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

# Reviewer 1

This is an interesting paper that summarizes MME in CMIP. I would recommend eventual publication subject to revision, particularly on the editorial side, with the aim of providing clearer goals and more focus.

We thank the reviewer for his/her time and the generally positive assessment of our paper. We appreciate the suggestion to improve clarity and focus of the manuscript and revised the manuscript accordingly - including substantial shortening and sharpening efforts. Below, we provide a point-by-point answer to the individual reviewer’s comments.

Major suggestions worth considering:

(1) While this is a very nice, thorough comprehensive review of various methods/ideas, the paper is far too long and needs some heavy trimming. I would ask the authors to dig deep and try to cut out some sections and shoot for 30-40 pages max (in the current format, it is 52 pages). Otherwise, I fear that the length will scare off readers and it will not have the desired impact. Some questions I would ask, for each section is, “is this section essential to the key goal of this paper?” “Would our key goal be harmed if it were cut?” Part of doing this may be scoping out the key goals a bit more by providing very clear questions this team wants to answer. Having a trusted colleague whose #1 goal would be to cut words/sections, and who is not part of the original authorship team, may also be helpful. Some suggestions as I was reading:

We thank the reviewer for sharing his/her perspective and the constructive propositions to improve clarity and focus of our paper. The key goal of this paper is to support researchers working with MMEs with a literature overview synthesizing existing practices. With this key goal in mind, we went through the entire manuscript again and shortened it substantially, by removing content or moving it to the appendix, e.g. in the Model Bias section and the Machine Learning section, see details below. This resulted in a substantially shortened and sharpened revised manuscript - reducing page number from 57 to 41.

— I’m a little biased (pun intended) but the bias section (page 12- 15, section 2.2) seems ripe for the chopping block is it is not immediately relevant to MME but models in general and is well covered by past literature. It comes off as a laundry list of random studies and is a bit tedious.

We moved this section to the appendix of the manuscript, which also helps contribute to a shortening of the paper. We believe that a summary of persistent model biases is still relevant to include in the paper, since it exposes limitations of ESMs, and thereby gives a direction for future ESM development. This should eventually improve climate MME.

— I also found pages 29-30 (section 3.3) to be mostly redundant with some of the messages already covered earlier in the paper (e.g. check that it performs well for a certain region/variable, MME allow for structural differences, uncertainty needs to be accounted for, etc). That these lessons also apply to extremes could be briefly mentioned in another section.

We thank the reviewer for this constructive feedback and have revised the section by reducing redundancy and emphasizing aspects that have to be kept in mind for extreme-event analyses (Extreme Value Theory, bias correction techniques for extremes, etc). We decided overall to retain the section, as we believe that a dedicated overview can be useful for authors working on extremes, allowing them to quickly access key considerations without having to search for related information scattered throughout the manuscript. We revised the introduction to section 3 to better clarify the goals of this section:

Building on the general workflow involved in MME studies in section 2, we draw on the experience within the Fresh Eyes community to identify common topics and challenges that arise in this context. All of these aspects are also relevant to the subsections in section 2; however, our aim here is to provide a dedicated overview of specific topics, allowing researchers to access the most relevant information in one place.

— There is a very long section on treatment of outliers (page 30-34) which I suspect could be much reduced.

We condensed this section substantially to almost half of its original length.

— I generally found the ML section to be too detailed and too lengthy. Pages 38-45 are dedicated to a deep dive on ML methods which seems to deviate from the overall emphasis of the paper to provide info on how to work with MMEs in CMIP. Truthfully, it feels like it belongs to a separate paper. I'm not proposing ML to be removed but perhaps a more top level overview of the pros and cons of ML to evaluate and work with MME.

Thanks for this feedback. We incorporate it by re-shaping this section substantially following the key goals we want to achieve in this paper. This resulted in a substantially shortened section.

(2) In addition to reducing the scope and length of this paper, I would also ask the authors to consider mentioning that there are other "climate MMEs" in the form of initialized climate models used for seasonal climate prediction. These are not the same as "initial condition ensembles" (ICE)— in prediction, there is more emphasis on assimilating the most accurate ocean/atmosphere/land state (this does not require computing an assimilation system from scratch, which is time/computing intensive. Some prediction models are therefore initialized from reanalysis that are separate from the prediction system).

