Answers to Reviews regarding manuscript “Developing Guidelines for Working with Multi-Model-Ensembles”

# Reviewer 2

General Remarks:
First of all, I want to congratulate the authors on a very comprehensive treatise on the past, present, and future of Earth System Modeling. An enormous amount of ground is being covered, and it is clear that a lot of hard work has been put in thus far. Based on the scope, which is more or less much of the climate model diagnostic research done in the past 35 years, it is clear that this will be a long review paper. With reaching readers in mind, I think it is important to distill things down in many places. It is very difficult to effectively review a paper with this much information in it. The following topics are covered in the introduction alone:

- History of Earth System Modeling
- Components of an ESM
- Complexity and parameterizations in ESMs
- Why multi-model ensembles?
- History of model intercomparison and CMIP
- MME design
- Other types of model ensembles

Some of the distillation can be achieved with supplementary material. I've noted down sections I think could move to an appendix to pace things up a bit, without completely discarding anyone's contribution. Additionally, the authors might consider synthesizing some information into tables or graphics that users can quickly reference (a good example of a summary table is given in Simpson et al. 2025). I have also made an attempt at a recommended restructuring, identifying sections that could be merged. Please take what is useful and leave the rest. Ultimately, you know the paper best.

We thank the reviewer for their time and thoughtful evaluation of our manuscript. We appreciate the constructive suggestions to improve clarity and focus and have revised the manuscript accordingly, including substantial efforts to harmonize and sharpen the text. We moved material that contributes less to the key goal of the paper to the appendix and also condensed remaining sections substantially. Below, we respond point by point to the reviewer’s comments. We believe that this resulted in a much clearer revised version of the manuscript.

There are only a handful of places that I feel warrant further discussion. In the introduction, expanding on how it has been determined that multi-model ensembles (MMEs) outperform single models would be a useful addition.

As the motivation to use MMEs is key for this study, we agree that expanding this paragraph is a good proposition and revised it accordingly:

“Besides the possibility to quantify uncertainty and increase robustness, MMEs have been found to generally outperform projections from individual models. Within the weather forecasting community, numerous studies have shown that ensemble predictions are more reliable than individual predictions (Doblas-Reyes et al., 2003; Krishnamurti et al., 1999), e.g. the north american multi model ensemble showed improvements in various skill metrics (correlation, RMSE, RPSS, and reliability) compared to individual models used before (Kirtman et al 2014). Inspired by these findings, studies in the climate context also analyzed the potential benefits from working with MMEs for projections. In climate model evaluation, the MME projections have proven to outperform individual model projections in numerous studies e.g. regarding the mean (Gleckler et al., 2008; Knutti et al., 2010a; Lambert and Boer, 2001; Palmer et al., 2005; Phillips and Gleckler, 2006; Pincus et al., 2008; Reichler and Kim, 2008) and variability (Zhang et al., 2007), further strengthening the motivation to use MMEs. The enhancement of the signal and cancellation of errors contribute to these advantages compared to individual models (X). Becker et al. (2022) highlight the practical advantage of the continuous operation of MMEs, which can be maintained even when individual modeling centers are temporarily unable to contribute, for example due to technical or political constraints. They further provide an example where the use of an

MME enabled the identification of outlier behavior in ENSO predictions, which could subsequently be traced back to previously unknown deficiencies in the underlying reanalysis dataset. Furthermore, an ensemble approach reduces the risk of selecting a model outlier with particularly large biases. Given these benefits, MME projections have become an established tool for climate studies addressing a broad range of research questions, also being the standard method to analyze and present results in the Assessment Reports (ARs) of the Intergovernmental Panel on Climate Change (IPCC), where the state-of-the-art knowledge on climate change is reviewed. For researchers, MMEs provide an efficient way to get an overview of general tendencies for specific questions. Also for non-experts, presenting results in a synthesised format as e.g. in the context of MME also facilitates accessibility and interpretation (Knutti et al., 2010a), underlining the benefits of MMEs for the users. "

Also missing is a discussion of CMIP as an ensemble of opportunity (e.g., Tebaldi and Knutti, 2007, Sanderson et al. 2012, Merrifield et al., 2023). CMIP is not explicitly designed, as a whole ensemble, to provide an estimate of robustness, for example.

