Earth Syst. Dynam. Discuss., author comment AC1
https://doi.org/10.5194/esd-2022-15-AC1, 2022

Reply on RC1

Adrienne M. Wootten et al.

Author comment on "Assessing sensitivities of climate model weighting to multiple methods, variables, and domains in the south-central United States" by Adrienne M. Wootten et al., Earth Syst. Dynam. Discuss., https://doi.org/10.5194/esd-2022-15-AC1, 2022

We thank the reviewer for taking the time to thoroughly review our manuscript. The reviewer's comments are included below in quotes. Our responses to the comments provided are in italics at points in the reviewer's comments.

"The authors offer that the value of this work is in this systematic exploration of the effects of weighting, but I am sorry to say that, aside from some very nice and thoughtful discussion of general issues (which by the way have been treated in some depth by a guidance document for the IPCC AR5 report as early as 2010, available here https://www.wcrp-climate.org/wgcm/references/IPCC_EM_MME_GoodPracticeGuidancePaper.pdf, and more recently in a review paper by Abramowitz et al. (2019) https://doi.org/10.5194/esd-10-91-2019), and the appreciation of the large amount of work that the authors have undertaken, I come away from this study only reinforcing what we all already knew: that different weighting schemes produce different results and nobody knows how to interpret the real value of those differences and what to do about it."

We respectfully disagree with the assessment of the reviewer. The debate over climate model weighting is precisely why we chose to invest the time and energy into this extensive study. Nobody knows what to do about the differences in results between different methods and applications of climate model weighting. So, we underwent this extensive and comprehensive research matrix of results and answered some of these outstanding questions in the community. In addition, several authors are involved in producing the Fifth National Climate Assessment (NCA) in the United States. The NCA effort includes a series of group discussions on downscaling and model weighting, which are the same questions of interest to this study. No other study has comprehensively answered these many questions in one study, such as applying model weighting based on the climate variables of interest, the domain of interest, different model weighting strategies, or the dataset used (GCM or downscaled).

While the IPCC AR5 report from 2010 provides general guidance, it does not include analysis or investigation into the recommendations provided, whereas our study does. For example, in Section 3.5 of the IPCC AR5 report, the authors discuss recommendations for regional assessments, and conclude that "Particular climate projections should be *assessed against the broader context of multiple sources (e.g., regional climate models, statistical downscaling) of regional information on climate change (including multi-model global simulations), recognizing that real and apparent contradictions may exist between information sources which need physical understanding." This is precisely what our study aims to do, by utilizing data from both global models as well as their downscaled counterparts. This is just one example of how the IPCC AR5 report makes recommendations and does not perform any investigations, where our study does apply the research needed to address such recommendation. In addition, our study does so and goes on to make several recommendations on the appropriate use of multi-model ensemble weighting, which is summarized in the abstract and conclusions sections of our study.*

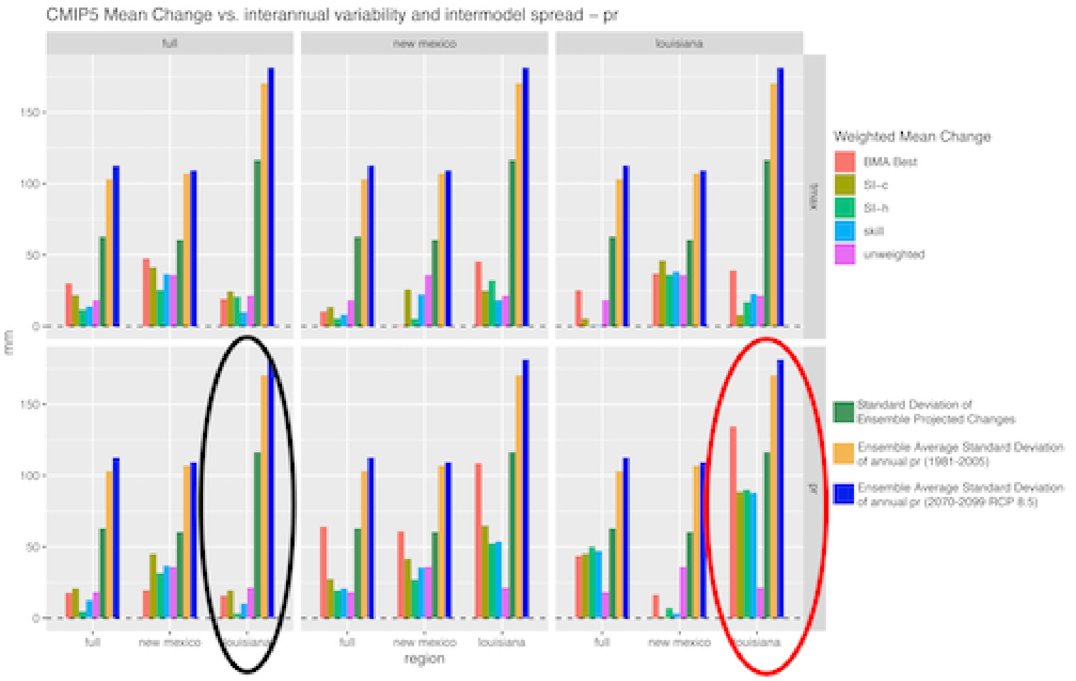
While Abramowitz et al. (2019) covers the concept of model dependence, our manuscript goes much further. Through the various weighting schemes, this manuscript covers different approaches to dealing with model independence. Two of the weighting schemes account for model dependence using the method created by Sanderson et al (2017) that accounts for model dependence in the historical simulation, and a variation of the Sanderson et al. (2017) method that accounts for model dependence in the future climate change signal. The former method has been used in previous studies, but the latter method is not a common approach to dealing with model independence and is not covered in the paper by Abramowitz et al. 2019. In addition, this study also includes a Bayesian Model Averaging (BMA) weighting scheme that approaches the model dependence problem over multiple moments of the distribution. Bayesian approaches are only mentioned in passing by Abramowitz et al. 2019.

