

June 16, 2026

Editors for *Geoscientific Model Development*

Dear Editors:

Revised Manuscript for Paper EGUSPHERE-2025-2685

My co-authors and I are pleased to finally submit this revised manuscript entitled, “Rapid Evaluation Framework for the CMIP7 Assessment Fast Track,” paper EGUSPHERE-2025-2685. We apologize for the multiple requests we made for extensions in providing referee responses and this revised manuscript. As a community development activity, the Rapid Evaluation Framework (REF) has evolved since it was originally conceptualized. Moreover, the underlying diagnostics packages and datasets have also changed and improved over the course of the time we have revised the manuscript, necessitating significant changes in the details provided in Tables 1, A1, and B1. We are exceedingly grateful for your patience and understanding as our large group of internationally diverse authors have reached the point where we can provide a “snapshot” of the configuration of what we expect will be a continually improving research community resource.

Attached to this letter is a document containing author responses to all referee comments. In addition, we have uploaded a clean version of the revised manuscript and a version of the manuscript that tracks changes made since the original submission. Because information in the four tables has changed significantly, and because the columns and orientation were changed in some cases, we have appended onto the track changes version of the manuscript a document that shows the pages containing all four tables in their original forms and in their revised forms. We hope by doing so that editors and referees will be easily able to identify any changes they wish to explore.

We look forward to continuing to work with you to get this manuscript accepted for publication. Thank you for your kind assistance and attention to this paper.

Sincerely,



Forrest M. Hoffman

Corporate Fellow and Computational Earth System Scientist

Group Leader, ORNL Integrated Computational Earth Sciences (ICES) Group

Joint Faculty Professor, Civil & Environmental Engineering, University of Tennessee, Knoxville

Fellow, American Association for the Advancement of Science (AAAS)

Senior Member, Institute of Electrical and Electronics Engineers (IEEE)

Response to Reviewer Comments for Manuscript EGUSPHERE-2025-2685

April 7, 2026

In this document, reviewer comments are shown in plain text, author responses are shown in bold text, and quoted/modified writing extracted from the revised manuscript accompanying author responses is shown in indented and italicized bold text.

We, the authors, thank Anonymous Referee #1 for the thoughtful comments, which identified areas that required additional clarity, and for the minor edits, which assisted us in making the article more readable. The manuscript was improved as a result of revisions in response to this helpful review. Responses to each point of the review are interspersed with the review comments below.

Section 4.3: I'm not exactly sure of the difference between the 'variables' described here and the 'diagnostics' referred to in Table 1. For example, all the diagnostics in the Land and Land Ice Realm are objects I would think of as variables output by a land model (GPP, LAI etc). I do see how some of the diagnostics (like ECS/TCR etc in the ES realm) would be computed from variables and not be variables themselves. Perhaps it would be good to clarify that there might or might not be duplication/cross-over between 'diagnostics' and 'variables'.

The definitions of Model Variables and Diagnostics provided in Section 1 distinguish the two terms from each other. In particular, some diagnostics employ a single variable, like GPP or LAI, while others, like Precipitation minus Evaporation (P-E), require multiple variables. However, to better clarify this point, the first sentence of the second paragraph of Section 4.3 was modified, and a second sentence was added as follows.

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 modelling 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, others require two or more model variables, as described in the definitions in Section 1.

Section 5.3: this is more of a philosophical thought about subsetting models for particular scientific purposes and/or regional downscaling – for an end user, it would be tempting to just grab the models that perform well on certain regions/realms of interest, but if those same models are way off in other regions/realms, would that be a fair way to subset models? Perhaps this is more of a question for the community as a whole to come up with the best way of pulling a

subset of models to look at specific questions (e.g. have some kind of penalty for not doing well for other regions/variables).

This is an important consideration, and to assist analysts in making such decisions, most of the REF diagnostics include multiple metrics and graphical output, like contour maps and Taylor diagrams, instead of only a single scalar score that can mask such details. To be more explicit, the following sentence was added to the paragraph in Section 5.3.

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.

Table B1: (i) I am curious why cSoil was included but not cVeg? Is there a plan to include more datasets such as those from GEDI? (ii) would it be possible to add the dates over which the reference datasets are available? I am familiar with the ILAMB land datasets and I think they only go up to 2014 – have there been extensions to these for more recent years?