Thank you for this important suggestion. We have pointed this out in the revised paragraph as follows:

"It is important to recognize that CMIP constitutes an "ensemble of opportunity" (Tebaldi and Knutti, 2007; Sanderson et al., 2012; Merrifield et al., 2023), as it reflects the collection of readily available simulations rather than a systematically designed sample. Contributing institutions range from long-established, well-resourced climate modeling centers to newer groups with sufficient computational resources to run adapted versions of existing models. While this inclusivity broadens participation, such ensembles of opportunity are not designed to constitute a statistically representative sample of multi-model uncertainty (Merrifield et al., 2023)."

And finally, if possible, please deputize a single author to read through the whole manuscript and harmonize contributions. "Multi-model ensembles (MMEs)" is defined at least twice (pg.3 L70, pg.4 L11), as is "Multi-Model Large Ensemble Archive (MMLEA)" (pg.46 L53, pg.47 L86). Sections are previewed multiple times back-to-back. There are several instances where things are discussed prior to being introduced as well.

This was indeed necessary and has been addressed in the revised manuscript. A single author reviewed the entire manuscript to harmonize terminology, structure, and cross-references, eliminating redundancies and ensuring that concepts are introduced consistently and in the appropriate order.

Specific Comments:

Pg. 1 L21: "… a key tool"

Corrected.

Pg. 2 L48: "Concurrently, the volume of ESM simulation output data…"

Corrected.

Pgs. 2-3 L50-59: This is a candidate for adaptation into a table or graphic.

Thank you for this suggestion. There already exist figures that illustrate various components of ESMs and also figures which show differing characteristic spatial and temporal scales of various atmospheric and oceanic phenomena are pretty common in educational academic textbooks and scientific literature. Unfortunately we currently don't have the capacity to produce a qualitative and novel graphic out of this entire paragraph, which is also difficult to fit into the table, therefore we decided to retain the text format.

Pg. 3 L74-77: I think this is under-appreciated and worth elaborating on. How does the MME outperform an individual model? Are there cases where all the models are wrong together?

Regarding first part of the question, see answer above. Regarding second part: we understand that it refers to cases where individual models outperform the MME. We think such cases could occur in cases where the smoothing effect of the mean is counterproductive, e.g. when an individual model captures a dynamic process particularly well, while the majority of models do not. Also in regional cases, when the resolution of the majority of models in a MME are not able to resolve mountain in sufficient detail, individual models with finer native grid would outperform the associate MME. And as discussed in the manuscript already, use cases as extremes or related to variability are also to be considered in individual models. Also, if most models share a specific component that leads to certain bias, that would affect the MME while individual models with other components would not be affected. We added these reflections to the main paper:

"At the same time, the superiority of MMEs is not universal. There are cases in which individual models can outperform the ensemble mean, for instance when the averaging inherent to MMEs suppresses dynamically relevant signals that are well represented in only a subset of models. This can occur for specific physical processes, or extremes, where ensemble averaging may smooth physically meaningful variability or dampen circulation-driven responses. Moreover, if most models in an MME share common structural components, parameterizations, or tuning strategies, systematic biases can persist in the ensemble mean. In such cases, individual models with alternative formulations may provide more accurate representations for specific variables, regions, or applications."

Pg. 4 L01: Is this the same AMIP experiment defined and discussed above?

It is fundamentally the same experiment, even though in CMIP6 it is more fully integrated in the intercomparison framework with updated protocols and as part of the broader set of coordinated simulations.

Pg. 4 L05: CO2 subscript

Corrected.

Pg. 4 L06: quotes not needed around definition of SSP

Corrected.

Pg. 5 L29-39: I recommend moving this to an appendix that includes the sections on SMILEs as well.

We thank the reviewer for this proposition. However, as SMILEs are mentioned already before being discussed in detail in section 4.2., we kept this to introduce them and also clarify the scope of our paper.

