CMIP
Coupled Model Intercomparison Project

RIfS
Regional Information for Society

WCRP

CLIMATE-RAPID EVALUATION FRAMEWORK (REF)

BEST PRACTICE FOR USING THE CMIP REF OUTPUT

March 2026 Version 1 DOI: 10.5281/zenodo.18775782

Terms and Conditions

Bibliographic Information

This report should be cited as: Morrison, M. A., Daba, D., Abrha, H., Chung, C., Samanta, D., Hoffman, F. M., Swaminathan, R., Brands, S., and Bonou, F. (2026), *Best Practice for Using the CMIP Rapid Evaluation Framework Output*. https://doi.org/10.5281/zenodo.18775782

Disclaimer

The designations employed in the WCRP Coupled Model Intercomparison Project, RIfS or CMIP, publications and the presentation of material in this publication do not imply the expression of any opinion whatsoever on the part of CMIP, RIfS, the World Climate Research Programme (WCRP) – including its Sponsor Organizations the World Meteorological Organization (WMO), the Intergovernmental Oceanographic Commission (IOC) of UNESCO and the International Science Council (ISC) – or the European Space Agency (ESA) in its role as host organization of the CMIP International Project Office (CMIP-IPO), or Ouranos Inc. in its role as host organization of the RIfS International Project Office (RIfS-IPO), concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

The findings, interpretations and conclusions expressed in CMIP and RIfS publications with named authors are those of the authors alone and do not necessarily reflect those of RIfS, CMIP, WCRP, its Sponsor Organizations, ESA or of their Members.

This document is not an official publication of the WCRP and has been issued without formal editing.

The views expressed herein do not necessarily have the endorsement of WCRP or its Members.

Any potential mention of specific companies or products does not imply that they are endorsed or recommended in preference to others of a similar nature which are not mentioned or advertised.

Copyright notice

publications@wcrp-rifs.org

RIfS International Project Office

Ouranos Inc.
550 Sherbrooke Street W
Montreal, Quebec
H3A 1B9, Canada

Authorship and publisher's notice

This report was authored by: Monica Ainhorn Morrison, Demiso Daba, Haftu Abrha, Christine Chung, Dhrubajyoti Samanta, Forrest M. Hoffman, R. Swaminathan, Swen Brands, Frederic Bonou.

Acknowledgements

This best practice note was prepared by members of the CMIP and RIfS joint task team on Responsible Data Use and members of the CMIP Model Benchmarking Task Team. The content was reviewed by members of the CMIP Model Benchmarking Task Team, the REF Delivery Team and the CMIP Panel.

CMIP is a project of the World Climate Research Programme (WCRP) Earth System Modelling and Observations (ESMO) Core Project . RIfS is one of the Core Projects of the WCRP.

The Rapid Evaluation Framework (REF) is a community project developed by CMIP. Development of the CMIP7 ready REF has been overseen by members of the CMIP Model Benchmarking Task Team and CMIP Panel. It has been funded by the European Space Agency and the U.S. Department of Energy.

The REF has been built by a delivery team of community members from the following organizations: Climate Resource, Netherlands eScience Center, DLR, Oak Ridge National Laboratory and Lawrence Livermore National Laboratory.

Contents

Document Purpose

This document offers guidance for users of the CMIP Rapid Evaluation Framework (REF)[1] on how to interpret REF outputs and apply them effectively. It is designed to help users from the CMIP community get the most out of both CMIP data and the REF by providing best practices for rigorous and informed use of the evaluation tool and results. The guidance addresses how to account for model uncertainty, reference data uncertainty, and internal variability, and helps users identify the most relevant metrics and diagnostics for their purposes.

Scope Statement

This document provides non-binding community guidance on how to interpret REF outputs in a rigorous and informed manner. It does not mandate model evaluation or selection procedures (i.e., it does not serve as fit-for-all purposes model evaluation or selection procedures) and does not replace domain-specific scientific judgment and analysis. The guidance reflects the current capabilities and limitations of the REF at the time of writing and may evolve as the framework develops. Within this document, REF diagnostics[2] are considered as providing an initial, standardized comparison step that may inform, but does not replace, subsequent validation, attribution, and application-specific evaluation. REF outputs are best interpreted as diagnostics for further investigation, not as endpoints.