(i) As explained in Section 2.2, the diagnostics and therefore, the corresponding variables included in the initial version of the REF were prioritized through a community survey process. The objective was to have the community identify the top four-to-seven diagnostics to include for each of the five Earth system realms. The results of the community survey identified the soil carbon (cSoil) diagnostic as having the highest priority for the Land & Land Ice realm. Total carbon in vegetation or above- plus belowground biomass (cVeg) was not prioritized in the top seven diagnostics for Land & Land Ice. However, adding this diagnostic could be easily accomplished in a future version of the REF since the ILAMB package already includes a biomass diagnostic. Plans for future versions also include bringing in more reference datasets, including newer remote sensing products.

(ii) Reference data used in the REF span multiple time ranges, and the expectation is that these data will be regularly updated by extending their time ranges as appropriate or by replacing them with improved data sets as they become available and evaluated by the research community. In Table A1, we will add time ranges for reference data included in the first version of the REF.

Section C1 1.1: what is the frequency of this calculation? i.e. annual/decadal/something else?

Both the Antarctic annual mean and Arctic September rate of sea ice area loss per degree of warming are calculated using a single value each year. The description Section C1 1.1 was modified to be explicit as follows.

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; both are regressed against annual global mean temperature.

Minor Edits

Line 114: “produced, processes” -> “produced, and processes”

Corrected.

Line 117: what is a “realm” referring to here? These are defined in line 295 but should be introduced earlier here.

Definition was moved and introduced at this earlier location.

Figure 1 caption: spell out the acronyms (e.g. I am not sure what ‘DAG’ is)

Abbreviations and acronyms are now defined in the figure caption.

Line 161: “a” -> “an”

Corrected.

Line 175: “at” -> “by the”

This was intended to precede a citation, which has now been corrected.

Line 259 -> add comma between “(ENSO)” and “CLIVAR”

Corrected.

Line 734: ref should be in parentheses

Corrected.

Line 743: insert “by” after “influenced”

Corrected.

Line 773: insert “by” after “obtained”

Corrected.

Line 780: I am not sure what is meant by “brought direction”

Corrected to “broad direction”.

Line 792: “patterns” -> “pattern”

Corrected.

Line 802: insert “in” after “patterns”

Corrected.

'Comment on egusphere-2025-2685', Anonymous Referee #2, 17 Oct 2025

We, the authors, thank Anonymous Referee #2 for the helpful comments that identified the need for improved descriptions of a few elements of the REF. We also appreciate the suggested minor edits. The manuscript was improved as a result of revisions in response to this review. Responses to each point of the review are interspersed with the review comments below.

This paper presents a newly developed tool to enable a rapid assessment of simulations produced in the context of CMIP7. The primary objective is to provide a minimum set of diagnostics on model simulations published for the AR7 Fast Track intercomparison exercise. Such an assessment will be of great value for CMIP model data users, particularly those involved in climate services. However, the scope of the tool is considerably broader. It is intended for use in a wider range of contexts - for example, by model developers during the model development phase - and could be extended to other contexts: future CMIP phases, regional intercomparison such as CORDEX, ISIMIP, etc... The tool appears to have been designed in a safe conceptual basis, with a clear separation of concerns, adaptability to diverse computing environments and ease of integration for new diagnostics. Furthermore, the paper outlines several promising perspectives for extending its capabilities.

This tool is undoubtedly of great interest for the community and the authors should be acknowledged for this significant community effort. This tool's focus on interoperability with existing metrics packages represents a particularly valuable contribution to the modelling community, and this aspect could be further emphasized in the paper.

Thank you for the comment and for recognizing the importance of designing the REF to be interoperable with existing evaluation packages. Indeed, the aspect of interoperability among community efforts could be better emphasized. The abstract has been slightly extended as follows to strengthen this point.

Production of reproducible diagnostics and community-based assessments are the key features of the REF that help to inform mitigation and adaptation policies.

Furthermore, providing interoperability with existing evaluation packages assures that contributions from previous community efforts will be available for use in future model intercomparison projects.

Such a tool is desirable; however, it should also be emphasized that its adoption may change the paradigm of multi-model intercomparison exercises. Indeed, once modellers are aware of which metrics users will focus on to select models, it is almost inevitable that they will concentrate their efforts on improving model performance against those specific metrics at the expense of other processes. This dynamic needs to be explicitly acknowledged. However, modellers must avoid excessive model tuning (“overtuning”).