Section 2.1: It would be good to open with a short explanation of what model evaluation is (e.g., benchmarking aspects of historical model simulations using observations, etc.). It also makes sense to define reanalysis here.

We added explanations on model evaluation and reanalysis as proposed:

“Model evaluation refers to the systematic assessment of climate model simulations against observational reference data in order to compare model performance and identify biases. In practice, this involves benchmarking historical simulations with respect to observed climate statistics, such as mean states, variability, spatial patterns, and relevant physical processes.”

“Reanalysis datasets are physically consistent products produced by assimilating diverse observational data into a numerical weather or climate model. They combine the broad spatial and temporal coverage of models with observational constraints and are therefore widely used as reference datasets”

Pg. 6 L65: You could highlight Sippel et al. 2024 as an example of disadvantages.

Done.

Pg. 6 L80: “Reanalysis” comes in a bit abruptly here; it has not really been defined.

This has been resolved now by adding an explanation on reanalysis data, following the respective previous reviewer comment.

Pg. 7 L84-85: There is a whole section on regridding. Can it be referenced after “This can also be achieved by appropriate regridding methods” instead of discussing it again here?

This is a great proposition. We revised the manuscript accordingly.

Pg 7. L86-89: I would recommend combining all mentions throughout of “model tuning” and “emergent constraints” (they disappear as a consequence of model tuning) into one section to avoid repetition. Including: Pg. 8-9 L20-30, Pg. 22 L96-09

We revised the manuscript such that model tuning and calibration are now discussed in a single, consolidated paragraph within the Model Evaluation section, thereby avoiding repetition and clarifying their role in model assessment. We chose to retain the discussion of emergent constraints in the Uncertainty section, as emergent constraints represent a distinct post-processing method for uncertainty reduction rather than a model development or evaluation technique.

Pg 8. L01-02: Yes, to some extent. It is often thought of in reverse: models that fail to capture features of historical climate are unlikely to capture them in the future.

Thanks for this clarification. We adapted accordingly.

Pg 8. L03: A brief description of what a Taylor diagram is seems like it would be in the spirit of what you are trying to achieve with this paper.

We agree and elaborated the part of Taylor diagram more accordingly.

Pg 9. L34: Fix citation

Done.

Pg. 10 L71: Fix citation

Done.

Pg. 10 L79 - Pg. 12 L36: Regarding “Below, we highlight examples of process-oriented analysis applied to CMIP models.” I recommend picking one example for the main text and moving the rest of the examples to supplementary material.

Thanks for mentioning that. We have moved most examples, as suggested, to the Supplementary section and have also condensed the few remaining examples with regard to their conceptual contributions.

Section 2.2 – This whole section is another great candidate for a summary table or graphic. The written details can then move to supplementary material.

Thanks for this proposition. We moved this section from the main text to the Appendix. Thus, page length of the original paper is also further reduced.

Pg. 12 L43: The ESMValTool has yet to be introduced. I know there is a whole section on tools, which I recommend moving up before the ESMValTool is mentioned. Otherwise, the fact that there are tools, e.g., the ESMValTool, could be mentioned in the introduction, such that this is not the first mention.

This subsection on model bias was moved to the appendix, thus the respective sentence is located after the ESMValTool is introduced in subsection 2.5 in the revised manuscript.

Section 2.3: Model dependence has come further into the mainstream than is credited here (e.g., Kuma et al. 2023, Merrifield et al. 2023). Additionally, I would highlight that the metadata reporting requirements in CMIP6 made comprehensive model dependence assessments possible, a real step forward for the field in transparency and reproducibility.

We thank the reviewer for this valuable perspective and rephrased accordingly, including a statement regarding the metadata: “The metadata reporting requirements introduced in CMIP6 have made comprehensive assessments of model dependence possible, thereby representing a meaningful advance in transparency.”
Specifically, ClimSIPS is noted as a valuable model selection method in section 2.3. The text reads “Recent model selection methods also emphasize model independence (Snyder et al., 2024) with tools being developed that account for model dependence such as ClimSIPS (Merrifield et al., 2023).” Additionally, Kuma et al., 2023 is incorporated into this review as their impressive research on finding model interdependencies is used to create our Figure 2 (with the permission of the authors).

