingley, B., Durack, P. J., Hassler, B., Juckes, M., Miyakawa, T., Mizielinski, M., Naik, V., Nicholls, Z., O'Rourke, E., Pincus, R., Sanderson, B. M., Simpson, I. R., and Taylor, K. E.: An evolving Coupled Model Intercomparison Project phase 7 (CMIP7) and Fast Track in support of future climate assessment, Geoscientific Model Development, 18, 6671–6700, https://doi.org/10.5194/gmd-18-6671-2025, 2025.

Döscher, R., Acosta, M., Alessandri, A., Anthoni, P., Arsouze, T., Bergman, T., Bernardello, R., Boussetta, S., Caron, L.-P., Carver, G., Castrillo, M., Catalano, F., Cvijanovic, I., Davini, P., Dekker, E., Doblas-Reyes, F. J., Docquier, D., Echevarria, P., Fladrich, U., Fuentes-Franco, R., Gröger, M., v. Hardenberg, J., Hieronymus, J., Karami, M. P., Keskinen, J.-P., Koenigk, T., Makkonen, R., Massonnet, F., Ménégoz, M., Miller, P. A., Moreno-Chamarro, E., Nieradzik, L., van Noije, T., Nolan, P., O'Donnell, D., Ollinaho, P., van den Oord, G., Ortega, P., Prims, O. T., Ramos, A., Reerink, T., Rousset, C., Ruprich-Robert, Y., Le Sager, P., Schmith, T., Schrödner, R., Serva, F., Sicardi, V., Sloth Madsen, M., Smith, B., Tian, T., Tourigny, E., Uotila, P., Vancoppenolle, M., Wang, S., Wårlind, D., Willén, U., Wyser, K., Yang, S., Yepes-Arbós, X., and Zhang, Q.: The EC-Earth3 Earth system model for the Coupled Model Intercomparison Project 6, Geoscientific Model Development, 15, 2973–3020, https://doi.org/10.5194/gmd-15-2973-2022, 2022.

Eaton, B., Gregory, J., Drach, B., Taylor, K., Hankin, S., Caron, J., Signell, R., Bentley, P., Rappa, G., Höck, H., Pamment, A., Juckes, M., Raspaud, M., Blower, J., Horne, R., Whiteaker, T., Blodgett, D., Zender, C., Lee, D., Hassell, D., Snow, A. D., Kölling, T., Allured, D., Jelenak, A., Soerensen, A. M., Gaultier, L., Herlédan, S., Manzano, F., Bärring, L., Barker, C., and Bartholomew, S. L.: NetCDF Climate and Forecast (CF) Metadata Conventions, https://doi.org/10.5281/zenodo.14275599, report produced by the CF Community, 2024.

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

Eyring, V., Cox, P. M., Flato, G. M., Gleckler, P. J., Abramowitz, G., Caldwell, P., Collins, W. D., Gier, B. K., Hall, A. D., Hoffman, F. M., Hurtt, G. C., Jahn, A., Jones, C. D., Klein, S. A., Krasting, J. P., Kwiatkowski, L., Lorenz, R., Maloney, E., Meehl, G. A., Pendergrass, A. G., Pincus, R., Ruane, A. C., Russell, J. L., Sanderson, B. M., Santer, B. D., Sherwood, S. C., Simpson, I. R., Stouffer, R. J., and Williamson, M. S.: Taking Climate Model Evaluation to the Next Level, Nature Climate Change, 9, 102–110, https://doi.org/10.1038/s41558-018-0355-y, 2019.

Eyring, V., Bock, L., Lauer, A., Righi, M., Schlund, M., Andela, B., Arnone, E., Bellprat, O., Brötz, B., Caron, L.-P., Carvalhais, N., Cionni, I., Cortesi, N., Crezee, B., Davin, E. L., Davini, P., Debeire, K., de Mora, L., Deser, C., Docquier, D., Earnshaw, P., Ehbrecht, C., Gier, B. K., Gonzalez-Reviriego, N., Goodman, P., Hagemann, S., Hardiman, S., Hassler, B., Hunter, A., Kadow, C., Kindermann, S., Koirala, S., Koldunov, N., Lejeune, Q., Lembo, V., Lovato, T., Lucarini, V., Massonnet, F., Müller, B., Pandde, A., Pérez-Zanón, N., Phillips, A., Predoi, V., Russell, J., Sellar, A., Serva, F., Stacke, T., Swaminathan, R., Torralba, V., Vegas-Regidor, J., von Hardenberg, J., Weigel, K., and Zimmermann, K.: 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, 3383–3438, https://doi.org/10.5194/gmd-13-3383-2020, 2020.

Fang, H., Baret, F., Plummer, S., and Schaepman-Strub, G.: An Overview of Global Leaf Area Index (LAI): Methods, Products, Validation, and Applications, Reviews of Geophysics, 57, 739–799, https://doi.org/https://doi.org/10.1029/2018RG000608, 2019.

FAO and IIASA: Harmonized World Soil Database Version 2.0, Food and Agriculture Organization (FAO) and International Institute for Applied Systems Analysis (IIASA), ISBN 978-92-5-137499-3, https://doi.org/10.4060/cc3823en, 2023.

Fasullo, J. T., Phillips, A. S., and Deser, C.: Evaluation of Leading Modes of Climate Variability in the CMIP Archives, Journal of Climate, 33, 5527–5545, https://doi.org/10.1175/JCLI-D-19-1024.1, 2020.

Fernández-Granja, J. A., Bedia, J., Casanueva, A., Brands, S., and Fernández, J.: The Signature of the Main Modes of Climatic Variability as Revealed by the Jenkinson-Collison Classification over Europe, International Journal of Climatology, 44, 4076–4088, https://doi.org/10.1002/joc.8569, 2024.

Ferraro, R., Waliser, D. E., Gleckler, P., Taylor, K. E., and Eyring, V.: Evolving Obs4MIPs to Support Phase 6 of the Coupled Model Intercomparison Project (CMIP6), Bulletin of the American Meteorological Society, 96, ES131–ES133, https://doi.org/10.1175/BAMS-D-14-00216.1, 2015.

Fu, W., Moore, J. K., Primeau, F., Collier, N., Ogunro, O. O., Hoffman, F. M., and Randerson, J. T.: 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, e2022JC018 965, https://doi.org/10.1029/2022JC018965, 2022.

Giorgi, F. and Francisco, R.: Uncertainties in Regional Climate Change Prediction: A Regional Analysis of Ensemble Simulations with the HADCM2 Coupled AOGCM, Climate Dynamics, 16, 169–182, https://doi.org/10.1007/PL00013733, 2000.

Gleckler, P., Ferraro, R., and Waliser, D.: Improving Use of Satellite Data in Evaluating Climate Models, Eos Transactions American Geophysical Union, 92, 172–172, https://doi.org/10.1029/2011EO200005, 2011.

Gleckler, P., Taylor, K. E., Durack, P. J., Nadeau, D., Biard, J. C., Elsaesser, G., Ferraro, R., Finkensieper, S., Hassler, B., Manaster, A., Mears, C., Pinnock, S., Stevens, S., Tuma, M., Turner, B., Waterfall, A., and Willet, K. M.: Obs4MIPs Data Specifications Version 2.5 (ODS2.5), https://doi.org/10.5281/zenodo.11500474, report produced by the Earth System Model and Observations (ESMO), 2024.

Gleckler, P. J., Taylor, K. E., and Doutriaux, C.: Performance Metrics for Climate Models, Journal of Geophysical Research: Atmospheres, 113, https://doi.org/10.1029/2007JD008972, 2008.

Gong, D. and Wang, S.: Definition of Antarctic Oscillation index, Geophysical Research Letters, 26, 459–462, https://doi.org/https://doi.org/10.1029/1999GL900003, 1999.