Initialized climate models have the advantage of having more frequent predictions that can be more immediately verified and are still climate models in the sense that correctly implementing the right radiative/boundary forcings is essential, especially beyond the influence of synoptic weather/subseasonal variability like the MJO. While I realize the main goals of CMIP are distinct, I think trying to merge "MME" lessons from these two worlds would be valuable— mostly because some challenges are the same, such as how to optimally combine MME and validating against observations. I would recommend reading the references below and perhaps adding a few lessons from this "initialized climate prediction MME" community that may be shared with CMIP.

Kirtman, B. P., and coauthors, 2014: The North American Multimodel Ensemble: Phase-1 Seasonal-to-Interannual Prediction; Phase-2 toward Developing Intraseasonal Prediction. Bull. Amer. Meteor. Soc., 95, 585–601.

DelSole, T., J. Nattala, and M. K. Tippett (2014), Skill improvement from increased ensemble size and model diversity, Geophys. Res. Lett., 41, 7331–7342.

Becker, E., Kirtman, B. P., & Pegion, K. (2020). Evolution of the North American multi-model ensemble. Geophysical Research Letters, 47, e2020GL087408.

Buontempo, C., and coauthors, 2022: The Copernicus Climate Change Service: Climate Science in Action. Bull. Amer. Meteor. Soc., 103, E2669–E2687.

Becker, E. J., and co authors, 2022: A Decade of the North American Multimodel Ensemble (NMME): Research, Application, and Future Directions. Bull. Amer. Meteor. Soc., 103, E973–E995.

Min, Y.-M., Lim, C.-M., Yoo, J.-H., Kim, H.-J., Kryjov, V. N., Jeong, D., et al. (2025). A diachronic assessment of advances in seasonal forecasting: Evolution of the APCC multi-model ensemble prediction system over the last two decades. Geophysical Research Letters, 52, e2025GL116416.

There are indeed interesting complementary learnings within this community. We scanned the listed references and added the following points in the revised manuscript, along with the references:

- a brief introduction on initialized climate models in the Introduction: "While the focus of this paper is on the challenges associated with combining various ESMs within a MME, it should be pointed out there are other types of climate ensembles. Besides such uninitialized simulations, there are initialized climate model ensembles that are routinely used for seasonal prediction. These systems emphasize accurate initialization and thus have an emphasis on assimilation procedures to capture the atmosphere, ocean and land conditions accurately. While their goals differ from those of CMIP, initialized prediction ensembles face similar challenges related to ensemble design, model weighting, and evaluation against observations. "

- that also for initialized models, the ensemble mean outperforms predictions from individual models (Introduction)

- some discussion on the advantages of initialized models via the possibility of direct verification (see also answer to comment 3)

- we learnt from this listed literature that the value of MME stems from the enhancement of signal, cancellation of errors, and the improved ability to characterize the uncertainty of model forecasts, and added this to the explanation why MMEs outperform individual models (Introduction)

- we added about advantages of MME, that a MME can compensate for individual models not being available e.g. due to technical or political issues and that they can help in identifying errors e.g. in data sets used for validation (Introduction)

Given the already tight length constraints of this manuscript, we believe that a more in-depth discussion of this topic would detract from the overall goal of the paper, which is to remain focused on the key challenges while keeping the manuscript concise.

(3) Somewhat related to #2, I think the CMIP community needs to think more about how to link climate scenarios with actionable climate services, where societal and financial decisions are being made based on their trust (or lack thereof) in initialized climate models. I would like this smart group of authors to ponder the idea that one of the reasons that IPCC and others may have lost some clout in recent years is that not enough people see this activity as immediately relevant to what they experience. For a subset of people, CMIP feels too far off to be relevant and too uncertain to make decisions. Therefore, I see opportunity to build additional credibility/trust by linking initialized MME world with CMIP/LE MME world. A helpful example is GFDL SPEAR which provides seasonal forecasts once a month and also provides a large ensemble (LE) as well.

https://www.gfdl.noaa.gov/spear/

https://www.gfdl.noaa.gov/spear_large_ensembles/

A user can conceivably then, make linkages (and build trust) based on the performance of this model that they use for practical decisions with the more far off scenarios that this model produces. My recommendation of this paper is not conditional on implementing this suggestion— if this group isn't comfortable addressing this issue, then it is fine to ignore.