Pg. 15 L33: I would not go as far as saying this is “intriguing or concerning”, it’s a consequence of limited resources (-) and collaboration (+) and has been known for a long time.

Agreed. We rephrased accordingly.

Pg. 16 L54: Wrong figure referenced.

Corrected.

Pg. 23 L15-29: I recommend this moves to the supplementary material.

Done.

Section 2.6: This is a good section length for a paper with this much material in it.

Thanks for the feedback.

Section 3: If you are going for an FAQ, then I would recommend a condensed answer and a reference to more information in the supplement for every one of the subsections in Section 3. No more than a page for each would be my preference.

We thank the reviewer for this proposition and condensed the subsections of section 3 substantially. We also rephrased the introduction of section 3 to clarify the aim and scope of this section.

Pg. 29 L 47: CORDEX is yet to be defined.

Clarified.

Section 3.4 – I recommend this section be combined with model evaluation / model weighting.

We understand the reviewer's suggestion to combine Subsection 3.4 (outliers) with the model evaluation/model weighting section and agree that this would be a reasonable alternative. We have reconsidered this structure carefully. However, this would also apply for other subsections (e.g. moving 3.1 "how many models" to the model evaluation section in 2, etc) We believe that retaining Section 3 as a dedicated section offers several advantages. In particular, it helps avoid further lengthening the already extensive sections in Section 2, where additional material might become less visible to readers. The format in section 3 allows central methodological considerations to be addressed more directly and accessibly. In response to this comment, we have substantially condensed Section 3.4 to improve clarity and reduce redundancy.

Pg. 32 L38-41: Pleased to report outlier models were included in CH2025. See Chapter 2 of the Scientific Report for more information.

Thanks for pointing this out. We added another sentence to clarify this: "*The subsequent CH2025 scenarios included outlier models (MeteoSwiss & ETH 2025).*"

I think Section 4.1 should be its own paper.

We thank the reviewer for her/his perspective. In response, we have moved a substantial part of Section 4.1 to the Appendix, resulting in a significant reduction in length. While we acknowledge that the content of this section could form the basis of a separate paper, we believe that outlining future directions in the development of multi-model ensembles is an integral component of the comprehensive review we aim to provide. We therefore consider it important to retain this section in the manuscript and have not removed it fully.

References

Isla R. Simpson et al., Confronting Earth System Model trends with observations.Sci. Adv.11,eadt8035(2025).DOI:10.1126/sciadv.adt8035

Claudia Tebaldi, Reto Knutti; The use of the multi-model ensemble in probabilistic climate projections. Philos Trans A Math Phys Eng Sci 15 August 2007; 365 (1857): 2053–2075.

Sanderson, B. M., and R. Knutti (2012), On the interpretation of constrained climate model ensembles, Geophys. Res. Lett., 39, L16708, doi:10.1029/2012GL052665.

Merrifield, A. L., Brunner, L., Lorenz, R., Humphrey, V., and Knutti, R.: Climate model Selection by Independence, Performance, and Spread (ClimSIPS v1.0.1) for regional applications, Geosci. Model Dev., 16, 4715–4747, https://doi.org/10.5194/gmd-16-4715-2023, 2023.

Sippel, S., Kent, E.C., Meinshausen, N. et al. Early-twentieth-century cold bias in ocean surface temperature observations. Nature 635, 618–624 (2024). https://doi.org/10.1038/s41586-024-08230-1

Simpson, I. R., K. A. McKinnon, F. V. Davenport, M. Tingley, F. Lehner, A. Al Fahad, and D. Chen, 2021: Emergent Constraints on the Large-Scale Atmospheric Circulation and Regional Hydroclimate: Do They Still Work in CMIP6 and How Much Can They Actually Constrain the Future?. J. Climate, 34, 6355–6377, https://doi.org/10.1175/JCLI-D-21-0055.1.

Finally, we went through this list of references, checked whether they were already included in the original manuscript and incorporated the missing references in respective sections.