Responses to referees' comments about the manuscript
"Distribution-based pooling for combination and multi-model bias correction of climate simulations"

by M. Vrac, D. Allard, G. Mariéthoz, S. Thao and L. Schmutz

First of all, we express our gratitude to the two referees for their thorough review and valuable feedback. We made efforts to incorporate all of their suggestions, and we believe that the manuscript has been enhanced as a result.

Below, we provide a detailed response to the reviewers' comments, with the comments presented in black and our responses in blue.

Anonymous Referee #1

Comment: I find that this work can be published after minor, if any, revisions. This paper proposes a novel tool (the authors have christened it: a-pooling) that solves the problem of weighing, or somehow combining, several climate models, and, at the same time, the problem of correcting the statistical distribution of model output (the focus here is not precipitation and temperature). I would have expected the authors to give a bit more relevance to this point since new bias correction methods, that always outperform past methods, are developed every year, while a method that bypasses the model weighing problem is truly unique. To be sure, the authors study their novel method in three settings. In the first, they test their method's performance using reanalysis. In the second, they test the method using models as reality (perfect model experiment). And, in the last, they study the sensitivity of the method to model choice. The result here is a particularly happy one because the method itself filters out redundant information making ulterior additions of model output unhelpful. This is an important paradigm shift as one is no longer worried about a "measure-of-goodness" or a "weight" for a given model which used to be a problem in ensemble averaging. I only have minor and minimal comments which the authors are free to ignore in their revisions. Because the revisions are so minor, I do not see the point in reviewing the paper again.

Response: We thank the reviewer for his/her kind remarks. We confirm that we study our novel method according to three different settings (reanalysis, perfect model experiment, sensitivity). We appreciate the suggestion to emphasize more the key-point of the method: the combination of the bias-correction question with the model weighting problem. This is now improved in the revised version.

My line-by-line minor comments:

Comment: Line 16- The authors state that the new method is "efficient" which is a very precise physical concept. How do define "efficiency" and show that a-pooling is efficient? I did not find the concept again in the manuscript.

Response: Indeed, we did not define it. We have modified the sentence to avoid mentioning a concept that we do not define: "*… indicating that alpha-pooling presents a potent way to combine GCMs' CDFs.*"

Comment: Line 16-18 "*The results of this study also show that the CDFs pooling strategy for "multi-model bias correction" is a credible alternative to usual GCM-by-GCM bias correction methods, by allowing to handle and consider several climate models at once.* " I would have given this concept more than one sentence but this is not very important…

Response: We only slightly emphasized more that this is a unique concept in the abstract. Indeed, the abstract is probably not the right place to add more text and descriptions. However, we strongly appreciate the enthusiasm of the reviewer!

Comment: Line 83-85 "*we bring together, in an original way, "bias correction" and "model combination", which are usually seen as different categories of methods, employed by separate scientific communities*." And again…

Response: This sentence is now followed by this additional text in the revised article: "*We stress that our proposed alpha-pooling method hinges on a unique concept that allows the simultaneous bias correction of multiple climate model simulations. This is accomplished through the innovative combination of model CDFs, which stands as an original concept in its own right.*"

Comment: Line 165-What are the problems associated with linear and log pooling? I think the authors mentioned the issues with log pooling but what is the issue with linear pooling? Perhaps I just missed the discussion.

Response: While there is not an inherent problem with linear pooling, like any linear approach, the method may lack flexibility and thus fail to capture the necessary non-linearity required to adjust to the data and their CDF. That is why non-linear (e.g., log-linear or alpha-pooling) methods have been / are developed. This has been added to sub-section 3.2 on "Linear pooling", as well as in the first sentence of sub-section 3.4 on "alpha-pooling".

Comment: Figure 1 What are the L2 (norm in CDF space?) Z -score. Is panel C truly necessary? I did not find it referenced in the text.

Response: Indeed, L2 is the L2 norm in the CDF space. This is now indicated in the caption of Figure 1. Regarding panel 1(c), it was briefly mentioned in line 216 (of the initial submission). Panel (c) is a zoom on panel (b) and, as such, it is true that it does not bring much additional information. Hence, panel (c) is removed for the revised article.

Comments:

- Line 297 "In this experiment, only 5 GCMs are used. This is partly constrained by the α-pooling method that can have stability issues to infer the parameters when combining a large number of models." Some explanation here might help. Stability issues may, or may not, be unimportant.

- Line 298 "Although it has been tested with more than 10 models, 5 GCMs appeared as a good compromise between a reasonable computation time," Again, some explanation here would not do any harm. What is a "good' compromise?

Responses:

Regarding "stability issues when combining a large number of models": This means that, when a relatively high number of models (i.e., CDFs) are combined, such as 10, depending on the initialization values of the parameters in the inference algorithm, the "optimal" final parameters may vary. In essence, the optimized parameters are unstable in such a case. This is because many local minima attain undistinguishable L2 distances. Indeed, while final parameters may differ between initializations, the minimized criterion values – specifically, the L2 distance in the CDF-space outlined in Eq. (10) – remain relatively consistent, often converging to similar or nearly identical values.