Thank you for sharing these reflections on the trust beyond the scientific context. We appreciate the thought and incorporate this in the context of discussing model evaluation based on performance in section 2.1, distinguishing between uninitialized and initialized climate model simulations in the revised manuscript:

"For shorter timescale forecasts, predictions can be verified within days as observations become available. This is typical of weather forecasting and initialized climate model simulations, in which models are started from observation-constrained initial conditions. Such near-term verifiability offers an opportunity to build confidence in models, particularly for climate services and decision-relevant applications. Although initialized and uninitialized climate projections address different time horizons, linking insights from both may help contextualize uncertainties and enhance trust in long-term projections. Climate projections addressing longer time scales cannot be directly verified in real time, as the relevant time scales (decades to centuries) preclude immediate verification. This is the case for uninitialized climate model simulations, which represent the standard approach for long-term climate projections and are the focus in this review. Accordingly, climate model performances are evaluated with reference to past and present-day climatology (Knutti, 2010)."

Minor notes:

Overarching comment: I'm not sure if this is a journal requirement, but it is fairly challenging dealing with line numbers that only go from 0 to 100 and repeat. I would follow standard practice and only have one set of line numbers that do not repeat.

We absolutely agree. Unfortunately, the first digits of the line numbering in our original manuscript were cut off in the process of creating a PDF for journal submission. We apologize for the inconvenience this causes in referencing and make sure to have appropriate line numbering in the revised manuscript.

Page 3, Line 74-75: "Climate model evaluation" should distinguish between projections and predictions.

We have revised the paragraph to clearly distinguish between “predictions” and “projections” in the context of climate model evaluation.

Page 5, lines 29: “climate MME” isn’t sufficiently clear (see #2 above).

Clarified, see answer to #2 above.

Page 5-6: The subsection breakdown seems to be listed out twice (40-42 and 55-58) which is a bit confusing. I would cover the outline in one place.

Thanks for pointing this out. We resolved this in the revised manuscript.

Page 7, line 81: I would argue that NCEP/NCAR should not be used b/c it is very dated and old. In the stratospheric community, in particular, the use of NCEP/NCAR has been discouraged.

Thanks for raising this point, we revised accordingly.

Page 7, line 98-00: See point #2 above. We can verify climate models with some frequency.

We revised the paragraph to distinguish between the role of initialized (short time scales, can be verified) and uninitialized climate model projections (longer time scales, cannot be verified in real time).

“For shorter timescale forecasts, predictions can be verified within days as observations become available. This is typical of weather forecasting and initialized climate model simulations, in which models are started from observation-constrained initial conditions. In contrast, climate projections addressing longer time scales cannot be directly verified in real time, as the relevant time scales (decades to centuries) preclude immediate verification. This is the case for uninitialized climate model simulations, which represent the standard approach for long-term climate projections. Therefore, climate model performances are evaluated with reference to past and present-day climatology [...]”

Page 7, lines 98-Page 8, 44: “Performance oriented evaluation” Needs to be some commentary on the type of things that can be compared against observations. Since these are free running, there is no restriction that it be tethered to observations, thus limiting this sort of comparison. Can only make comparisons in general sense… e.g. what a pattern looks like, in the mean/climo, etc.

We adapted accordingly:

“Because uninitialized climate model simulations are free-running and not constrained by observations, performance-oriented evaluation cannot rely on a direct comparison of individual events or temporal trajectories. Instead, model evaluation is necessarily based on climatological characteristics, such as mean states or spatial patterns. Evaluating this climatological performance comes down to the choice of appropriate metrics. [...]”

Page 8, Line 20-30: what parameters are adjusted? Is this calibration the same as model tuning? Or something different. Would distinguish this if so.

Yes, calibration and tuning can be used here as synonyms. We clarified which parameters are typically adjusted: parameters “typically associated with unresolved processes such as clouds, convection, or boundary-layer dynamics”

Page 9, line 31: Need to remind the reader what “this assumption” is here. What assumption.

Thanks for pointing out this inconsistency. This issue was resolved through restructuring of the section.

Page 9, line 39: See #2-3 above. One alternative is to actually run these models in a way that can be validated against observations. Those that do well in this space may build trust for their use in projections.

We removed the respective sentence. We also hope that this possibility becomes more clear with our changes regarding your comment on page 7, line 98-00.

Page 10, line 77-79: Could imagine that more coordination on processes could result in larger group of scientists developing preferences for certain methods against others. This may also result in more model similarity. Could reduce diversity of models so how to protect against that? [future question?]

— Reading ahead to page 15 it seems that this problem may be observed, so it does suggest some thought should be given to how to protect against too much model convergence.

