Response to Reviewers

Geoscientific Model Development

Manuscript:	egusphere-2024-1518
Title:	Monitoring and benchmarking Earth system model simulations with ESMValTool v2.12.0
Authors:	Axel Lauer, Lisa Bock, Birgit Hassler, Patrick Jöckel, Lukas Ruhe, Manuel Schlund
Date:	19 September 2024

---

*We thank the two reviewers for their helpful comments. In this document, we answer each point raised by the reviewers. The original reviewers' comments are given in **black**, our answers in **blue**. A "track changes" version of the revised manuscript highlighting all changes is available.*

Reviewer #1

This paper is documenting the recent updates of one of climate and Earth system model evaluation software package, named ESMValTool. Considering the popularity of the tool, it is important to keep the capabilities updated and well-documented as a community resource. I think the paper is well organized. I only have a few minor comments as follows.

We thank Reviewer #1 for providing helpful comments to improve the manuscript.

I wonder if authors could clarify more explicitly what are the new capabilities in this specific version of the tool, compared to the previous version with the published paper. Are those metrics in section 2.2 all new metrics that were not available in the previous version?

In the previous version, only 'bias' was available as a preprocessing function that can be applied to an ensemble of models. Correlation and RMSE were only calculated within selected diagnostics, which did not allow generic application to arbitrary quantities, dimensions (time, longitude latitude, level), geographical regions, etc. A visual comparison of the metric results among an ensemble of models was therefore not possible with previous versions. The metric Earth mover's distance has been newly added to ESMValTool v2.12.0. We made this clearer by adding the following paragraph to section 2.2:

"The metrics have been implemented as generic preprocessing functions that are newly available in v2.12.0. In contrast to previously available diagnostic-specific implementations of such metrics, the preprocessing functions can be applied to ensembles of models and arbitrary variables and dimensions providing the flexibility needed for the new benchmarking and monitoring capabilities of ESMValTool described here."

Line 44 to 47 "For this, for example results from the Coupled Model Intercomparison Phase 5 and 6 (Eyring et al., 2016; Taylor et al., 2012) can be used to get an overview of which biases can be

considered “acceptable for now”, and which would need more attention and more detailed analysis and comparisons with observations.”: As there have been several tools being developed for such purposes, I wonder if it would be beneficial to provide a few references as examples: e.g., PCMDI Metrics Package (Lee et al., 2024, GMD), ILAMB (with proper reference), etc.

The reviewer has a good point. We added references to PCMDI Metric Package and ILAMB to the introduction:

”A number of software tools for model evaluation has been developed over the recent years. Examples include, for instance, the PCMDI Metrics Package (PMP, Lee et al. (2024)), the International Land Model Benchmarking (ILAMB) system (Collier et al., 2018), or the Earth System Model Evaluation Tool (ESMValTool, Righi et al. (2020)).”

Line 88 “For all metrics, an unweighted and weighted version exists” and sections 2.2.2 through 2.2.4: I wonder what the rationale was to include unweighted metrics. While it is fair to include both methods as options, I think weighted metrics might better considered for the “default” method. Is there any practical use case of unweighted metrics?

We agree with the reviewer that applications of weighted metrics are by far the most common use case. We implemented also the option to calculate unweighted metrics for example for application to station data, for which individual model grid cells are selected that contain the measurement station. In this case, weighting with the gridbox area instead of giving each station pixel the same weight might distort results. We added the following sentence to section 2.2:

“While the weighted version is the preferred option for most use cases, an unweighted option is available for cases where weighing with the gridbox area might distort the results. Examples of such cases include, for instance, extracting individual model grid cells containing a measurement station and giving the same weight to each station, independent of the model gridbox area.”

Line 149 “The default value in ESMValTool is n=100”: Does the bin distributed equally and sized evenly by min/max of the PDF? I guess this might be the case but it won’t be harmful to clarify it.

Yes, the reviewer is right, the bins are distributed automatically. We clarified this by adding “equally sized” to the corresponding sentence.

Line 155 “2.3.1 Observation datasets”: I think the word “reference” might better inclusively represent datasets listed in the section. Some reanalysis datasets were discussed along, but often they are preferred to be differentiated from instrument-based observation.

Agreed, we changed “observational data” to “reference data” as suggested.

Line 170 “(Ecmwf, 2000)”: Please capitalize all letters for ECMWF.

Thanks for spotting this. This is done automatically by the EndNote style “Copernicus_Publications” for Word and will hopefully be fixed during type setting.

Line 557: The ERA5 reference is placed where it is not in right alphabetical order in the reference list.

This is also caused by the formatting of the EndNote style “Copernicus_Publications” omitting the authors of this website “Copernicus Climate Change Service (C3S)”. This will hopefully also be fixed during type setting.