Grams, C. M., Beerli, R., Pfenninger, S., Staffell, I., and Wernli, H.: Balancing Europe's Wind-power Output Through Spatial Deployment Informed by Weather Regimes, Nature Climate Change, 7, 557–562, https://doi.org/10.1038/nclimate3338, 2017.

Gregory, J. M., Ingram, W. J., Palmer, M. A., Jones, G. S., Stott, P. A., Thorpe, R. B., Lowe, J. A., Johns, T. C., and Williams, K. D.: A New Method for Diagnosing Radiative Forcing and Climate Sensitivity, Geophysical Research Letters, 31, https://doi.org/10.1029/2003GL018747, 2004.

Gregory, J. M., Jones, C. D., Cadule, P., and Friedlingstein, P.: Quantifying Carbon Cycle Feedbacks, Journal of Climate, 22, 5232–5250, https://doi.org/10.1175/2009JCLI2949.1, 2009.

Hannachi, A., Finke, K., and Trendafilov, N.: Common EOFs: A Tool for Multi-model Comparison and Evaluation, Climate Dynamics, 60, 1689–1703, https://doi.org/10.1007/s00382-022-06409-8, 2023.

Harris, I., Osborn, T. J., Jones, P., and Lister, D.: Version 4 of the CRU TS Monthly High-resolution Gridded Multivariate Climate Dataset, Scientific Data, 7, 109, https://doi.org/10.1038/s41597-020-0453-3, 2020.

Hassler, B., Hoffman, F. M., Beadling, R., Blockley, E., Huang, B., Lee, J., Lembo, V., Lewis, J., Lu, J., Madaus, L., Malinina, E., Medeiros, B., Pokam, W., Scoccimarro, E., and Swaminathan, R.: Systematic Benchmarking of Climate Models: Methodologies, Applications, and New Directions, Reviews of Geophysics, 64, e2025RG000 891, https://doi.org/https://doi.org/10.1029/2025RG000891, e2025RG000891 2025RG000891, 2026a.

Hassler, B., Lewis, J., Lee, J., Andela, B., Swaminathan, R., Hoffman, F. M., Turner, B., and O'Rourke, E.: Climate - Rapid Evaluation Framework CMIP7 Quality Checklist, https://doi.org/10.5281/zenodo.18915234, 2026b.

Hawkins, E. and Sutton, R.: Connecting Climate Model Projections of Global Temperature Change with the Real World, Bulletin of the American Meteorological Society, 97, 963–980, https://doi.org/10.1175/BAMS-D-14-00154.1, 2016.

Hegedűs, D., Swaminathan, R., and Woolliams, E.: Climate - Rapid Evaluation Framework Uncertainty Checklist, https://doi.org/10.5281/zenodo.17312883, 2025.

Hellinger, E.: Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen, Journal für die reine und angewandte Mathematik, 1909, 210–271, https://doi.org/10.1515/crll.1909.136.210, 1909.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 Global Reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Hobeichi, S., Abramowitz, G., Evans, J., and Beck, H. E.: Linear Optimal Runoff Aggregate (LORA): A Global Gridded Synthesis Runoff Product, Hydrology and Earth System Sciences, 23, 851–870, https://doi.org/10.5194/hess-23-851-2019, 2019.

Hoffman, F.: Gross Primary Production (GPP) example of REF interactive dashboard plots, https://doi.org/10.5281/zenodo.20162791, 2026.

Hoffman, F., Hassler, B., Model Benchmarking Task Team, and Dingley, B.: Rapid Evaluation Framework Overview, https://doi.org/10.5281/zenodo.20053946, 2026.

Hoffman, F. M., Hargrove, W. W., Erickson, D. J., and Oglesby, R. J.: Using Clustered Climate Regimes to Analyze and Compare Predictions from Fully Coupled General Circulation Models, Earth Interactions, 9, 1–27, https://doi.org/10.1175/EI110.1, 2005.

Hoffman, F. M., Randerson, J. T., Arora, V. K., Bao, Q., Cadule, P., Ji, D., Jones, C. D., Kawamiya, M., Khatiwala, S., Lindsay, K., Obata, A., Shevliakova, E., Six, K. D., Tjiputra, J. F., Volodin, E. M., and Wu, T.: Causes and Implications of Persistent Atmospheric Carbon Dioxide Biases in Earth System Models, Journal of Geophysical Research: Biogeosciences, 119, 141–162, https://doi.org/10.1002/2013JG002381, 2014.

Hoffman, F. M., Koven, C. D., Keppel-Aleks, G., Lawrence, D. M., Riley, W. J., Randerson, J. T., Ahlström, A., Abramowitz, G., Baldocchi, D. D., Best, M. J., Bond-Lamberty, B., De Kauwe, M. G., Denning, A. S., Desai, A. R., Eyring, V., Fisher, J. B., Fisher, R. A., Gleckler, P. J., Huang, M., Hugelius, G., Jain, A. K., Kiang, N. Y., Kim, H., Koster, R. D., Kumar, S. V., Li, H., Luo, Y., Mao, J., McDowell, N. G., Mishra, U., Moorcroft, P. R., Pau, G. S. H., Ricciuto, D. M., Schaefer, K., Schwalm, C. R., Serbin, S. P., Shevliakova, E., Slater, A. G., Tang, J., Williams, M., Xia, J., Xu, C., Joseph, R., and Koch, D.: International Land Model Benchmarking (ILAMB) 2016 Workshop Report, Tech. Rep. DOE/SC-0186, U.S. Department of Energy, Office of Science, Germantown, Maryland, USA, https://doi.org/10.2172/1330803, 2017.

Hurrell, J. W., Kushnir, Y., and Visbeck, M.: The North Atlantic Oscillation, Science, 291, 603–605, https://doi.org/10.1126/science.1058761, 2001.

IPCC: Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, 1st edn., ISBN 978-1-009-15789-6, https://doi.org/10.1017/9781009157896, 2023.

Irving, D. B., Wijffels, S., and Church, J. A.: Anthropogenic Aerosols, Greenhouse Gases, and the Uptake, Transport, and Storage of Excess Heat in the Climate System, Geophysical Research Letters, 46, 4894–4903, https://doi.org/https://doi.org/10.1029/2019GL082015, 2019.

Iturbide, M., Gutiérrez, J. M., Alves, L. M., Bedia, J., Cerezo-Mota, R., Cimadevilla, E., Cofiño, A. S., Di Luca, A., Faria, S. H., Gorodetskaya, I. V., Hauser, M., Herrera, S., Hennessy, K., Hewitt, H. T., Jones, R. G., Krakovska, S., Manzanas, R., Martínez-Castro, D., Narisma, G. T., Nurhati, I. S., Pinto, I., Seneviratne, S. I., van den Hurk, B., and Vera, C. S.: An Update of IPCC Climate Reference Regions for Subcontinental Analysis of Climate Model Data: Definition and Aggregated Datasets, Earth System Science Data, 12, 2959–2970, https://doi.org/10.5194/essd-12-2959-2020, 2020.

Juckes, M., Taylor, K. E., Antonio, F., Brayshaw, D., Buontempo, C., Cao, J., Durack, P. J., Kawamiya, M., Kim, H., Lovato, T., Mackallah, C., Mizielinski, M., Nuzzo, A., Stockhause, M., Visioni, D., Walton, J., Turner, B., O'Rourke, E., and Dingley, B.: Baseline Climate Variables for Earth System Modelling, Geoscientific Model Development, 18, 2639–2663, https://doi.org/10.5194/gmd-18-2639-2025, 2025.