We agreed and have revised this paragraph to outline this risk accordingly:

"By incorporating process-oriented analysis into diagnostic packages (examples in section 2.5), evaluations become reproducible, accelerating model improvements and establishing benchmarks for progress. As with any standardization effort, however, such benchmarks must be applied with care, as they have the potential to promote model similarity. "

We also want to point out, that the original manuscript addressed this aspect in performance-oriented model evaluation (see below) and that we discuss model dependency in a dedicated section (section 2.2.):

"Given the diversity of possible research questions, there is no single or combined performance metric that can reliably identify the "best" model independent of the research question. While this may sound disappointing since it prevents the standardization of model evaluation, it also has the advantage of reducing the effect of model convergence due to tuning (Knutti 2010), which allows for a more reliable representation of future uncertainty and decreases the likelihood of making overconfident predictions."

Section 2.2. I know every study cannot be included but this one feels fairly critical to mention given the importance of ENSO. Planton et al. (2021) https://journals.ametsoc.org/view/journals/bams/102/2/BAMS-D-19-0337.1.xml

We added this reference in process-oriented model evaluation section, as the section 2.2. has been moved to the Appendix.

Section 2.2. I have to admit this section seems unfocused. I thought the goal was to explain how using multi model approaches can help diagnose and resolve climate biases. This comes off as a bit of laundry list of model biases, so I would suggest starting over and writing a more concise section with a few (maybe 2-3) concrete examples of how folks leveraged multiple models to diagnose these biases. I other words, how could they see and understand the bias better using multiple models vs. just a single model.

The goal of this section was to give some examples of persistent model biases in different components of ESMs (atmosphere, ocean, cryosphere, ...) to expose limitations of MMEs. This simultaneously also provides insight into which phenomena are generally poorly represented in a multitude of ESMs, which informs a broader community of climate model developers about future work endeavours to resolve climate biases. However, we understand the concern and we moved this section to the Appendix also to shorten the main part of the paper as suggested. We decided not to create an additional section on the topic as proposed, also as the overlap with the Model evaluation section would have been too substantial.

It feels like there is some repetition on page 16 (47-54) and page 18 (62-70). I also don't see the Figure 3 which is referenced (just Figure 2). I would clean this up a bit.

Figure 3 was incorrectly referenced, and should have been Figure 2. We corrected this in the revised manuscript.

We thank the reviewer for pointing out the repetition on pages 16 and 18. These paragraphs have been combined into 1 paragraph, avoiding repetition and streamlining the ideas. The text now reads:

"The lack of a universally accepted and unambiguous definition of model independence complicates efforts to systematically account for model dependence in MME studies. Some definitions focus on the conceptual idea of whether or not a model adds novel additional information to the MME (Masson and Knutti, 2011). Others adopt a more analytical approach to understanding model dependence, offering examples for evaluating model dependence and using their framework (e.g., Annan and Hargreaves, 2017). Despite such advances, no broadly accepted solution has yet emerged. Further approaches, such as weighting schemes (Section 2.3) have been proposed, but these tend to be problem-specific and struggle to capture the full complexity of model dependencies. The metadata reporting requirements introduced in CMIP6 have made comprehensive assessments of model dependence possible, thereby representing a meaningful advance in transparency. As new model generations are developed and incorporated to

CMIP, continued efforts to quantify and correct for model dependence will be essential to ensure robust ensemble projections that reflect true uncertainty.”

Page 20 (lines 21-30) Feels like there should be a clear statement in this paper about how retrospectively picking winners is tricky unless proper cross validation procedures are applied. I’m also curious how much more accurate the predictions are, so can something be cited here that quantifies this a bit more (line 23).

We have included further quantification of the potential to improve accuracy as requested:

“Another approach to account for model performance is the selection of a subset of models. This can also be considered as a weighting method, which uses the weight 1 for included models, and the weight 0 for excluded models. MMEs with optimized sub-selection can reduce the computational load and have been shown to decrease the ensemble-mean RMSE, e.g. by roughly 10–20% for air temperature and approximately 12% for precipitation relative to the full multi-model mean (Hamed et al., 2021; Herger et al., 2018; Snyder et al., 2024). “

Further quantification can be found in the references, e.g. Herger et al. 2018.

The “retrospective” process is discussed already in our manuscript, in a slightly different context, however the central problem of using the same observational data set is the same. We also discuss possible approaches to address this issue in the subsection of Model Evaluation on observational datasets.

“When evaluating models, it is important to bear in mind that model parameters are often adjusted to match the same observational datasets that are subsequently used for model evaluation.”

Page 20 line 29: typo— capitalized O.

Corrected.