[1] Find out more about the REF here: https://climate-ref.org/

[2] Diagnostic Collection List is available here: DOI 10.5281/zenodo.14284374.

Understanding the REF and its Purpose

The REF is:	The REF is not:
A rapid, standardized entry point for evaluating CMIP models behavior against reference datasets.	A validation of model correctness.
Designed to support early-stage assessment, exploration, and ranking and selection of models for further assessment.	A definitive or universal ranking of model performance.
Aimed at highlighting relative multi-model differences, strengths, and weaknesses across multiple diagnostics.	A substitute for expert-driven, process-based, regional or application specific evaluation.

Note: Certain metrics, such as TCR, TCRE, and ECS are included on the dashboard, but do not have a reference dataset, hence, not all metrics have observations to compare against, and certain evaluations in the REF are for the sake of model intercomparison.

The following table illustrates how undesirable approaches can be strengthened with consideration of best practices:

Undesirable Approach	Best Practice Alternative
Treating REF summary scores as a universal ranking of model fidelity	Use multiple diagnostics to understand which models are more suitable for a specific application or region
Eliminating a large fraction of the ensemble based on one or two diagnostics	Retain model diversity and document the reasoning for any weighting or narrowing of the ensemble
Describing models as “validated” based on REF diagnostics alone	Describe REF results as benchmarking against selected reference datasets, noting that evaluation involves additional steps beyond the REF’s current scope

Primary Purpose of the REF

The REF serves three primary purposes for the CMIP user community:

1. ***Systematic comparison across the CMIP ensemble.*** The REF provides systematic, open-source, community-driven comparison tools that can help with the identification of relative model strengths and weaknesses, as well as potential errors in model output, across CMIP participating models and the broader ensemble. Drawing on diverse data sources and software, this allows users to better understand the range of model behaviors and make informed decisions about the CMIP ensemble.

2. ***Identifying model subsets for specific research questions.*** REF diagnostics can help identify a subset of models for a specific study by examining diagnostic criteria related to model performance in key processes or regions relevant to the research question. For example, results for global diagnostics (such as global mean temperature) are not necessarily indicative of results for regional-scale diagnostics. Application specific metrics are therefore valuable for determining which models are well-suited for which research questions.

3. ***Informing model selection for downscaling and weighting.*** REF outputs can inform model selection for downscaling applications (Sobolowski et al. 2025) and multi-model weighting schemes (Merrifield et al. 2023), provided that evaluation metrics are matched to the specific variables and processes most relevant to the intended use. Fidelity of standard climate measures is related to, but does not guarantee, fidelity for specialized applications like hydroclimate extremes or agricultural risk assessments.

The REF is most appropriately used at the early stages of an analysis workflow, i.e., as an entry point for deeper investigation into CMIP models rather than as a definitive judgment of model quality. Because all models have varying strengths and limitations for particular research applications, REF results are best understood as a first look into the larger context of challenges in climate modeling and ongoing need for observational and inter-model comparison. The relevance of model fidelity over the historical period to future projections requires comprehensive analysis beyond the REF (see Eyring et al. 2019).

Some REF diagnostics provide spatially explicit or regionally resolved information, while others are global or aggregated by design. Cautious interpretation is warranted when inferring regional model performance solely from global metrics. Confirming that the spatial and temporal resolution of selected diagnostics is appropriate for a given application is an important part of rigorous and informed use.

Ten Considerations for Best Practices

1. **Choose models based on multiple metrics and diagnostics, not a single indicator.** Model performance varies across regions, variables, and temporal scales. A model that appears to be the best fit for one application can perform poorly for other applications. Using a suite of complementary metrics to examine model performance across multiple relevant dimensions provides a more robust basis for decision-making (Sobolowski et al. 2025).