In our study, we conducted three experiments using five models. This configuration ensured not only stability in the quadratic criterion but also consistency in the final optimized parameters. In addition to a reasonable computation time (e.g., no more than a few minutes of computations for each location/variable), this represents our definition of a "good compromise".

This is now clarified in Section 4.1:

> "*When a relatively high number of models (i.e., CDFs) are combined, such as 10, depending on the initialization values of the parameters in the inference algorithm, the "optimal" final parameters may vary. In essence, the optimized parameters are unstable in such a case. This is because many local minima attain undistinguishable L2 distances. Indeed, while final parameters may differ between initializations, the minimized criterion values – specifically, the quadratic distance in the CDF-space outlined in Eq. (10) – remain relatively consistent, often converging to similar or nearly identical values. Although it has been tested with more than 10 models, the use of 5 GCMs appeared as a good compromise in the sense that (i) it ensured not only stability in the quadratic criterion but also consistency in the final optimized parameters, (ii) it allows a reasonable computation time (e.g., no more than a few minutes of computations for each location/variable), and (iii) it employs a sufficient number of simulations to get robust results.*"

Comment: Line 360 The explanation about pixelated nature of figure 2 was unclear to me. Why should we not be worried about the large changes between adjacent grid-points?

Response: As precipitation has a well-known spatial heterogeneity – both in occurrence and intensity – that is often difficult to represent by climate models, it is expected, as a result, to retrieve this spatial variability in the estimated alpha-pooling parameters. This is now clarified in the first paragraph of sub-section 5.1:

> "*This can be explained by the widely recognized spatial variability of precipitation, encompassing both occurrence and intensity, which is often challenging to accurately capture in climate models and thus reflected in the spatial diversity in the estimated alpha-pooling parameters.*"

Comment: Line 445 Sensitivity experiments results: If I understand correctly, this suggests that no justification is needed for the choice of models. I find that this is a paradigm-shifting result and I would give it more prominence.

Response: We do not think that "no justification is needed for the choice of the models". On the contrary, we do think that a careful choice of models is and will always be key to obtain relevant and robust projections, even with the use of the alpha-pooling approach.

Anonymous Referee #2

General comments:

Comment: This paper proposes a new method for combining climate model bias correction and model combination into a single methodology dubbed alpha-pooling. The method is shown to be robust to model selection sensitivity experiment framework, and somewhat superior to the other tested methods in a perfect model framework. However, at least for the region and variables investigated, when verified against ERA5 reanalysis, little added value is seen for this method when compared to other, simpler methods. This result is explained as being due to the proximity of the calibration and validation periods, and so I therefore recommend separating these periods in time to check this explanation. In practice, I fear this method may not be particularly useful if it fails to provide any added value in the one experiment that can be verified against observations. I am, however, intrigued by the methodology and look forward to seeing the authors' responses.

Response: We appreciate the remarks and questions asked by the anonymous reviewer. We answer below to every specific comment, including the general comment about the ERA5 experiment.

Specific comments:

Comment: Section 5.1: What does the similarity of performance between alpha pooling and the other methods tell us? You briefly mention an explanation for their similar performance: that the calibration and evaluation periods are adjacent in time, but this is the only experiment you show that in grounded by verification against observations, and it shows little difference compared to other methods. What would happen if you chose time periods further apart in time, such as 1951-1971 compared to 2001-2020? Do you think the results would be substantially different, and if so, would alpha pooling show the best performance? If not, then what is the advantage of alpha pooling? I appreciate that it is a less constrained method that allows more flexibility in the resulting GCM CDF pooling, but if, in practice, the result is indistinguishable from linear pooling, then what is the point?

Response: We believe that repeating the ERA5 experiment over a calibration period further apart in time does not yield additional information. Indeed, it is evident that the climate changes (in temperature and precipitation) from 1980 to 2100 in the SSP8.5 CMIP6 simulations are significantly more pronounced than from 1950 to 2000 in the ERA5 reanalysis data. Therefore, since our primary objective is to assess and compare our various pooling strategies in the context of a strong climate change (see beginning of section 4.2 in the initial submission), the "perfect model experiment" (PME) effectively and more clearly fulfills this purpose already.

From a conceptual perspective, we propose a new method that extends the linear approach to offer more flexibility through the possibility of non-linear combinations via the alpha parameter. In other words, our suggested alpha-pooling method "contains" the linear one as the special case alpha=1. Therefore, alpha-pooling can do at least as good as linear-pooling.

In practice, the ERA5 experiment indicates that linear pooling yields satisfactory results over consecutive calibration and projection periods. This suggests that the added flexibility of our method may not be required in a climate that is relatively stable or undergoing minimal change. However, findings from the PME experiment, particularly during the 2081-2100 period (and earlier periods depending on the criterion) where significant climate change is evident, demonstrate a clear distinction between the results obtained through alpha-pooling and linear pooling. The additional parameter introduced by alpha-pooling, which allows for non-linearity, positively influences the outcomes, resulting in superior results with our proposed approach.