Karlsson, K.-G., Stengel, M., Meirink, J. F., Riihelä, A., Trentmann, J., Akkermans, T., Stein, D., Devasthale, A., Eliasson, S., Johansson, E., Håkansson, N., Solodovnik, I., Benas, N., Clerbaux, N., Selbach, N., Schröder, M., and Hollmann, R.: CLARA-A3: The Third Edition of the AVHRR-based CM SAF Climate Data Record on Clouds, Radiation and Surface Albedo Covering the Period 1979 to 2023, Earth System Science Data, 15, 4901–4926, https://doi.org/10.5194/essd-15-4901-2023, 2023.

Keenan, T. F. and Williams, C. A.: The Terrestrial Carbon Sink, Annual Review of Environment and Resources, 43, 219–243, https://doi.org/10.1146/annurev-environ-102017-030204, 2018.

Kelley, D., Swaminathan, R., and Lenhardt, J.: Data for CMIP AR7 REF fire controls assessment, https://doi.org/10.5281/zenodo.14917245, 2025.

Keppel-Aleks, G., Randerson, J. T., Lindsay, K., Stephens, B. B., Moore, J. K., Doney, S. C., Thornton, P. E., Mahowald, N. M., Hoffman, F. M., Sweeney, C., Tans, P. P., Wennberg, P. O., and Wofsy, S. C.: Atmospheric Carbon Dioxide Variability in the Community Earth System Model: Evaluation and Transient Dynamics during the Twentieth and Twenty-First Centuries, Journal of Climate, 26, 4447–4475, https://doi.org/10.1175/JCLI-D-12-00589.1, 2013.

Kumar, P. B., Vialard, J., Lengaigne, M., Murty, V. S. N., and McPhaden, M. J.: TropFlux: Air-sea Fluxes for the Global Tropical Ocean - Description and Evaluation, Climate Dynamics, 38, 1521–1543, https://doi.org/10.1007/s00382-011-1115-0, 2012.

Lauer, A., Eyring, V., Bellprat, O., Bock, L., Gier, B. K., Hunter, A., Lorenz, R., Pérez-Zanón, N., Righi, M., Schlund, M., Senftleben, D., Weigel, K., and Zechlau, S.: 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, 4205–4228, https://doi.org/10.5194/gmd-13-4205-2020, 2020.

Lauer, A., Bock, L., Hassler, B., Jöckel, P., Ruhe, L., and Schlund, M.: Monitoring and Benchmarking Earth System Model Simulations with ESMValTool v2.12.0, Geoscientific Model Development, 18, 1169–1188, https://doi.org/10.5194/gmd-18-1169-2025, 2025.

Lauritzen, P. H., Kevlahan, N. K.-R., Toniazzo, T., Eldred, C., Dubos, T., Gassmann, A., Larson, V. E., Jablonowski, C., Guba, O., Shipway, B., Harrop, B. E., Lemarié, F., Tailleux, R., Herrington, A. R., Large, W., Rasch, P. J., Donahue, A. S., Wan, H., Conley, A., and Bacmeister, J. T.: Reconciling and Improving Formulations for Thermodynamics and Conservation Principles in Earth System Models (ESMs), Journal of Advances in Modeling Earth Systems, 14, e2022MS003 117, https://doi.org/10.1029/2022MS003117, 2022.

Lavergne, T., Sørensen, A. M., Kern, S., Tonboe, R., Notz, D., Aaboe, S., Bell, L., Dybkjær, G., Eastwood, S., Gabarro, C., Heygster, G., Killie, M. A., Brandt Kreiner, M., Lavelle, J., Saldo, R., Sandven, S., and Pedersen, L. T.: Version 2 of the EUMETSAT OSI SAF and ESA CCI Sea-Ice Concentration Climate Data Records, The Cryosphere, 13, 49–78, https://doi.org/10.5194/tc-13-49-2019, 2019.

Lawrence, D. M., Fisher, R. A., Koven, C. D., Oleson, K. W., Swenson, S. C., Bonan, G., Collier, N., Ghimire, B., van Kampenhout, L., Kennedy, D., Kluzek, E., Lawrence, P. J., Li, F., Li, H., Lombardozzi, D., Riley, W. J., Sacks, W. J., Shi, M., Vertenstein, M., Wieder,

W. R., Xu, C., Ali, A. A., Badger, A. M., Bisht, G., van den Broeke, M., Brunke, M. A., Burns, S. P., Buzan, J., Clark, M., Craig, A., Dahlin, K., Drewniak, B., Fisher, J. B., Flanner, M., Fox, A. M., Gentine, P., Hoffman, F. M., Keppel-Aleks, G., Knox, R., Kumar, S., Lenaerts, J., Leung, L. R., Lipscomb, W. H., Lu, Y., Pandey, A., Pelletier, J. D., Perket, J., Randerson, J. T., Ricciuto, D. M., Sanderson, B. M., Slater, A., Subin, Z. M., Tang, J., Thomas, R. Q., Val Martin, M., and Zeng, X.: The Community Land Model Version 5: Description of New Features, Benchmarking, and Impact of Forcing Uncertainty, Journal of Advances in Modeling Earth Systems, 11, 4245–4287, https://doi.org/10.1029/2018MS001583, 2019.

Le Bras, I. A.-A., Willis, J., and Fenty, I.: The Atlantic Meridional Overturning Circulation at 35°N From Deep Moorings, Floats, and Satellite Altimeter, Geophysical Research Letters, 50, e2022GL101 931, https://doi.org/https://doi.org/10.1029/2022GL101931, 2023.

Le Quéré, C., Andrew, R. M., Friedlingstein, P., Sitch, S., Hauck, J., Pongratz, J., Pickers, P. A., Korsbakken, J. I., Peters, G. P., Canadell, J. G., Arneth, A., Arora, V. K., Barbero, L., Bastos, A., Bopp, L., Chevallier, F., Chini, L. P., Ciais, P., Doney, S. C., Gkritzalis, T., Goll, D. S., Harris, I., Haverd, V., Hoffman, F. M., Hoppema, M., Houghton, R. A., Hurtt, G., Ilyina, T., Jain, A. K., Johannessen, T., Jones, C. D., Kato, E., Keeling, R. F., Goldewijk, K. K., Landschützer, P., Lefèvre, N., Lienert, S., Liu, Z., Lombardozzi, D., Metzl, N., Munro, D. R., Nabel, J. E. M. S., Nakaoka, S.-I., Neill, C., Olsen, A., Ono, T., Patra, P., Peregon, A., Peters, W., Peylin, P., Pfeil, B., Pierrot, D., Poulter, B., Rehder, G., Resplandy, L., Robertson, E., Rocher, M., Rödenbeck, C., Schuster, U., Schwinger, J., Séférian, R., Skjelvan, I., Steinhoff, T., Sutton, A., Tans, P. P., Tian, H., Tilbrook, B., Tubiello, F. N., van der Laan-Luijkx, I. T., van der Werf, G. R., Viovy, N., Walker, A. P., Wiltshire, A. J., Wright, R., Zaehle, S., and Zheng, B.: Global Carbon Budget 2018, Earth System Science Data, 10, 2141–2194, https://doi.org/10.5194/essd-10-2141-2018, 2018.

Lee, J.: CMIP6 Community Survey: Model Benchmarking and Evaluation Results, https://doi.org/10.5281/zenodo.15478482, 2024.

Lee, J.: Northern Annular Mode (NAM) example of REF interactive dashboard plots, https://doi.org/10.5281/zenodo.20162570, 2026.

Lee, J., Sperber, K. R., Gleckler, P. J., Bonfils, C. J. W., and Taylor, K. E.: Quantifying the Agreement Between Observed and Simulated Extratropical Modes of Interannual Variability, Climate Dynamics, 52, 4057–4089, https://doi.org/10.1007/s00382-018-4355-4, 2019.