Page 21: I feel like a sentence is needed to explain *why* SMILEs improves uncertainty in climate projections. Otherwise it’s unclear how it’s an improvement over polynomial fit.

We added an explanation how SMILEs can improve uncertainty partitioning by using model uncertainty as example:

“More recently, exploiting the increasing computational capabilities, Lehner et al. (2020) overcame the assumption of the polynomial fit (i) from Hawkins and Sutton (2009), which produced significant regional biases, by using several SMILEs. Instead of calculating the variance of the polynomials as in Hawkins and Sutton (2009), in this approach, the model uncertainty is calculated as the variance across ensemble means from the available SMILEs. This reduces methodological assumptions and thereby improves the results, making SMILEs currently a broadly used tool to partition uncertainty in climate projections. It is important to note that the lack of independence between models (section 2.2), and the methods to account for it (section 2.3) must also be considered in this context.”

Page 22 (lines 99-05): I think we need a clearer definition of what is meant by an “emergent constraint.” As written, it appears to just be plotting up x and y in models. What’s missing is that the relationship would need to be strong to be an effective constraint (could imagine a situation with low correlations) and that the current day observations should be plotted in the figure to help us understand how the real world is actually operating in the context of the various models (or model ensemble).

We rephrased accordingly:

“An evaluation and uncertainty reduction technique that avoids this bias is the development of emergent constraints (Hall et al., 2019). An emergent constraint refers to a statistically robust relationship across a MME between an observable present-day quantity (x) and a projected future change in a quantity (y), typically approximated as linear (Simpson et al., 2021). When this relationship is robust, observations of x can be used to constrain the plausible range of y, thereby reducing uncertainty. This is commonly achieved by analyzing the probability distribution function of y conditioned on the observed value of x.“

Page 27 (11-13): I’m not clear why it needs to be less than half the initial sample. Is this test among multiple model ensembles? So you don’t overly favor one model? If so, this needs to be stated up front.

We added further background on this:

“This requirement is introduced to avoid resampling bias, as random subsets close to the full ensemble share many members, are no longer independent, and therefore tend to reproduce the full-ensemble signal by construction rather than providing an unbiased estimate of the required ensemble size.“

Page 27 (line 19) Alternative to “internal variability is low” is when “signal to noise ratio is high.”

Added. Please note that this paragraph has been moved to section 4.2 in the revised manuscript.

Page 28 (line 31-32): As written, this is a bit unclear. Are you saying average together each distinct model ensemble to create an an ensemble mean and THEN average across multiple models? Or just average together all available members across multiple models pooled together? There is a distinction.

Thanks for pointing out the ambiguity. We have rewritten the statement for clarification.

“When multiple realizations (or variants) for a given simulation are available for the same model, it is considered good practice to first average all members within each model to obtain a single ensemble model mean and incorporate such means into the equally weighted MME (Knutti et al., 2010b).”

Page 34 (lines 03-06): I’m not totally sure what this is saying and I think this needs to be rewritten. How does identifying values in the tails of the distribution help find models that represent more extreme events? All models with ensembles produce distributions that have tails. Are you talking about single member models?

Thanks for this comment. We revised the lines accordingly.

Page 38 (line 19-20): In the leading/Intro section to ML, I think it’s worth explicitly mentioning that the advances in the modeling space have really been for weather timescales, in part because, there are sufficiently large training datasets. Climate and especially climate change provide fewer samples overall and more “out of sample” situations, which may be limited if the training dataset does not contain the extremes. So, while, yes, ML has potential to improve these models this may be much harder in the climate change domain than it is when applied to weather.

This is now addressed at the beginning of Sect. 4.1. Thank you.

Page 42-43: Some font issues.

Corrected.

Page 46 (line 47-49). I might mention “Single forcing large ensembles” (SFLE) which partitions the forcings into different components (GHG, aerosols, biomass, etc). https://www.cesm.ucar.edu/working-groups/climate/simulations/cesm2-single-forcing-le . Otherwise it can be difficult to diagnose what aspect of the forcings are leading to a response.

We thank the reviewer for pointing this out and added this valuable point.

“In this context, single-forcing large ensembles (SFLEs), such as those derived from the CESM2 framework (Simpson et al. 2023), provide a complementary approach by partitioning the forced response into contributions from individual forcings (e.g. greenhouse gases, aerosols, biomass burning). These ensembles evolve only one forcing at a time while holding all others fixed, thereby enabling attribution of the drivers underlying responses identified in all-forcing SMILEs.”