Responses to Reviewer 1

Author Comments: We thank the Reviewer for taking the time to review our manuscript a second time. Our responses to each comment follow the written text of the reviewer.

Review of "To weight or not to weight: assessing sensitivities of climate model weighting to multiple methods, variables, and domains in the south-central United States" by Wootten et al.

This is my second review of this manuscript. First of all I would like to thank the authors for their extensive answers to my comments and their work on the manuscript. I think I now understand better what they want to achieve with this study. However, I still see a mismatch between the answers the authors give to their research questions and the actual analysis done. In the first round of revisions I commented that the author's analysis is not sufficient to properly answer the posed research questions and this criticism still holds.

As I see it the manuscript has two parts:
1. an (extensive) analysis of weighting scheme impacts presented in section 3. and
2. recommendations in section 4 .

Both parts individually are mostly fine (for example I generally agree with most of the recommendations given – just not with the way they are intended to be based on the analysis). I don't really know how to resolve this larger issue I see with the manuscript other than focusing either only at 1. (and doing a purely physical analysis of weighting effects) OR only at 2. (writing a perspective style manuscript with general recommendations).

If the authors want to do both they need to cleanly connect both, clearly showing a chain of analysis-result-interpretation-recommendation. This is not done at the moment and therefore I can not recommend this manuscript for publication at this point.

Author Comments: We thank the reviewer for the comments and thorough review of our manuscript. In line with the comments above we have made significant changes to the manuscript. The new version of the paper now focuses on the extensive quantitative analysis of weighting strategies (which includes the weighting scheme and the choices surrounding variable[s] and domain of interest). The second portion of the manuscript, which used to include a qualitative perspective with general recommendations, is now removed from the paper. For this reason, we think that several of the comments below no longer apply with this revision.

Major comment on the analysis / Discussion from the last round of revisions

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?

The authors go on to write ‘In this study, we show that, yes, for different variables and domains that different weights need to be estimated.’

I am sorry but this is not shown. The authors merely show that the weights differ for different cases. There are some figures showing the change in RMSE in the historical period but they are obviously not the focus of the analysis as they are placed in the supplement and not really discussed (figures S3-S6). And looking at them I can find some examples where using information from the same region leads to better historical RMSE scores but there are also counter examples to the overall interpretation remains unclear.

Author Comments: In response to comments from Reviewer 2, Tables 2-4 are added to the text to incorporate a clear discussion of the improvements offered by weighting strategies that also clarifies that there are differences in the improvement offered dependent on the domain, ensemble, and variable used. The following is added prior to the Discussion on Lines 413-425: “Figures S3-S6 indicate that all the weighting strategies used in this study resulted in higher skill for both high temperature and precipitation in all three domains. To summarize the results for skill, the RMSE of each weighting strategy is shown for all three domains for precipitation and high temperature in Table 2 and Table 3 and the RMSE for the unweighted cases are in Table 4. Of the weighting strategies using the CMIP5 ensemble 92%, 92%, and 75% have lower RMSE for precipitation than their unweighted counterparts for the full, New Mexico, and Louisiana domains. Similarly for the high temperature, 96%, 100%, and 79% of weighting strategies have lower RMSE than their unweighted counterparts for the full, New Mexico, and Louisiana domains. Therefore, most weighting strategies have higher skill than the unweighted CMIP5 ensemble. However, there is a similar pattern for weighting strategies using the LOCA ensemble. For precipitation, 79%, 58%, and 67% of weighting strategies using the LOCA ensemble have a lower RMSE than their unweighted counterparts for the full, New Mexico, and Louisiana domains. Similarly for high temperature, 88%, 88%, and 83% of weighting strategies using the LOCA ensemble have a lower RMSE than their unweighted counterparts.”

Major comment on the recommendations

In the abstract the authors write ‘From the results of our analysis, we summarize our recommendations concerning multi-model ensemble weighting as follows: […]’ followed by a list of recommendations. This to me clearly indicates that the recommendations are based on the analysis and I (still) disagree with that to a large extend. A similar situation arises in section 4 but for brevity I will focus mainly on the abstract here to make my point.

- 'That model weighting, if used, be derived using both common (e.g., precipitation) and stakeholder-specific (e.g., streamflow) variables […]'
How can this statement follow from the analysis if the authors do not look at any stakeholder-specific variables in their analysis?

Author Comments: Thank you. This comment has been removed from the text.

- 'That weighting is derived for individual sub-regions in addition to what is derived for the continental United States […]'
I completely agree with this statement from an expert perspective. But again, how is this derived from the results presented in section 3?
I assume some of the results shown in the RMSE figures S3-S6 could be used to make this case but this is not done in the manuscript.

Author Comments: This comment has also been removed from the text.

- 'Multiple strategies for model weighting are employed when feasible, to assure that uncertainties from various sources (e.g., weighting strategy used, domain or variable of interest applied, etc.) are considered.'
I find it hard to support this statement so generally. My argument would rather be that each weighting strategy used (and I agree that using multiple can be advantageous) should be justified. Just using multiple strategies with weights based on various regions and variables does not lead to any obvious benefits I can see if not supplemented by a more in-depth analysis of method skill or system understanding.
To connect this to the analysis take for example figure 8: here weighted mean values from 24 different strategies are presented. But what is the added value of this alone? There are quite considerable differences in the weighted means depending on the strategy and they cover both, higher and lower values than in the unweighted case. This actually leaves me with the slightly uncomfortable impression that it might be better to just use the unweighted case.