Lee, J., Gleckler, P. J., Ahn, M.-S., Ordonez, A., Ullrich, P. A., Sperber, K. R., Taylor, K. E., Planton, Y. Y., Guilyardi, E., Durack, P., Bonfils, C., Zelinka, M. D., Chao, L.-W., Dong, B., Doutriaux, C., Zhang, C., Vo, T., Boutte, J., Wehner, M. F., Pendergrass, A. G., Kim, D., Xue, Z., Wittenberg, A. T., and Krasting, J.: Systematic and Objective Evaluation of Earth System Models: PCMDI Metrics Package (PMP) Version 3, Geoscientific Model Development, 17, 3919–3948, https://doi.org/10.5194/gmd-17-3919-2024, 2025.

Lembo, V., Folini, D., Wild, M., and Lionello, P.: Inter-hemispheric Differences in Energy Budgets and Cross-equatorial Transport Anomalies During the 20th Century, Climate Dynamics, 53, 115–135, https://doi.org/10.1007/s00382-018-4572-x, 2019.

Lewis, J., Andela, B., Collier, N., Lee, J., Hegedűs, D., Pflüger, M., and Xu, M.: Rapid Evaluation Framework, All versions, https://doi.org/10.5281/zenodo.15103441, 2025.

Lewis, J., Andela, B., Collier, N., Lee, J., Hegedūs, D., Pflüger, M., Xu, M., Hoffman, F., Swaminathan, R., Hassler, B., and Price, Z.: Rapid Evaluation Framework, https://doi.org/10.5281/zenodo.15103441, 2026.

Loeb, N. G., Doelling, D. R., Wang, H., Su, W., Nguyen, C., Corbett, J. G., Liang, L., Mitrescu, C., Rose, F. G., and Kato, S.: Clouds and the Earth's Radiant Energy System (CERES) Energy Balanced and Filled (EBAF) Top-of-Atmosphere (TOA) Edition-4.0 Data Product, Journal of Climate, 31, 895–918, https://doi.org/10.1175/JCLI-D-17-0208.1, 2018.

Loeb, N. G., Wang, H., Allan, R. P., Andrews, T., Armour, K., Cole, J. N. S., Dufresne, J.-L., Forster, P., Gettelman, A., Guo, H., Mauritsen, T., Ming, Y., Paynter, D., Proistosescu, C., Stuecker, M. F., Willén, U., and Wyser, K.: New Generation of Climate Models Track Recent Unprecedented Changes in Earth's Radiation Budget Observed by CERES, Geophysical Research Letters, 47, e2019GL086 705, https://doi.org/https://doi.org/10.1029/2019GL086705, 2020.

Lucarini, V., Ragone, F., and Lunkeit, F.: Predicting Climate Change Using Response Theory: Global Averages and Spatial Patterns, Journal of Statistical Physics, 166, 1036–1064, https://doi.org/10.1007/s10955-016-1506-z, 2017.

Luo, Y. and Hoffman, F. M.: Benchmark Analysis, in: Land Carbon Cycle Modeling: Matrix Approach, Data Assimilation, & Ecological Forecasting, edited by Luo, Y. and Smith, B., pp. 157–162, CRC Press, Boca Raton, FL, USA, ISBN 978-1-4987-3701-2, https://doi.org/10.1201/9780429155659-24, 2022.

Luo, Y. Q., Randerson, J. T., Abramowitz, G., Bacour, C., Blyth, E., Carvalhais, N., Ciais, P., Dalmonech, D., Fisher, J. B., Fisher, R., Friedlingstein, P., Hibbard, K., Hoffman, F. M., Huntzinger, D., Jones, C. D., Koven, C., Lawrence, D., Li, D. J., Mahecha, M., Niu, S. L., Norby, R., Piao, S. L., Qi, X., Peylin, P., Prentice, I. C., Riley, W., Reichstein, M., Schwalm, C., Wang, Y. P., Xia, J. Y., Zaehle, S., and Zhou, X. H.: A Framework for Benchmarking Land Models, Biogeosciences, 9, 3857–3874, https://doi.org/10.5194/bg-9-3857-2012, 2012.

MacDougall, A. H., Frölicher, T. L., Jones, C. D., Rogelj, J., Matthews, H. D., Zickfeld, K., Arora, V. K., Barrett, N. J., Brovkin, V., Burger, F. A., Eby, M., Eliseev, A. V., Hajima, T., Holden, P. B., Jeltsch-Thömmes, A., Koven, C., Mengis, N., Menviel, L., Michou, M., Mokhov, I. I., Oka, A., Schwinger, J., Séférian, R., Shaffer, G., Sokolov, A., Tachiiri, K., Tjiputra, J., Wiltshire, A., and Ziehn, T.: Is There Warming in the Pipeline? A Multi-model Analysis of the Zero Emissions Commitment from CO_2, Biogeosciences, 17, 2987–3016, https://doi.org/10.5194/bg-17-2987-2020, 2020.

Martens, B., Miralles, D. G., Lievens, H., van der Schalie, R., de Jeu, R. A. M., Fernández-Prieto, D., Beck, H. E., Dorigo, W. A., and Verhoest, N. E. C.: GLEAM v3: Satellite-based Land Evaporation and Root-zone Soil Moisture, Geoscientific Model Development, 10, 1903–1925, https://doi.org/10.5194/gmd-10-1903-2017, 2017.

Mauritsen, T., Stevens, B., Roeckner, E., Crueger, T., Esch, M., Giorgetta, M., Haak, H., Jungclaus, J., Klocke, D., Matei, D., Mikolajewicz, U., Notz, D., Pincus, R., Schmidt, H., and Tomassini, L.: Tuning the Climate of a Global Model, Journal of Advances in Modeling Earth Systems, 4, https://doi.org/https://doi.org/10.1029/2012MS000154, 2012.

Mauritsen, T., Bader, J., Becker, T., Behrens, J., Bittner, M., Brokopf, R., Brovkin, V., Claussen, M., Crueger, T., Esch, M., Fast, I., Fiedler, S., Fläschner, D., Gayler, V., Giorgetta, M., Goll, D. S., Haak, H., Hagemann, S., Hedemann, C., Hohenegger, C., Ilyina, T., Jahns, T., Jimenéz-de-la Cuesta, D., Jungclaus, J., Kleinen, T., Kloster, S., Kracher, D., Kinne, S., Kleberg, D., Lasslop, G., Kornblueh, L., Marotzke, J., Matei, D., Meraner, K., Mikolajewicz, U., Modali, K., Möbis, B., Müller, W. A., Nabel, J. E. M. S., Nam, C. C. W., Notz, D., Nyawira, S.-S., Paulsen, H., Peters, K., Pincus, R., Pohlmann, H., Pongratz, J., Popp, M., Raddatz, T. J., Rast, S., Redler, R., Reick, C. H., Rohrschneider, T., Schemann, V., Schmidt, H., Schnur, R., Schulzweida, U., Six, K. D., Stein, L., Stemmler, I., Stevens, B., von Storch, J.-S., Tian, F., Voigt, A., Vrese, P., Wieners, K.-H., Wilkenskjeld, S., Winkler, A., and Roeckner, E.: Developments in the MPI-M Earth System Model version 1.2 (MPI-ESM1.2) and Its Response to Increasing CO_2, Journal of Advances in Modeling Earth Systems, 11, 998–1038, https://doi.org/10.1029/2018MS001400, 2019.

Mayer, M., Kato, S., Bosilovich, M., Bechtold, P., Mayer, J., Schröder, M., Behrangi, A., Wild, M., Kobayashi, S., Li, Z., and L'Ecuyer, T.: Assessment of Atmospheric and Surface Energy Budgets Using Observation-Based Data Products, Surveys in Geophysics, 45, 1827–1854, https://doi.org/10.1007/s10712-024-09827-x, 2024.