Thus, considering this outcome and the likelihood that climate changes since 1950 are significantly less pronounced than those depicted in SSP8.5 CMIP6 simulations, we believe that extending the ERA5 experiment is unnecessary to substantiate our argument. Regardless of the results obtained from additional ERA5 experiments covering more distant time periods, we assert that evaluating the added value of alpha-pooling within a context of climate change is more effectively accomplished through our PME experiment.

All this has been clarified in the revised article at the end of section 5.1:

> "*In the ERA5 experiment, the results are relatively similar for the four methods. This indicates that the added flexibility provided by alpha-pooling may not be required over the 1981-2020 period of ERA5. This can nevertheless be different when considering other projection periods and reference datasets. Furthermore, the evaluation (2001-2020) and calibration (1981-2000) periods are quite close to each other, resulting in similar outcomes for both periods. These two results suggest that distinguishing between the different methods may be challenging in a climate that is relatively stable or undergoing minimal change. However, our primary objective is to assess and compare our various pooling strategies in the context of a significant climate change. Given that climate changes (in temperature and precipitation) from 1980 to 2100 in the SSP8.5 CMIP6 simulations are significantly more pronounced than what can be seen in the*

> *whole ERA5 reanalysis dataset over western Europe, the "perfect model experiment" (PME) will effectively and more clearly fulfill this purpose.***"**

Comment: L466-L469: It's true that due to the weighting can result in CDFs that are strongly affected by individual models, but isn't it also true that the weighting would decrease the influence of models with significantly different CDFs from the "truth" model?

Response: Yes, this is true over the calibration period. However, there is no guarantee that a model that is close to reference over calibration will still be close over the projection, or the other way around, that a model significantly different from the reference (i.e., with a small weight) over calibration will still be different over the projection (hence the perfect model experiment to test the robustness of the projections). Put differently, assigning a high weight to a model will inevitably impact the projection, regardless of whether the model closely aligns with reality during the projection period. This has been better clarified towards the end of sub-section 5.3:

> "*It was somewhat expected that the linear- and alpha-pooling have a larger uncertainty than MMM. Indeed, the use of weights means that models with higher weights will have a stronger influence on the resulting CDFs and bias corrections. Thus, even if these models do not closely align with reality during the projection period, their influence can lead to combined projections that can significantly deviate from the simple average performed by MMM.*"

Comment: Section 6: Some discussion of how alpha-pooling can retain uncertainty for future climate scenarios is warranted. Because MMM pooling can potentially result in overconfident projections, your new pooling method may more realistically portray scenario uncertainty. Though there is also the possibility that your pooling method results in unrealistic scenario uncertainty. Either way, I'd be interested to hear your thoughts on the matter and see some discussion of this in the manuscript.

Response: We fully agree that it is an interesting question, albeit beyond the scope of the current article. We thus mentioned this important point in section 6:

> "*While MMM-pooling has the potential to lead to overly confident projections, our novel pooling method may offer a more realistic representation of scenario uncertainty. Nevertheless, it is crucial to acknowledge the potential for our α-pooling method to introduce unrealistic scenario uncertainty. This aspect warrants further investigation in future studies, especially for practical applications.*"

Comment: Overall: Your method is certainly intriguing, and I think the PME experiment demonstrates its potential, but in practice, I wonder how useful this method really is. Assuming you would calibrate this method against a reference CDF from an observational dataset in order to attain a projection envelope for the future century, doesn't your first experiment demonstrate that in practice, at least for these (important) variables and region, your method isn't any different from MMM or linear pooling? You also state several times that this is a computationally costly method, doesn't that compromise its applicability if MMM/linear pooling can achieve such similar results at lower cost?

Response: Please, note that all outcomes from the sensitivity analysis section are obtained after calibrating the different pooling methods (+ CDF-t) using ERA5 data as reference, i.e., mirroring a practical application. It means that, while the results exhibit some degree of similarity among the three pooling methods over the first 20-year time period (2001-2020), significant discrepancies may emerge during the 2081-2100 timeframe, since the uncertainty ranges are quite different. Therefore, in practice, it becomes evident that alpha-pooling significantly differs from MMM or linear-pooling.

In terms of computation time, it is obvious that alpha-pooling is more computationally demanding than linear- or MMM-pooling. This is in part due to the additional parameter α, but mostly to the nonlinearity induced by α-pooling. However, for the combination of up to 10 climate models (i.e., CDFs), the computational time for each location and variable time series typically does not exceed a few minutes. Given the substantial computational demands associated with running individual climate models, the computational aspect of combining them is trivial by comparison. Moreover, considering that this post-processing of climate simulations does not need to be performed on a daily basis but rather once for all, we believe that this represents a reasonable computational cost, ensuring the method's practical applicability without compromise.

This last paragraph has been added to the conclusion section 6 for clarification.

Technical corrections:

L351: Space before period

L370: Noth-East -> northeast

L381-383: no need to capitalize directional terms in my opinion

L405, L444, and L466: somehow -> somewhat

L455: smaller -> lower

Section 5.3: envelop/envelops -> envelope/envelopes

L463: “an important information is their size” -> “their size is also important”

L463: I’d probably call it envelope width or breadth rather than length

Response: All technical corrections have been done.