Author Comments: As mentioned above, these recommendations were removed from the paper. The paper now focuses on purely a quantitative analysis of the extensive results. As such, we hope this removal of text satisfies the concerns mentioned above by the reviewer.

Minor comments

17: I am sorry but I still find the convention used here unclear. The authors write that '...model weights and the corresponding weighted model means are highly sensitive to the weighting scheme that is applied'. But then the examples given here do not address different weighting schemes but only differences in the weights when they are based on different regions.

Author Comments: Thank you for this comment. We corrected this sentence in our abstract to read 'weighting strategies' instead of 'weighting scheme', and we go on to explain on Lines 127-132:

“For reference for the reader, we define weighting schemes to refer to the numerical approach to weighting alone, such as Bayesian Model Averaging (BMA) or the approach defined by Sanderson et al. (2015, 2017). We define a weighting strategy as the weighting scheme and other choices made when using the weighting scheme to derive model weights. For example, a weighting strategy would be using the BMA weighting scheme to derive weights using the continental United States and daily high temperature alone and another weighting strategy would be using the BMA weighting scheme to derive weights using the Southern Great Plains of the United States and daily precipitation alone. Both such examples use the BMA weighting scheme, but with different choices made to derive weights, making the two examples different weighting strategies.”

- ‘when estimating model weights based on Louisiana precipitation, the weighted projections show a wetter and cooler south-central domain in the future compared to other weighting schemes” Do the authors mean ‘other regions’ instead of ‘ other weighting schemes’? Otherwise I do not understand the sentence.

Author Comments: In line with the previous comment, we corrected this sentence in our abstract to read ‘weighting strategies’.

- ‘Alternatively, for example, when estimating model weights based on New Mexico temperature, the weighted projections show a drier and warmer south-central domain in the future. However, when considering the entire south-central domain in estimating the model weights, the weighted
future projections show a compromise in the precipitation and temperature estimates.’

My viewpoint is that there are, one the one hand, different weighting schemes and, on the other hand, these weighting schemes can be used to calculate weights based on different variables and regions. But in the second case (at least for me) it is still the same weighting scheme. This is what I tried to convey with my comment on the same topic in the last round of revisions and in fact the authors themselves make the same differentiation later (line 89) but not here it seems.

Author Comments: Thank you again. The point the reviewer made is correct, in that the different weighting schemes can be applied to calculate weights based on different variables and regions, and this is what a weighting strategy is. We have added a statement to clarify this in the introduction on Lines 127-132 to make our definition clear for the reader.

327: ‘The internal variability is represented by the ensemble average of the standard deviation of each variable from each ensemble member (per Hawkins and Sutton, 2009; 2011)’
Just to make sure I understand correctly what was done here: the authors have removed the estimated forced response using a 4th order polynomial and then calculated the time standard deviation as an estimate of internal variability as in the Hawkins and Sutton studies?

Maybe the authors want to mention the caveats of this approach? Mainly that the polynomial fit is not a very good estimation of the actual forced response (in particular regionally) as shown, for example, in a recent related publication using large ensembles: 10.5194/esd-11-491-2020

Author Comments: In this case, the calculation is over time, but we did not remove the estimated forced response since we could not estimate the polynomial regression from the historical period (1981-2005) all the way through the future period (2070-2099), as the period of the data was not continuous to make a robust regression. The sentence in question is revised on Lines 382-385 to read the following with an additional reference in support:
"The internal variability of the historical and future period is represented by the ensemble average of the standard deviation of each variable from each ensemble member over time (per Hawkins and Sutton, 2009; 2011, Maher et al. 2020) for each of the three domains (full, Louisiana, and New Mexico). However, we note that the forcing response is not removed given the temporal period is not continuous which is a caveat for this analysis."

Figure 8-13: This is just a suggestion but would plotting the change to the unweighted case in the small panels not show the effect of the weighting strategies better?

Author Comments: Yes, doing so may show more of the differences between the unweighted case and the individual weighting strategies in those Figures. However, we think that these figures together with Figures 6 and 7 (which show each weighting strategy compared to the unweighted case in a different format) make these differences apparent and allow comparisons between weighting strategies. As such, no change is made to Figures 8-13.

Responses to Reviewer 2

Author Comments: We thank the Reviewer for taking the time to review our manuscript a second time. Our responses to each comment follow the written text of the reviewer. There's a note that must be made at this point. In response to Reviewer 1 the manuscript was changed significantly to focus on the analytical analysis rather than the original questions and recommendations. For this reason, many comments made with respect to previous editions of this manuscript may no longer apply. We have responded to comments that remain in the manuscript and noted those that no longer apply below.