Meehl, G. A., Senior, C. A., Eyring, V., Flato, G., Lamarque, J.-F., Stouffer, R. J., Taylor, K. E., and Schlund, M.: Context for Interpreting Equilibrium Climate Sensitivity and Transient Climate Response from the CMIP6 Earth System Models, Science Advances, 6, eaba1981, https://doi.org/10.1126/sciadv.aba1981, 2020.

Miller, S. M., Hayek, M. N., Andrews, A. E., Fung, I., and Liu, J.: Biases in Atmospheric CO_2 Estimates from Correlated Meteorology Modeling Errors, Atmospheric Chemistry and Physics, 15, 2903–2914, https://doi.org/10.5194/acp-15-2903-2015, 2015.

Miralles, D. G., Holmes, T. R. H., De Jeu, R. A. M., Gash, J. H., Meesters, A. G. C. A., and Dolman, A. J.: Global Land-surface Evaporation Estimated from Satellite-based Observations, Hydrology and Earth System Sciences, 15, 453–469, https://doi.org/10.5194/hess-15-453-2011, 2011.

Moat, B. I., Smeed, D., Rayner, D., Johns, W. E., Smith, R. H., Volkov, D. L., Elipot, S., Petit, T., Kajtar, J. B., Baringer, M. O., and Collins, J.: Atlantic Meridional Overturning Circulation Observed by the RAPID-MOCHA-WBTS Array at 26N from 2004 to 2023 (v2023.1a), https://doi.org/10.5285/33826d6e-801c-b0a7-e063-7086abc0b9db, Archive Location: North Atlantic Ocean, Straits of Florida, Atlantic Ocean, Northeast Atlantic Ocean (40°W), Northwest Atlantic Ocean (40°W), published by NERC EDS British Oceanographic Data Centre NOC, 2025.

Morice, C. P., Kennedy, J. J., Rayner, N. A., Winn, J. P., Hogan, E., Killick, R. E., Dunn, R. J. H., Osborn, T. J., Jones, P. D., and Simpson, I. R.: An Updated Assessment of Near-Surface Temperature Change From 1850: The HadCRUT5 Data Set, Journal of Geophysical Research: Atmospheres, 126, e2019JD032 361, https://doi.org/10.1029/2019JD032361, 2021.

Notz, D. and SIMIP Community: Arctic Sea Ice in CMIP6, Geophysical Research Letters, 47, e2019GL086 749, https://doi.org/https://doi.org/10.1029/2019GL086749, 2020.

Nowicki, S. M. J., Payne, A., Larour, E., Seroussi, H., Goelzer, H., Lipscomb, W., Gregory, J., Abe-Ouchi, A., and Shepherd, A.: Ice Sheet Model Intercomparison Project (ISMIP6) contribution to CMIP6, Geoscientific Model Development, 9, 4521–4545, https://doi.org/10.5194/gmd-9-4521-2016, 2016.

Ordonez, A., Ullrich, P., and Lee, J.: cmecmetrics/cmec-driver: v1.1.9-release, https://doi.org/10.5281/zenodo.15548748, 2025.

O'Rourke, E.: CMIP AR7 Fast Track modelling centres survey: Spin up metrics, ocean regridding, BECCS and Rapid Evaluation Tool, https://doi.org/10.5281/zenodo.12731068, 2024.

O'Rourke, E.: CMIP6 Community Survey Results, https://doi.org/10.5281/zenodo.11654909, 2025.

Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y.-W., Poindexter, C., Chen, J., Elbashandy, A., Humphrey, M., Isaac, P., Polidori, D., Reichstein, M., Ribeca, A., van Ingen, C., Vuichard, N., Zhang, L., Amiro, B., Ammann, C., Arain, M. A., Ardö, J., Arkebauer, T., Arndt, S. K., Arriga, N., Aubinet, M., Aurela, M., Baldocchi, D., Barr, A., Beamesderfer, E., Marchesini, L. B., Bergeron, O., Beringer, J., Bernhofer, C., Berveiller, D., Billesbach, D., Black, T. A., Blanken, P. D., Bohrer, G., Boike, J., Bolstad, P. V., Bonal, D., Bonnefond, J.-M., Bowling, D. R., Bracho, R., Brodeur, J., Brümmer, C., Buchmann, N., Burban, B., Burns, S. P., Buysse, P., Cale, P., Cavagna, M., Cellier, P., Chen, S., Chini, I., Christensen, T. R., Cleverly, J., Collalti, A., Consalvo, C., Cook, B. D., Cook, D., Coursolle, C., Cremonese, E., Curtis, P. S., D'Andrea, E., da Rocha, H., Dai, X., Davis, K. J., Cinti, B. D., Grandcourt, A. d., Ligne, A. D., De Oliveira, R. C., Delpierre, N., Desai, A. R., Di Bella, C. M., Tommasi, P. d., Dolman, H., Domingo, F., Dong, G., Dore, S., Duce, P., Dufrêne, E., Dunn, A., Dušek, J., Eamus, D., Eichelmann, U., ElKhidir, H. A. M., Eugster, W., Ewenz, C. M., Ewers, B., Famulari, D., Fares, S., Feigenwinter, I., Feitz, A., Fensholt, R., Filippa, G., Fischer, M., Frank, J., Galvagno, M., Gharun, M., Gianelle, D., Gielen, B., Gioli, B., Gitelson, A., Goded, I., Goeckede, M., Goldstein, A. H., Gough, C. M., Goulden, M. L., Graf, A., Griebel, A., Gruening, C., Grünwald, T., Hammerle, A., Han, S., Han, X., Hansen, B. U., Hanson, C., Hatakka, J., He, Y., Hehn, M., Heinesch, B., Hinko-Najera, N., Hörtnagl, L., Hutley, L., Ibrom, A., Ikawa, H., Jackowicz-Korczynski, M., Janouš, D., Jans, W., Jassal, R., Jiang, S., Kato, T., Khomik, M., Klatt, J., Knohl, A., Knox, S., Kobayashi, H., Koerber, G., Kolle, O., Kosugi, Y., Kotani, A., Kowalski, A., Kruijt, B., Kurbatova, J., Kutsch, W. L., Kwon, H., Launiainen, S., Laurila, T., Law, B., Leuning, R., Li, Y., Liddell, M., Limousin, J.-M., Lion, M., Liska, A. J., Lohila, A., López-Ballesteros, A., López-Blanco, E., Loubet, B., Loustau, D., Lucas-Moffat, A., Lüers, J., Ma, S., Macfarlane, C., Magliulo, V., Maier, R., Mammarella, I., Manca, G., Marcolla, B., Margolis, H. A., Marras, S., Massman, W., Mastepanov, M., Matamala, R., Matthes, J. H., Mazzenga, F., McCaughey, H., McHugh, I., McMillan, A. M. S., Merbold, L., Meyer, W., Meyers, T., Miller, S. D., Minerbi, S., Moderow, U., Monson, R. K., Montagnani, L., Moore, C. E., Moors, E., Moreaux, V., Moureaux, C., Munger, J. W., Nakai, T., Neirynck, J., Nesic, Z., Nicolini, G., Noormets, A., Northwood, M., Nosetto, M., Nouvellon, Y., Novick, K., Oechel, W., Olesen, J. E., Ourcival, J.-M., Papuga, S. A., Parmentier, F.-J., Paul-Limoges, E., Pavelka, M., Peichl, M., Pendall, E., Phillips, R. P., Pilegaard, K., Pirk, N., Posse, G., Powell, T., Prasse, H., Prober, S. M., Rambal, S., Rannik, U., Raz-Yaseef, N., Rebmann, C., Reed, D., Dios, V. R. d., Restrepo-Coupe, N., Reverter, B. R., Roland, M., Sabbatini, S., Sachs, T., Saleska, S. R., Sánchez-Cañete, E. P., Sanchez-Mejia, Z. M., Schmid, H. P., Schmidt, M., Schneider, K., Schrader, F., Schroder, I., Scott, R. L., Sedlák, P., Serrano-Ortíz, P., Shao, C., Shi, P., Shironya, I., Siebicke, L., Šigut, L., Silberstein, R., Sirca, C., Spano, D., Steinbrecher, R., Stevens, R. M., Sturtevant, C., Suyker, A., Tagesson, T., Takanashi, S., Tang, Y., Tapper, N., Thom, J., Tomassucci, M., Tuovinen, J.-P., Urbanski, S., Valentini, R., van der Molen, M., van Gorsel, E., van Huissteden, K., Varlagin, A., Verfaillie, J., Vesala, T., Vincke, C., Vitale, D., Vygodskaya, N., Walker, J. P., Walter-Shea, E., Wang, H., Weber, R., Westermann, S., Wille, C., Wofsy, S., Wohlfahrt, G., Wolf, S., Woodgate, W., Li, Y., Zampedri, R., Zhang, J., Zhou, G., Zona, D., Agarwal, D., Biraud, S., Torn, M., and Papale, D.: The FLUXNET2015 Dataset and the ONEFlux Processing Pipeline for Eddy Covariance Data, Scientific Data, 7, 225, https://doi.org/10.1038/s41597-020-0534-3, 2020.

