Response to Reviewer Comments

I acknowledge the efforts the authors seem to have put into revising their paper. However, I also have to stress that they seem to have ignored large parts of my first major comment from the last round of revisions (which in turn was already a follow-up from the first round).

We apologize that the reviewer believes that we ignored portions of their comments from the revisions. We would like to offer clarifications to address these concerns and suggestions for additions to further address the reviewer's remarks. Our comments in response to the reviewer are in blue and follow each section of reviewer comments.

It reads: "In answer to my comment on a missing skill analysis the authors write in their answer 'The added value is shown in the reduction in bias in the historical period (e.g., Figure S3), and in the quantification and ultimately the reduction of uncertainty in the estimated climate change signal (e.g., Figures 5 and 6).'

I am not convinced by either of these arguments:
- 'reduction in bias in the historical period': ultimately (I assume) the weighting is mainly intended to improve assessments of the future and a reduced bias in the historical period (in sample) does not necessarily mean better model performance in the future (e.g. Sanderson et al. 2017). This is why I brought up the perfect model test.
- 'reduction of uncertainty in the estimated climate change signal': A reduction of model uncertainty does not necessarily mean a better representation of future climate. Indeed, the opposite could be the case if the raw model distribution was already overconfident (in this case an increase in uncertainty would be beneficial at least from a reliability perspective).

In figures 3 and 4 we see that the weighting strategies effectively reduce the model ensemble to only 3-5 models with non-zero weights. I wonder if such a reduction does not indeed rather lead to worse, i.e. overconfident/too narrow future projections?"

Again, we apologize that the reviewer believes that we ignored these remarks. This was not our intention. We very much agree with the reviewer that the reduced bias in the historical period does not mean better model performance in the future period. We also agree that if the raw model distribution is itself overconfident that increasing the model uncertainty would be beneficial.

On the first point, our intention with this manuscript was to assess the sensitivity of future projections to the various weighting strategies. We do not think that our historical weights justify using any or all the weighting strategies for the future projections. To the best of our knowledge, we have not said or implied this in our manuscript. However, we understand that our manuscript may not have gone far enough in discussing this point. We have included the following specific statements to this point at the end of Section 3 on Line 362 of the revised manuscript.

"It is important to note that this analysis of RMSE and bias is for the historical period only. Prior studies have noted that reducing historical biases does not mean better performance during the future period (Dixon et al. 2016; Sanderson et al. 2017). Therefore, historical skill alone does not justify the use of any weighting strategy. In what follows, we do not recommend using any specific weighting strategy based on the historical skill. Rather, we focus on the sensitivity of the projected changes to the various weighting strategies."

On the second point, we again agree with the reviewer that (theoretically) increasing the uncertainty would be appropriate if the raw ensemble is already overconfident. However, this is not the case in our analysis. The right-hand panels of Figure 5 in the manuscript show the spread of projected changes around the mean from the raw ensemble for both variables, while Figure 6 show the reduced spread of projected changes from the application of each weighting strategy, particularly from the BMA results. We recognize that we did not make this statement explicitly in previous iterations, but the reason for including these figures and the associated text was precisely to address this concern. The following is included at the end of Section 3.3 on Line 291:

"In addition, the right-hand panels of Figure 5 show that the projected changes around the mean from the raw ensemble are significantly larger than the reduced spread in Figure 6 (particularly from the BMA results) in the

weighted ensembles. This suggests that the raw ensemble has less confidence for both variables, both ensembles, and all three regions compared to the weighted ensembles."

On the final point above, the reviewer is correct that the several models regularly have large weights, but no models are given zero weight regardless of the weighting strategy used. We have provided the weights themselves in Table S2 in the supplemental material to clarify this. In addition, we have included the following text in the manuscript at the end of Section 3.3 on Line 296 to direct readers to the model weights.

"The weights for each model from each multi-model weighting strategy are given in Tables S2-S7."

We agree with the reviewer that having a few models with non-zero weights could make future projections overconfident, but this is unlikely in this case given that the spread of the raw ensemble for both variables was quite wide, even in the case of projected changes from the downscaled ensemble, as we have discussed above.

So, in short, my criticism was that the authors use the historical (in sample) RMSE improvement to justify weighting future changes by putting most of the weight on only a few models. I also pointed out that this might lead to overconfident results for projections of future climate.

Obviously it is, generally speaking, up to the authors to implement reviewer comments or argue against them. I think there might be arguments why implementing my comment is not relevant or feasible for this case but what the authors seem to do in their answer is simply reiterating their results.

Basically, they refer to additional tables and figures showing improvements in historical RMSE, if I am not mistaken. I am sorry, but this answer just does not address my comment. It is not surprising that historical bias is reduced when weighting based on historical bias but this does not justify (without further analysis and discussion) to apply the same weights to projections of future change.