I continue to remain very doubtful about the value of this study. The authors have chosen not to address any of my main suggestions concretely, and they continue to defend the normative quality of this study, with a list of specific questions that they set out to answer on the basis of their results. Except, nothing in their results supports any of their answers to the list of questions, I'm sorry to say.

Once again, their results only confirm something we have known for a while, that different weighting schemes deliver different outcomes. I'm going to argue that nothing about this – expected – outcome supports the authors' recommendations. Specifically (the authors' words in italics)

Author Comments: We thank the reviewer for the comments. The new version of the paper now focuses on the extensive quantitative analysis of weighting strategies (which includes the weighting scheme and the choices surrounding variable[s] and domain of interest). The second portion of the manuscript, which used to include a qualitative perspective with general recommendations, is now removed from the paper.

That model weighting, if used, be derived using both common (e.g., precipitation) and stakeholder-specific (e.g., streamflow) variables to produce relevant analysis for impact assessments or using multiple climate variables relevant for a national assessment region (Question 1).
The authors don't use stakeholder-specific variables, or multiple variables, so nothing from their study shows the added value of doing that and therefore supports this first recommendation.

Author Comments: Thank you. This comment has been removed from the text.

That weighting is derived for individual sub-regions in addition to what is derived for the continental United States or other nations and that weighting for impact assessment is also derived for a domain relevant to the impact assessment (Question 2).
The authors results show that the outcome will be different depending on what domain is used. What would the final outcome of doing this multiple types of weighting be (besides once again discovering that the results vary)? Which one is right and which one is wrong?

Author Comments: This comment has also been removed from the text.

Weighted ensemble means should be used not only for national and international assessments but also for regional impacts assessments and planning (Question 3).
What, from the authors' results, support this recommendation? What would impact assessment and planning do with the results of weighting, which either will be fraught by the knowledge that the specific result is very sensitive to the actual choices made for the weighting schemes, or will be delivering multiple results in front of which the planner will be left with having to make a decision? What is the added value of these weighted schemes if we have no idea of their skills?

Author Comments: This comment has also been removed from the text.

Multiple strategies for model weighting are employed when feasible, to assure that uncertainties from various sources (e.g., weighting strategy used, domain or variable of interest applied, etc.) are considered (Question 4).
What concretely would the user do with the results of multiple weighting schemes? Again, why do this if nothing has demonstrated the added value of using a weighting scheme? I could argue that throwing out models at random will produce different results from these weighting schemes. Should that be done just to demonstrate that if CMIP had included different models we would have gotten different results? I would argue that is as important a recognition as any. But nobody sets out to do that for the sake of it.

Author Comments: This comment has also been removed from the text.

Future efforts should examine the weighting of impacts model outputs from climate model inputs (Question 5).
We don't know what to do with climate model inputs weighting…how are we going to combine that with impacts model weighting? Wouldn't it be better to first figure out what skills model weighting has, if any at all? And again, and for the last time, where is this recommendation coming from, on the basis of the paper results?

Author Comments: This comment has also been removed from the text.

Latching on to this last comment, I would like to underline that these recommendations would make any sense only if the authors had shown any skill in the weighted projections compared to the unweighted, but none of that is shown in this paper. I would find the publication of these recommendations actually problematic and harmful, not being substantiated in any way.

Author Comments: Thank you. Given the response to Reviewer 1 and the resulting changes to the manuscript many of the above comments no longer apply with respect to the manuscript. However, the reviewers' final comment above does merit a response. The reviewer claims that we did not demonstrate the skill of the weighted projections. Figures S3-S6 show the RMSE of the weighted ensembles compared to the unweighted case for each of the weighting strategies considered in the study. These four figures clearly show that the weighted projections have greater skill than the unweighted case for both high temperature and precipitation in all three domains. In this revised manuscript, we have noted this in the main manuscript and refer the reader to the Supplemental Figures.

Furthermore, Tables 2-4 are added to the text to incorporate a clear discussion of the improvements offered by weighting strategies that also clarifies that there are differences in the improvement offered dependent on the domain, ensemble, and variable used. The following is added prior to the Discussion on Lines 413-425:

“Figures S3-S6 indicate that all the weighting strategies used in this study resulted in higher skill for both high temperature and precipitation in all three domains. To summarize the results for skill, the RMSE of each weighting strategy is shown for all three domains for precipitation and high temperature in Table 2 and Table 3 and the RMSE for the unweighted cases are in Table 4. Of the weighting strategies using the CMIP5 ensemble 92%, 92%, and 75% have lower RMSE for precipitation than their unweighted counterparts for the full, New Mexico, and Louisiana domains. Similarly for the high temperature, 96%, 100%, and 79% of weighting strategies have lower RMSE than their unweighted counterparts for the full, New Mexico, and Louisiana domains. Therefore, most weighting strategies have higher skill than the unweighted CMIP5 ensemble. However, there is a similar pattern for weighting strategies using the LOCA ensemble. For precipitation, 79%, 58%, and 67% of weighting strategies using the LOCA ensemble have a lower RMSE than their unweighted counterparts for the full, New Mexico, and Louisiana domains. Similarly for high temperature, 88%, 88%, and 83% of weighting strategies using the LOCA ensemble have a lower RMSE than their unweighted counterparts.”