2. **Match evaluation metrics to the specific application.** Different research questions call for different evaluation criteria. Metrics relevant for global mean climate may not be informative for regional extremes, seasonal cycles, or process-specific applications. Selecting diagnostics that are fit for purpose is central to rigorous application of the REF.

3. **Understand that model performance and value are relative to intended use.** No model is universally "good" or "bad". Each model represents some components of the Earth system with more fidelity than others. Fitness-for-purpose is always relative to a specific research question, regional, variable, and temporal scale.

4. **Look beyond summary scores.** Investigating the underlying information used to generate a summary score, such as individual metrics scalars, spatial patterns, and variable-level performance, provides a more complete and nuanced picture than aggregated results alone.

5. **Compliment the REF with process-based diagnostics where available.** Underlying physical process representations drive model behavior, so process-based diagnostics tend to provide more reliable and interpretable guidance than purely statistical measures. Where available, these diagnostics are a valuable complement to standard metrics.

6. **Account for observational uncertainty and gaps in the REF.** All model-observation comparisons carry uncertainties from the reference datasets and models, the former of which deserve attention. Where possible, comparing results against multiple reference datasets (e.g. ERA5 and JRA-55) helps assess sensitivity to the choice of observational product (Brands 2022*;* Jain et al. 2023). For under-observed regions like West/Central Africa, prioritize regional diagnostics (e.g., Sahel precipitation biases, Gulf of Guinea upwelling) to reveal model dependencies missed by global metrics.

7. **Consider model dependencies and shared components.** Nominally different models often share code or components, meaning they are not fully independent (Knutti et al. 2013; Kuma et al. 2023). Where resources permit, accounting for model dependencies and shared components strengthens the robustness of ensemble-based analyses. When resources are limited and model dependencies cannot be assessed, institutional democracy (one model per institution) offers a practical default (however, there might be project specific reasons to have multiple models from an single institution represented) (Knutti et al. 2010; Boé 2018).

8. **Recognize that the REF currently analyzes only single ensemble members.** A full ensemble will likely increase the spread and uncertainty of a given metric. As a result, REF outputs do not fully represent ensemble uncertainty. Metrics sensitive to variability or extremes may exhibit a narrower spread than would be obtained from a multi-member ensemble for each model. Where a model's internal variability or structural uncertainty is important for an intended application, REF results can be complemented with ensemble-based analysis conducted outside the REF.

9. **Communicate uncertainty ranges rather than deterministic results.** Presenting information from selected models alongside uncertainty ranges and characterizations of ensemble spread, rather than single-point estimates or conclusive scores, gives a more transparent representation of the state of knowledge.

10. **Document methodology completely and transparently.** Clearly recording and communicating all choices made when using REF outputs, including which diagnostics were selected for one's purposes, which and why certain models were selected for downstream use, and what limitations apply, supports reproducibility and enables others to assess the basis for conclusions.

User Checklist

The following checklist can help users confirm that REF outputs are applied in an informed and rigorous manner. These items are drawn from the considerations above and reflect the collective experience of the REF development community.

Foundations

- ✓ Use the REF as a starting point for making informed decisions about model fitness, not as a final determination.
- ✓ Examine multiple complementary diagnostics and metrics rather than relying on a single indicator.
- ✓ Understand the experimental protocol the models were run with (i.e., historical, AMIP, OMIP, etc.).

Uncertainty and Context

- ✓ Check whether model errors are sensitive to the choice of observational or reanalysis reference dataset (e.g., by comparing results against ERA5 and JRA-55).
- ✓ Check whether model errors are sensitive to the effect of internal variability by evaluating multiple realization (or ensemble members) of the considered model and experiment (r1..., r2..., etc. from the historical experiment).
- ✓ Understand the relevance (for intended uses) of selected metrics for the global climate system or, in the case of regional diagnostics, for the climate in the region of interest (e.g., the PNA and NAO and their impact on North American and European climate).
- ✓ Understand the sensitivity of model evaluation results to the selection of metrics and criteria.