Planton, Y. Y., Guilyardi, E., Wittenberg, A. T., Lee, J., Gleckler, P. J., Bayr, T., McGregor, S., McPhaden, M. J., Power, S., Roehrig, R., Vialard, J., and Voldoire, A.: Evaluating Climate Models with the CLIVAR 2020 ENSO Metrics Package, Bulletin of the American Meteorological Society, 102, E193–E217, https://doi.org/10.1175/BAMS-D-19-0337.1, 2021.

Priestley, M. D. K., Ackerley, D., Catto, J. L., Hodges, K. I., McDonald, R. E., and Lee, R. W.: An Overview of the Extratropical Storm Tracks in CMIP6 Historical Simulations, Journal of Climate, 33, 6315–6343, https://doi.org/10.1175/JCLI-D-19-0928.1, 2020.

Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., Rowell, D. P., Kent, E. C., and Kaplan, A.: Global Analyses of Sea Surface Temperature, Sea Ice, and Night Marine Air Temperature Since the Late Nineteenth Century, Journal of Geophysical Research: Atmospheres, 108, https://doi.org/10.1029/2002JD002670, 2003.

Reagan, J. R., Boyer, T. P., García, H. E., Locarnini, R. A., Baranova, O. K., Bouchard, C., Cross, S. L., Mishonov, A. V., Paver, C. R., Seidov, D., Wang, Z., and Dukhovskoy, D.: World Ocean Atlas 2023, https://doi.org/10.25921/VA26-HV25, 2023.

Righi, M., Andela, B., Eyring, V., Lauer, A., Predoi, V., Schlund, M., Vegas-Regidor, J., Bock, L., Brötz, B., de Mora, L., Diblen, F., Dreyer, L., Drost, N., Earnshaw, P., Hassler, B., Koldunov, N., Little, B., Loosveldt Tomas, S., and Zimmermann, K.: Earth System Model Evaluation Tool (ESMValTool) v2.0 – Technical Overview, Geoscientific Model Development, 13, 1179–1199, https://doi.org/10.5194/gmd-13-1179-2020, 2020.

Roach, L. A., Dörr, J., Holmes, C. R., Massonnet, F., Blockley, E. W., Notz, D., Rackow, T., Raphael, M. N., O'Farrell, S. P., Bailey, D. A., and Bitz, C. M.: Antarctic Sea Ice Area in CMIP6, Geophysical Research Letters, 47, e2019GL086 729, https://doi.org/10.1029/2019GL086729, 2020.

Roberts, M. J., Camp, J., Seddon, J., Vidale, P. L., Hodges, K., Vannière, B., Mecking, J., Haarsma, R., Bellucci, A., Scoccimarro, E., Caron, L.-P., Chauvin, F., Terray, L., Valcke, S., Moine, M.-P., Putrasahan, D., Roberts, C. D., Senan, R., Zarzycki, C., Ullrich, P., Yamada, Y., Mizuta, R., Kodama, C., Fu, D., Zhang, Q., Danabasoglu, G., Rosenbloom, N., Wang, H., and Wu, L.: Projected Future Changes in Tropical Cyclones Using the CMIP6 HighResMIP Multimodel Ensemble, Geophysical Research Letters, 47, e2020GL088 662, https://doi.org/10.1029/2020GL088662, 2020.

Rutz, J. J., Shields, C. A., Lora, J. M., Payne, A. E., Guan, B., Ullrich, P., O'Brien, T., Leung, L. R., Ralph, F. M., Wehner, M., Brands, S., Collow, A., Goldenson, N., Gorodetskaya, I., Griffith, H., Kashinath, K., Kawzenuk, B., Krishnan, H., Kurlin, V., Lavers, D., Magnusdottir, G., Mahoney, K., McClenny, E., Muszynski, G., Nguyen, P. D., Prabhat, M., Qian, Y., Ramos, A. M., Sarangi, C., Sellars, S., Shulgina, T., Tome, R., Waliser, D., Walton, D., Wick, G., Wilson, A. M., and Viale, M.: The Atmospheric River Tracking Method Intercomparison

Project (ARTMIP): Quantifying Uncertainties in Atmospheric River Climatology, Journal of Geophysical Research: Atmospheres, 124, 13 777–13 802, https://doi.org/10.1029/2019JD030936, 2019.

Sanderson, B. M., Brovkin, V., Fisher, R. A., Hohn, D., Ilyina, T., Jones, C. D., Koenigk, T., Koven, C., Li, H., Lawrence, D. M., Lawrence, P., Liddicoat, S., MacDougall, A. H., Mengis, N., Nicholls, Z., O'Rourke, E., Romanou, A., Sandstad, M., Schwinger, J., Séférian, R., Sentman, L. T., Simpson, I. R., Smith, C., Steinert, N. J., Swann, A. L. S., Tjiputra, J., and Ziehn, T.: flat10MIP: an emissions-driven experiment to diagnose the climate response to positive, zero and negative CO_2 emissions, Geoscientific Model Development, 18, 5699–5724, https://doi.org/10.5194/gmd-18-5699-2025, 2025.

Scaife, A. A., Woollings, T., Knight, J., Martin, G., and Hinton, T.: Atmospheric Blocking and Mean Biases in Climate Models, Journal of Climate, 23, 6143–6152, https://doi.org/10.1175/2010JCLI3728.1, 2010.

Schlund, M., Lauer, A., Gentine, P., Sherwood, S. C., and Eyring, V.: Emergent Constraints on Equilibrium Climate Sensitivity in CMIP5: Do They Hold for CMIP6?, Earth System Dynamics, 11, 1233–1258, https://doi.org/10.5194/esd-11-1233-2020, 2020.

Schlund, M., Hassler, B., Lauer, A., Andela, B., Jöckel, P., Kazeroni, R., Loosveldt Tomas, S., Medeiros, B., Predoi, V., Sénési, S., Servonnat, J., Stacke, T., Vegas-Regidor, J., Zimmermann, K., and Eyring, V.: Evaluation of Native Earth System Model Output with ESMValTool v2.6.0, Geoscientific Model Development, 16, 315–333, https://doi.org/10.5194/gmd-16-315-2023, 2023.