Again, it was not our intention to ignore the reviewer's remarks and we apologize if we have inadvertently done so. To reiterate, the purpose of our manuscript is to examine the sensitivities of projected changes associated with weighting strategies. In a previous response to the reviewers (dated August 2022), we believed we had addressed the comment of the reviewer related to the aforementioned perfect model test and provided justification as to why we elected not to perform a perfect model test for this manuscript. The following is from the response to reviewers submitted in August 2022:

"While we agree that with the reviewer that this is a useful exercise, the question of the stationarity of weighting schemes is beyond the scope of this analysis and worthy of a manuscript in and of itself. In addition, the perfect model method assumes that the model that is chosen is a good approximation to the truth and that the future climate simulated in this model, along with the change in climate that occurs within its simulations, is representative of actual climate change and that the other models in the ensemble should have the same change signal. For example, in Brunner et al., (2019, ERL) it mentions 'For all regions there is also a chance that the skill decreases due to the weighting. This can happen if the perfect model has a very different response to future forcing compared to the other models, leading to the weighted multimodel ensemble moving further away from the 'truth'.' Furthermore, when applying weighting on the LOCA downscaled data, this perfect model test becomes irrelevant since all the models are bias-corrected to apply the downscaling. The author team is considering examining the question in a future study where we would also vary the model used as the absolute truth to examine some of the assumptions associated with the perfect model approach."

We recognize that the above may not have sufficiently answered the reviewers' concerns and suggestions and we have included the following language at the end of Section 3 on Line 362 to emphasize that we do not view the historical weights as justification for their future use.

"It is important to note that this analysis of RMSE and bias is for the historical period only. Prior studies have noted that reducing historical biases does not mean better performance during the future period (Dixon et al. 2016; Sanderson et al. 2017). Therefore, historical skill alone does not justify the use of any weighting strategy. In what

follows, we do not recommend using any specific weighting strategy based the historical skill. Rather, we focus on the sensitivity of the projected changes to the various weighting strategies."

In addition, the following language is included in the conclusions on Line 546 to note that further work should be done specifically on the question of the bias of future projections associated with multi-model weighting strategies.

"Future research will examine the accuracy and sensitivity using a perfect model exercise (such as what is described by Dixon et al. 2016 and Sanderson et al. 2017) to test the stationarity assumption associated with ensemble weighting. This is important since studies like Sanderson et al. (2017) show that a more skillful representation of the present-day state does not necessarily translate to a more skillful projection in the future. Our study does not consider the skill of the multi-model weighting strategies in the future projections, but rather it assesses the sensitivity of future projections to the various multi-model weighting strategies."

I find the lack of response to this point particularly striking because there is (among other work) a paper from Sanderson et al. (10.5194/gmd-10-2379-2017) which was done in the frame of the fourth National Climate Assessment for the United States (NCA4) which seems to be the same framework the authors work in. In the work from Sanderson et al. one can find statements like: "A more skillful representation of the present-day state does not necessarily translate to a more skillful projection in the future. In order to assess whether our metrics improve the skill of future projections at all, we consider a perfect model test where a single model is withheld from the ensemble and then treated as truth." So my criticism on the authors approach is not a personal opinion but reflected in the published literature and, in fact, in literature from the same climate assessment the authors work on.

Again, we apologize if the reviewer believes that we intended to ignore their remarks in previous reviews. We want to clarify that we don't think our historical weights are justified for the future. Instead, our paper shows the extensive matrix of results without choosing or recommending any specific weights. We understand that weights obtained using a historical bias do not necessarily mean a lower bias in the future. We are in complete agreement with the reviewer on this point. In our original edits, we tried to specifically address this comment with the following statement in the conclusions:

"Third, this study does assume stationarity in the multi-model ensemble weights and resulting weighted means. Future research will examine the accuracy and sensitivity using a perfect model exercise (such as what is described by Dixon et al. 2016) to test the stationarity assumption associated with ensemble weighting."

We acknowledge this might not be enough to address this concern, therefore we have extended the statement above to include the following additional statements inspired by the comments made by the reviewer:

"Third, this study does assume stationarity in the multi-model ensemble weights and resulting weighted means. Future research will examine the accuracy and sensitivity using a perfect model exercise (such as what is described by Dixon et al. 2016 and Sanderson et al., 2017) to test the stationarity assumption associated with ensemble weighting. This is important since studies like Sanderson et al. (2017) show that a more skillful representation of the present-day state does not necessarily translate to a more skillful projection in the future. Our study does not consider the skill of the multi-model weighting strategies in the future projections, but rather it assesses the sensitivity of future projections to the various multi-model weighting strategies."

We hope that including the above text will address the reviewer's concerns and show that we agree with the reviewer on this topic, and make it more clear to the reader what the actual aims of this study are.

So, given that I have now basically made the same comment twice and it has remained unanswered, I am sorry to say that my recommendation is now to reject this manuscript.

We apologize if we have ignored the reviewers' comments in our responses. It was not our intention to do so. To reiterate, the purpose of our manuscript is to assess the sensitivities of future projections to the numerous available options for weighting strategies. As our title states, we are "Assessing sensitivities of climate model weighting to

multiple methods, variables, and domains in the south-central United States". We completely agree that weights obtained using historical bias do not necessarily translate into a lower bias in future projections. We hope that the above comments address the reviewer's concerns.