This comment gives us the opportunity to stress once again that the community development of a standardized set of metrics and diagnostics does not imply a classification of “better” and “worse” models. The point here is to allow analysts to objectively characterize the strengths and weaknesses of models, so that modelers can work on improving the model where possible, and analysts can identify which models are more appropriate than others for the study of specific science questions. To address this question for the reader, we have added the following sentence in Section 5.3.

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.

It is therefore crucial that the package is overly clear on the characterisation of the target metric uncertainty. For mean climate aspects, this target metric uncertainty primarily depends on observational uncertainty. Although the paper briefly discusses observational uncertainty, it mainly approaches it as the uncertainty associated with individual datasets. In my view, the uncertainty of single observational products does not fully capture the range of observational uncertainty. A more robust estimate could be obtained by considering an ensemble of different observational products. For trend estimates, another important source of target uncertainty arises from internal variability. In future versions of the tool, I strongly recommend placing greater emphasis on the quantification and communication of this target metric uncertainty.

We thank the referee for this thoughtful suggestion. Obtaining quantified uncertainties for individual reference datasets is difficult; however, in the initial version of the REF, we are including at least one diagnostic that includes a quantified uncertainty. In exploring observational uncertainties, we have identified a variety of sources of uncertainty, most of which are difficult to constrain, even for a single variable. Future versions of the REF are likely to incorporate multiple reference datasets for a given variable, as is already common in some of the community packages included in the REF, offering another method for understanding uncertainty for a single variable. We agree that the REF should further evolve into developing more robust estimates of uncertainties from a wide variety of reference data and visualize these uncertainties alongside the reference data and model

results. The REF already permits multiple initial condition realizations to be analyzed as a way of quantifying internal model variability.

I would also recommend clarifying the types of simulations on which the REF focuses. My understanding is that it will primarily rely on historical-type simulations. However, since ECS and TCR are among the target metrics, this implies that academic simulations such as 1% CO₂ and abrupt4xCO₂ experiments are also expected. When modellers use the tool for model development, is it straightforward to select which simulations are to be evaluated? Additionally, is there a mechanism to ensure consistency between the simulation setup and the metric being calculated?

The list of simulations relevant to each of the diagnostics in the REF is now provided in Appendix B1. When modelers use the REF for model development, they can simply identify the model outputs that are to be evaluated. The REF uses the metadata provided in the CMORized netCDF files to ensure consistency between simulations and the metrics being calculated.

Specific questions:

Abstract

Sentence “For instance, the REF will facilitate improvements to models and reductions in uncertainties for projections since ESMs are the main tool for studying the global Earth system.”

I am not convinced that the REF will significantly reduce uncertainties in ESM projections. On the centennial timescale, uncertainties are primarily associated with model uncertainties and scenario uncertainties. In the case of model-related uncertainties, it has not been demonstrated that they can be substantially reduced through model selection. It may be more appropriate to rephrase the statement to suggest that the REF could contribute to enhancing the confidence in model projections and selection.

This is an excellent point, and one that many of the authors have experienced first hand. We have changed the sentence in the abstract as follows.

For instance, the REF will facilitate improvements to models and enhance the confidence in model projections through process-based selection of models based on their performance with respect to observations.

Introduction

In the first sentence, ESM are stated as the primary tools to study interactions between the atmosphere, biosphere,... It is correct as long as centennial to millennium time-scales are considered but probably not for longer time-scale. I would be worth indicated a time range.

We agree that running ESMs over paleoclimatic timescales is computationally difficult and that mesoscale models are excellent for higher resolutions over short timescales. We changed the first sentence of the introduction as follows.