Schlund, M., Andela, B., Benke, J., Comer, R., Hassler, B., Hogan, E., Kalverla, P., Lauer, A., Little, B., Loosveldt Tomas, S., Nattino, F., Peglar, P., Predoi, V., Smeets, S., Worsley, S., Yeo, M., and Zimmermann, K.: Advanced climate model evaluation with ESMValTool v2.11.0 using parallel, out-of-core, and distributed computing, Geoscientific Model Development, 18, 4009–4021, https://doi.org/10.5194/gmd-18-4009-2025, 2025.

Seneviratne, S. I., Corti, T., Davin, E. L., Hirschi, M., Jaeger, E. B., Lehner, I., Orlowsky, B., and Teuling, A. J.: Investigating Soil Moisture–Climate Interactions in a Changing Climate: A Review, Earth-Science Reviews, 99, 125–161, https://doi.org/https://doi.org/10.1016/j.earscirev.2010.02.004, 2010.

Senior, C. A., Jones, C. G., Wood, R. A., Sellar, A., Belcher, S., Klein-Tank, A., Sutton, R., Walton, J., Lawrence, B., Andrews, T., and Mulcahy, J. P.: U.K. Community Earth System Modeling for CMIP6, Journal of Advances in Modeling Earth Systems, 12, e2019MS002 004, https://doi.org/10.1029/2019MS002004, 2020.

Slivinski, L. C., Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Giese, B. S., McColl, C., Allan, R., Yin, X., Vose, R., Titchner, H., Kennedy, J., Spencer, L. J., Ashcroft, L., Brönnimann, S., Brunet, M., Camuffo, D., Cornes, R., Cram, T. A., Crouthamel, R., Domínguez-Castro, F., Freeman, J. E., Gergis, J., Hawkins, E., Jones, P. D., Jourdain, S., Kaplan, A., Kubota, H., Blancq, F. L., Lee, T.-C., Lorrey, A., Luterbacher, J., Maugeri, M., Mock, C. J., Moore, G. K., Przybylak, R., Pudmenzky, C., Reason, C., Slonosky, V. C., Smith, C. A., Tinz, B., Trewin, B., Valente, M. A., Wang, X. L., Wilkinson, C., Wood, K., and Wyszyński, P.: Towards a More Reliable Historical Reanalysis: Improvements for Version 3 of the Twentieth Century Reanalysis System, Quarterly Journal of the Royal Meteorological Society, 145, 2876–2908, https://doi.org/10.1002/qj.3598, 2019.

Sobolowski, S., Somot, S., Fernandez, J., Evin, G., Brands, S., Maraun, D., Kotlarski, S., Jury, M., Benestad, R. E., Teichmann, C., Christensen, O. B., Bülow, K., Buonomo, E., Katragkou, E., Steger, C., Sørland, S., Nikulin, G., McSweeney, C., Dobler, A., Palmer, T., Wilcke, R., Boé, J., Brunner, L., Ribes, A., Qasmi, S., Nabat, P., Sevault, F., and Oudar, T.: GCM Selection and Ensemble Design: Best Practices and Recommendations from the EURO-CORDEX Community, Bulletin of the American Meteorological Society, pp. E1834–E1850, https://doi.org/10.1175/BAMS-D-23-0189.1, 2025.

Sofieva, V. F., Szelag, M., Tamminen, J., Arosio, C., Rozanov, A., Weber, M., Degenstein, D., Bourassa, A., Zawada, D., Kiefer, M., Laeng, A., Walker, K. A., Sheese, P., Hubert, D., van Roozendael, M., Retscher, C., Damadeo, R., and Lumpe, J. D.: Updated Merged SAGE-

CCI-OMPS+ Dataset for the Evaluation of Ozone Trends in the Stratosphere, Atmospheric Measurement Techniques, 16, 1881–1899, https://doi.org/10.5194/amt-16-1881-2023, 2023.

Solberg, R., Rudjord, Ø., Salberg, A.-B., Killie, M. A., Eastwood, S., Sørensen, A., Marin, C., Premier, V., Schwaizer, G., and Nagler, T.: ESA Snow Climate Change Initiative (Snow_cci): Fractional Snow Cover in CryoClim, v1.0, https://doi.org/10.5285/F4654030223445B0BAC63A23AAA60620, 2023.

Stainforth, D. A., Allen, M. R., Tredger, E. R., and Smith, L. A.: Confidence, Uncertainty and Decision-support Relevance in Climate Predictions, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, https://doi.org/10.1098/rsta.2007.2074, 2007.

Stockhause, M., Huard, D., Al Khourdajie, A., Gutiérrez, J. M., Kawamiya, M., Klutse, N. A. B., Krey, V., Milward, D., Okem, A. E., Pirani, A., Sitz, L. E., Solman, S. A., Spinuso, A., and Xing, X.: Implementing FAIR Data Principles in the IPCC Seventh Assessment Cycle: Lessons Learned and Future Prospects, PLOS Climate, 3, e0000 533, https://doi.org/10.1371/journal.pclm.0000533, 2024.

Swaminathan, R., Parker, R. J., Jones, C. G., Allan, R. P., Quaife, T., Kelley, D. I., Mora, L. d., and Walton, J.: The Physical Climate at Global Warming Thresholds as Seen in the U.K. Earth System Model, Journal of Climate, 35, 29–48, https://doi.org/10.1175/JCLI-D-21-0234.1, 2022.

Séférian, R., Nabat, P., Michou, M., Saint-Martin, D., Voldoire, A., Colin, J., Decharme, B., Delire, C., Berthet, S., Chevallier, M., Sénési, S., Franchisteguy, L., Vial, J., Mallet, M., Joetzjer, E., Geoffroy, O., Guérémy, J.-F., Moine, M.-P., Msadek, R., Ribes, A., Rocher, M., Roehrig, R., Salas-y Mélia, D., Sanchez, E., Terray, L., Valcke, S., Waldman, R., Aumont, O., Bopp, L., Deshayes, J., Éthé, C., and Madec, G.: Evaluation of CNRM Earth System Model, CNRM-ESM2-1: Role of Earth System Processes in Present-Day and Future Climate, Journal of Advances in Modeling Earth Systems, 11, 4182–4227, https://doi.org/10.1029/2019MS001791, 2019.

Séférian, R., Berthet, S., Yool, A., Palmiéri, J., Bopp, L., Tagliabue, A., Kwiatkowski, L., Aumont, O., Christian, J., Dunne, J., Gehlen, M., Ilyina, T., John, J. G., Li, H., Long, M. C., Luo, J. Y., Nakano, H., Romanou, A., Schwinger, J., Stock, C., Santana-Falcón, Y., Takano, Y., Tjiputra, J., Tsujino, H., Watanabe, M., Wu, T., Wu, F., and Yamamoto, A.: Tracking Improvement in Simulated Marine Biogeochemistry Between CMIP5 and CMIP6, Current Climate Change Reports, 6, 95–119, https://doi.org/10.1007/s40641-020-00160-0, 2020.

Taylor, K., Doutriaux, C., and Peterschmitt, J.: Climate Model Output Rewriter (CMOR), Tech. Rep. UCRL-TR-204637, 15014202, Lawrence Livermore National Laboratory, https://doi.org/10.2172/15014202, 2004.

Taylor, K. E.: Summarizing Multiple Aspects of Model Performance in a Single Diagram, Journal of Geophysical Research: Atmospheres, 106, 7183–7192, https://doi.org/10.1029/2000JD900719, 2001.

Taylor, K. E., Troussellier, L., Ames, S., Hassell, D., Molina, M., Nicholls, Z., Schupfner, M., Anstey, J., Ellis, D., Dingley, B., Durack, P. J., LEVAVASSEUR, G., Mizielinski, M., and Moine, M.-P.: CMIP7 Global Attributes, DRS, Filenames, Directory Structure, and CVs, https://doi.org/10.5281/zenodo.17250296, 2026.