Selection and Diversity

- ✓ Confirm that the chosen diagnostics are well resolved by the ESMs under consideration.
- ✓ Maintain ensemble spread and multiple model families after selection.
- ✓ Acknowledge trade-offs inherent in different models' performance (e.g. a model that represents precipitation well might have limitations when representing sea ice).

Transparency and Communication

- ✓ Document clearly the reasons for every choice or decision made based on timestamped REF outputs.
- ✓ Communicate the scope, limitations, and appropriate interpretation of REF outputs when using or citing them in assessments, synthesis, presentations, or peer-reviewed publications.
- ✓ Present results with uncertainty ranges and ensemble spread rather than as deterministic single-model outputs.

A living Framework

The REF is expected to evolve, beyond the timeframe of the CMIP7 Assessment Fast Track, as new diagnostics, models, datasets, and methods become available. Limitations in the REF identified by users may also be addressed through future development of the underlying benchmarking tools and governance processes. Users are encouraged to consider the REF as a living and evolving framework rather than a static product.

References and Suggested Reading

Benestad, Rasmus, Erasmo Buonomo, José Manuel Gutiérrez, and al. et. 2022. Guidance for EURO-CORDEX Climate Projections Data Use. Community document. CORDEX community. https://euro-cordex.net/imperia/md/content/csc/cordex/guidance_for_euro-cordex_climate_projections_data_use__2021-02_1_.pdf.

Boé, Julien. 2018. "Interdependency in Multimodel Climate Projections: Component Replication and Result Similarity." Geophysical Research Letters 45 (6): 2771–79. https://doi.org/10.1002/2017GL076829.

Brands, Swen. 2022. "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 (4): 1375–411. https://doi.org/10.5194/gmd-15-1375-2022.

Eyring, V., Cox, P. M., Flato, G. M., Gleckler, P. J., Abramowitz, G., Caldwell, P., ... & Williamson, M. S. (2019). Taking climate model evaluation to the next level. Nature Climate Change, 9(2), 102-110.

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

Hoffman, Forrest M., Birgit Hassler, Ranjini Swaminathan, et al. 2026. "Rapid Evaluation Framework for the CMIP7 Assessment Fast Track." EGUsphere, July 11, 1–57. https://doi.org/10.5194/egusphere-2025-2685.

Jain, Shipra, Adam A. Scaife, Theodore G. Shepherd, et al. 2023. "Importance of Internal Variability for Climate Model Assessment." Npj Climate and Atmospheric Science 6 (1): 68. https://doi.org/10.1038/s41612-023-00389-0.

Knutti, Reto, Reinhard Furrer, Claudia Tebaldi, Jan Cermak, and Gerald A. Meehl. 2010. "Challenges in Combining Projections from Multiple Climate Models." Journal of Climate 23 (10): 2739–58. https://doi.org/10.1175/2009JCLI3361.1.

Knutti, Reto, David Masson, and Andrew Gettelman. 2013. "Climate Model Genealogy: Generation CMIP5 and How We Got There." Geophysical Research Letters 40 (6): 1194–99. https://doi.org/10.1002/grl.50256.

Kreienkamp, Frank, Heike Huebener, Carsten Linke, and Arne Spekat. 2012. "Good Practice for the Usage of Climate Model Simulation Results - a Discussion Paper." Environmental Systems Research 1 (1): 9. https://doi.org/10.1186/2193-2697-1-9.

Kuma, Peter, Frida A.-M. Bender, and Aiden R. Jönsson. 2023. "Climate Model Code Genealogy and Its Relation to Climate Feedbacks and Sensitivity." Journal of Advances in Modeling Earth Systems 15 (7): e2022MS003588. https://doi.org/10.1029/2022MS003588.

McSweeney, C. F., R. G. Jones, R. W. Lee, and D. P. Rowell. 2015. "Selecting CMIP5 GCMs for Downscaling over Multiple Regions." Climate Dynamics 44 (11): 3237–60. https://doi.org/10.1007/s00382-014-2418-8.