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 in the Earth system and how it responds to human-induced and natural forcings

Section 2.1

It is stated that “Releasing the REF before model outputs are submitted to ESGF offers modelling centres the opportunity to systematically assess their models and to make targeted improvements during the development phase” (ln 129-131)

This is certainly the case, but at the same time, it will most likely change the paradigm of model development for CMIP exercises. It is hardly avoidable that models will tend to match the REF metrics, leading to a reduction of ensemble errors on these specific metrics. However, does this necessarily imply a genuine reduction in uncertainty? Could it occur for undesirable reasons? Since the REF includes metrics on model trends, models will likely converge toward similar historical trends and TCR values. At the same time, estimations of historical trends are influenced by internal variability, while accurately assessing real-world internal variability remains a challenge and is often based on model-derived estimates (Simpson et al., 2025; Gyuleva et al., 2025). Is it truly desirable for models to converge on observed historical trends? Similarly, TCR remains an unknown parameter—does this approach risk pushing all models toward a TCR within the medium range of previous CMIP exercises? Is that desirable? While I recognize the necessity of the REF tool, I believe this paper should highlight that it will, to some extent, reshape the model development paradigm. This potential impact should be carefully considered when analyzing future model ensembles.

These are all excellent questions, many of which our authors have discussed. Since modeling centers have access to evaluation tools, datasets, and model output and assessments from CMIP6 and prior activities, they already routinely use that information to evaluate their new models. The existence of the REF does not change the potential for the undesirable social effects of model convergence. The beauty of the REF is that these sorts of benchmarking activities will be consistent, made centrally available for potential model users, and employ regularly updated and curated observational reference data. As noted above, the breadth of the diagnostics and metrics included in the REF will preclude large-scale convergence across model components and realms. Overtuning for a few related variables will almost certainly produce poor performance for other variables. To be more explicit on these points, we added the following text in Section 5.3.

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.

Section 3

Three primary potential applications are stated (In 196), but the text is not that clear on which are these three applications. It would help the readability if they were clearly stated here and then developed.

We agree. We restructured the text introducing Section 3 as follows.

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.

Section 4

Four open-source packages have been used as a starting point, yet no explanation is provided regarding the criteria for selecting these particular packages. Even if the choice was made purely for practical reasons, it would be valuable to include this information.

Excellent suggestion. We have modified the introductory sentence of Section 4.1 as follows.

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.

I consider CMEC to be a major development within the REF tool. I suggest emphasizing this aspect more prominently in the paper, particularly in the abstract. As noted, there exists a wide diversity of benchmarking packages in the community, but they often lack interoperability and can be challenging to install or implement. Consequently, CMEC has the potential to play a key role in establishing standards and facilitating their use for modelling centers.

We agree that CMEC has provided a key technological framework for combining disparate benchmarking packages together into a cohesive system. We have modified a sentence in the abstract as follows.

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.

Figure 2 is a bit unclear to me, the relationship between redis and database is particularly indirect. Why isn't there any arrow leaving the redis block?

Unfortunately, the reviewed manuscript included an early draft of Figure 2. It has been replaced with a more complete and accurate figure of the system design.

Miscellaneous questions/remarks:

In 68-69: Model variables are described as changing during the model execution. In that case, what about fixed fields? Probably there are not metrics scrutinized by the REF but they could be part of a diagnostic calculation. Do you consider that they are still "Variables" or do you have another another term to call them?

Diagnostics targeting fixed quantities and model parameters have not been considered. We think that comparisons of fixed quantities (e.g., topography, bathymetry, constant astronomical factors) should be treated as a matter for model design intercomparison, rather than for the evaluation of model performance.

In 140-141 : the sentence is a bit weird, consider rephrasing.

We agree. We restructured the sentence as follows.

These conventions ensure that data are consistently structured, with uniform variable naming, units, and metadata, facilitating easy comparison across model outputs.

In 207: "of the REF is that it be usable by" consider rephrasing.

The sentence has been restructured as follows.

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

In 237 : I have not seen CMEC defined before in the text.

Agreed. We have now spelled out Coordinated Model Evaluation Capabilities (CMEC).

In 275: Could you explain a bit what is "a functional relationship metric"

Certainly we can. We have simplified the identified sentence and added a second sentence that provides an illustrative example as follows.

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.

Gyuleva, G., Knutti, R., and Sippel, S.: Combination of Internal Variability and Forced Response Reconciles Observed 2023–2024 Warming, Geophysical Research Letters, 52, e2025GL115270, <https://doi.org/10.1029/2025GL115270>, 2025.

Simson et al., Confronting Earth System Model trends with observations | Science Advances: <https://www-science-org.insu.bib.cnrs.fr/doi/10.1126/sciadv.adt8035>, last access: 16 May 2025.

Citation: <https://doi.org/10.5194/egusphere-2025-2685-RC2>