Teixeira, J., Waliser, D., Ferraro, R., Gleckler, P., Lee, T., and Potter, G.: Satellite Observations for CMIP5: The Genesis of Obs4MIPs, Bulletin of the American Meteorological Society, 95, 1329–1334, https://doi.org/10.1175/BAMS-D-12-00204.1, 2025.

Thompson, D. W. J. and Wallace, J. M.: Annular Modes in the Extratropical Circulation. Part I: Month-to-Month Variability, Journal of Climate, 13, 1000 – 1016, https://doi.org/10.1175/1520-0442(2000)013<1000:AMITEC>2.0.CO;2, 2000.

Tjiputra, J. F., Negrel, J., and Olsen, A.: Early Detection of Anthropogenic Climate Change Signals in the Ocean Interior, Scientific Reports, 13, 3006, https://doi.org/10.1038/s41598-023-30159-0, 2023.

Trenberth, K. E. and Caron, J. M.: Estimates of Meridional Atmosphere and Ocean Heat Transports, Journal of Climate, 14, 3433–3443, https://doi.org/10.1175/1520-0442(2001)014<3433:EOMAAO>2.0.CO;2, 2001.

Trenberth, K. E., Branstator, G. W., Karoly, D., Kumar, A., Lau, N.-C., and Ropelewski, C.: Progress During TOGA in Understanding and Modeling Global Teleconnections Associated with Tropical Sea Surface Temperatures, Journal of Geophysical Research: Oceans, 103, 14 291–14 324, https://doi.org/10.1029/97JC01444, 1998.

Trenberth, K. E., Smith, L., Qian, T., Dai, A., and Fasullo, J.: Estimates of the Global Water Budget and Its Annual Cycle Using Observational and Model Data, Journal of Hydrometeorology, 8, 758–769, https://doi.org/10.1175/JHM600.1, 2007.

Trumbore, S. E. and Czimczik, C. I.: An Uncertain Future for Soil Carbon, Science, 321, 1455–1456, https://doi.org/10.1126/science.1160232, 2008.

Vaittinada Ayar, P., Battisti, D. S., Li, C., King, M., Vrac, M., and Tjiputra, J.: A Regime View of ENSO Flavors Through Clustering in CMIP6 Models, Earth's Future, 11, e2022EF003 460, https://doi.org/10.1029/2022EF003460, 2023.

Waliser, D., Gleckler, P. J., Ferraro, R., Taylor, K. E., Ames, S., Biard, J., Bosilovich, M. G., Brown, O., Chepfer, H., Cinquini, L., Durack, P. J., Eyring, V., Mathieu, P.-P., Lee, T., Pinnock, S., Potter, G. L., Rixen, M., Saunders, R., Schulz, J., Thépaut, J.-N., and Tuma, M.: Observations for Model Intercomparison Project (Obs4MIPs): Status for CMIP6, Geoscientific Model Development, 13, 2945–2958, https://doi.org/10.5194/gmd-13-2945-2020, 2020.

Wallace, J. M. and Gutzler, D. S.: Teleconnections in the Geopotential Height Field during the Northern Hemisphere Winter, Monthly Weather Review, 109, 784–812, https://doi.org/10.1175/1520-0493(1981)109<0784:TITGHF>2.0.CO;2, 1981.

Wan, H., Zhang, K., Rasch, P. J., Singh, B., Chen, X., and Edwards, J.: A New and Inexpensive Non-bit-for-bit Solution Reproducibility Test Based on Time Step Convergence (TSC1.0), Geoscientific Model Development, 10, 537–552, https://doi.org/10.5194/gmd-10-537-2017, 2017.

Wang, B., Kim, H.-J., Kikuchi, K., and Kitoh, A.: Diagnostic Metrics for Evaluation of Annual and Diurnal Cycles, Climate Dynamics, 37, 941–955, https://doi.org/10.1007/s00382-010-0877-0, 2011.

Wang, H., Pearson, B., Hou, A., Lanfer Marquez, A., and Bonnet, P.: Model Evaluation and Benchmarking: Community Survey Summary Report, https://doi.org/10.5281/zenodo.15212597, 2025.

Wang, Y. and Mao, J.: Global Multi-layer Soil Moisture Products, https://doi.org/10.6084/M9.FIGSHARE.13661312.V1, data set, 2021.

Wang, Y., Mao, J., Jin, M., Hoffman, F. M., Shi, X., Wullschleger, S. D., and Dai, Y.: Development of Observation-based Global Multilayer Soil Moisture Products for 1970 to 2016, Earth System Science Data, 13, 4385–4405, https://doi.org/10.5194/essd-13-4385-2021, 2021.

Warszawski, L., Frieler, K., Huber, V., Piontek, F., Serdeczny, O., and Schewe, J.: The Inter-Sectoral Impact Model Intercomparison Project (ISI–MIP): Project Framework, Proceedings of the National Academy of Sciences, 111, 3228–3232, https://doi.org/10.1073/pnas.1312330110, 2014.

Weigel, K., Bock, L., Gier, B. K., Lauer, A., Righi, M., Schlund, M., Adeniyi, K., Andela, B., Arnone, E., Berg, P., Caron, L.-P., Cionni, I., Corti, S., Drost, N., Hunter, A., Lledó, L., Mohr, C. W., Paçal, A., Pérez-Zanón, N., Predoi, V., Sandstad, M., Sillmann, J., Sterl, A., Vegas-Regidor, J., von Hardenberg, J., and Eyring, V.: 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, 3159–3184, https://doi.org/10.5194/gmd-14-3159-2021, 2021.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 'T Hoen, P. A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., Van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., Van Der Lei, J., Van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B.: The FAIR Guiding Principles for Scientific Data Management and Stewardship, Scientific Data, 3, 160 018, https://doi.org/10.1038/sdata.2016.18, 2016.

Winker, D.: CALIPSO Lidar Level 3 Ice Cloud Data, Standard V1-00, https://doi.org/10.5067/CALIOP/CALIPSO/L3_ICE_CLOUD-STANDARD-V1-00, published by NASA Langley Atmospheric Science Data Center Distributed Active Archive Center, 2024.

Winker, D., Cai, X., Vaughan, M., Garnier, A., Magill, B., Avery, M., and Getzewich, B.: A Level 3 Monthly Gridded Ice Cloud Dataset Derived from 12 Years of CALIOP Measurements, Earth System Science Data, 16, 2831–2855, https://doi.org/10.5194/essd-16-2831-2024, 2024.

Yang, X., Ricciuto, D. M., Thornton, P. E., Shi, X., Xu, M., Hoffman, F. M., and Norby, R. J.: The Effects of Phosphorus Cycle Dynamics on Carbon Sources and Sinks in the Amazon Region: A Modeling Study Using ELM v1, Journal of Geophysical Research: Biogeosciences, 124, 3686–3698, https://doi.org/10.1029/2019JG005082, 2019.

Yukimoto, S., Kawai, H., Koshiro, T., Oshima, N., Yoshida, K., Urakawa, S., Tsujino, H., Deushi, M., Tanaka, T., Hosaka, M., Yabu, S., Yoshimura, H., Shindo, E., Mizuta, R., Obata, A., Adachi, Y., and Ishii, M.: The Meteorological Research Institute Earth System Model Version 2.0, MRI-ESM2.0: Description and Basic Evaluation of the Physical Component, Journal of the Meteorological Society of Japan. Ser. II, 97, 931–965, https://doi.org/10.2151/jmsj.2019-051, 2019.

Zhu, Q., Riley, W. J., Tang, J., Collier, N., Hoffman, F. M., Yang, X., and Bisht, G.: Representing Nitrogen, Phosphorus, and Carbon Interactions in the E3SM Land Model: Development and Global Benchmarking, Journal of Advances in Modeling Earth Systems, 11, 2238–2258, https://doi.org/10.1029/2018MS001571, 2019.