Merrifield, Anna L., Lukas Brunner, Ruth Lorenz, Vincent Humphrey, and Reto Knutti. 2023. "Climate Model Selection by Independence, Performance, and Spread (ClimSIPS v1.0.1) for Regional Applications." Geoscientific Model Development 16 (16): 4715–47. https://doi.org/10.5194/gmd-16-4715-2023.

Räisänen, Jouni. 2007. "How Reliable Are Climate Models?" Tellus A 59 (1): 2–29. https://doi.org/10.1111/j.1600-0870.2006.00211.x.

Shukla, J., T. DelSole, M. Fennessy, J. Kinter, and D. Paolino. 2006. "Climate Model Fidelity and Projections of Climate Change." Geophysical Research Letters 33 (7). https://doi.org/10.1029/2005GL025579.

Sobolowski, Stefan, Samuel Somot, Jesus Fernandez, et al. 2025. "GCM Selection & Ensemble Design: Best Practices and Recommendations from the EURO-CORDEX Community." Bulletin of the American Meteorological Society. Bulletin of the American Meteorological Society (aop). https://doi.org/10.1175/BAMS-D-23-0189.1.

Taylor, Karl E. 2001. "Summarizing Multiple Aspects of Model Performance in a Single Diagram." Journal of Geophysical Research: Atmospheres 106 (D7): 7183–92. https://doi.org/10.1029/2000JD900719.

Addendum: Terminology

The following terms are used throughout this document with specific meanings in the context of the REF and climate model evaluation.

Benchmarking: the process where model simulations are evaluated with observations, reanalysis data or with other models often resulting in a statement made about the "goodness" of the simulation or model based on a predetermined set of standards or criteria (e.g., observations or other standards). The evaluation process normally occurs whenever new schemes are added to the model, but model benchmarking occurs only after all the pieces are assembled and climate simulations are produced (Hassler, Hoffman et al. 2026).

Diagnostic: A comparison of model outputs (individual variables or combinations of variables) with either a reference dataset or an intercomparison across models of a model variable or some combination of model variables. A diagnostic may also represent an evaluation of a relationship between multiple model variables and/or multiple reference datasets (i.e., relationship diagnostics). Each diagnostic comprises one or more model performance metrics (Hoffman et al. 2026).

Evaluation: the process of assessing simulations against one or more observational data sets. The necessity for observations means evaluation can only be done for the historical period, and only for variables or processes for which observations or reanalysis data are available. Model evaluation can be done for a single model or in a multi-model context. Incomplete observational records, including limited time series length, unobserved variables, biases due to specific instruments, and uncertainties in spatial and temporal coverage can make evaluation challenging for certain processes and realms of the climate system that are under-observed (Hassler, Hoffman et al. 2026).

Fidelity (aka Performance): A quantitative assessment of the degree to which model output corresponds to reference data in aggregate. One approach for deriving a fidelity metric is to aggregate relevant scores (Hoffman et al. 2026). The term "performance" is often used as a synonym for "fidelity" (Shukla et al. 2006; Taylor 2001).

Fitness-for-Purpose: The degree to which a model is suited to a specific application, research question, region, variable or temporal scale. Fitness-for-purpose is always relative to an intended use and cannot be reduced to a single universal ranking.

Metric: 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), and spatial or temporal correlations (Taylor 2001). 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 (Hoffman et al. 2026).

Metric Scalar: The numerical output resulting from the calculation of a performance metric (e.g., a calculated bias value) (Hoffman et al. 2026).

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 data, 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 (Hoffman et al. 2026).

(Process) Validation: The process of determining how well a model represents processes in the real-world, particularly for the intended uses of the model. Process Validation can include a broad range of aspects, from ensuring correct units and signs pf the data produced, to the interactions between model components or variables and process representations (Hassler, Hoffman et al. 2026). REF benchmarking represents an initial step toward validation but does not constitute a full validation exercise on its own.