Given the above, we believe that our study is timely given the debate in the community regarding the use of multi-model weighting.

"In my view, there would be two ways to make this exploration more useful. First, perform this exercise with a clear accounting of internal and inter-model variability. I don't know what to make of pictures that show me multimodel means and how they differ from one another. The question is, do they differ in a way that is significant, compared to internal variability? And do they differ in a way that is significant with respect to a measure of uncertainty around the multimodel mean, which could be taken (likely underestimating it and therefore possibly favoring the detection of significant differences, but that could be expressed as a caveat) as its standard deviation, computed by the inter-model standard deviation divided by the square root of the ensemble size (at each grid point)?"

We agree that this is a useful additional component and thank the reviewer for pointing out that we should address it, at least as a caveat, in our study. There are two points to be made on this issue raised by the reviewer. First, Figure 5 in our manuscript shows the inter-model variability for the projected changes of both variables for both ensembles and all domains and Figure 6 shows the resulting ensemble means from the various weighting schemes. These two figures clearly indicate that the differences between the means are not as large as the inter-model variability. Second, while these figures do offer some suggestion, we agree that this is not a robust and explicit treatment and does not include a discussion of internal variability. While a full treatment and discussion of the weighting compared to internal and inter-model variability is beyond the scope of this paper, we have produced an approximate analysis and an additional figure will be placed after Figures 5-6 with additional discussion.

The new Figure 7 includes the absolute value of the ensemble mean changes in comparison with the standard deviation of the projected changes of the ensemble *(representing the intermodel variability), the average of the standard deviation of the annual values of each ensemble member from the historical period, and the average of the standard deviation of the annual values of each ensemble member from the future period. These latter two items have been used as a rough approximation of the internal variability in multiple studies previously (e.g., Hawkins and Sutton, 2009; 2011). As such, the new Figure 7 is of a similar arrangement to Figure 6, but a component of the final figure is provided below (a high resolution copy is include as a pdf attachment to this comment). This component is for the precipitation projections from the CMIP5 ensemble, weighted based on tasmax in the top row and weighted based on precipitation in bottom row, and weighted based on the full (left), New Mexico (middle), and Louisiana (right) domains. Within a single plot of bars, the weighted means are for the weighting schemes applied to the full domain (left), New Mexico (middle), and Louisiana (right). For a single group of bars, the multi-model means from the various schemes are plotted directly against our proxies for the inter-model and internal variability. From this sample we can say that in most cases, the differences between ensemble means are not larger than the inter-model or internal variability. However, there are some cases where the differences between these means are comparable to or larger than inter-model variability or comparable to the internal variability. For example, for ensemble means weighted for Louisiana precipitation and applied to Louisiana precipitation (circled in red below for reference), the difference between the BMA ensemble mean and the unweighted mean is comparable to the inter-model variability. As a further example, the difference between BMA ensemble mean created based on Louisiana precipitation and all the weighted ensemble means created based on full domain precipitation (circled in black below for reference) is also comparable to inter-model variability and internal variability. The analysis discussed here will be included in the revised version of the manuscript to address the comments of the reviewer.*



Component of New Figure 7. *Absolute value of mean projected changes in precipitation from the CMIP5 ensemble using multi-model weights produced with all four weighting schemes applied to all three domains and both variables (tmax and pr) plotted alongside the standard deviation of the CMIP5 ensemble and the ensemble average standard deviation of annual precipitation for both the historical and future periods (no weighting is used to calculate any standard deviations). In the new Figure 7, which will take a similar form to Figure 6, this would be the top left group of plots. The top row is the results from weighting schemes derived with tmax, and the bottom row is the results from weighting* *schemes derived with pr. In addition, within an individual group, the left column is the results from weighting derived using the full domain, the middle column is the results for weighting derived using the New Mexico domain, the right column is the results for weighting derived using the Louisiana domain. Within a given domain and variable, the results are shown from left to right for the domain the weights are applied to. For simplicity, the BMA best weights are used (and the boxplots from Figure 6 are omitted). The standard deviations are in all cases from the unweighted ensemble.*

"Second, perform a perfect model exercise where one model furnishes the truth, current and future, and the rest of the models undergo this exercise in variation of weights (derived using the left-out model historical portion as observations), so that besides ascertaining that the weights have diverse effects, we can start seeing something about the value of applying them: Do they produce anything more accurate than the unweighted projection? Which of the choices does that better, if any?

The need to take into account internal variability requires the "true model" to be one that has produced initial condition ensembles, but there are plenty of CMIP5-era large ensembles now available through US CLIVAR SMILEs (https://www.cesm.ucar.edu/projects/community-projects/MMLEA/), and the authors could easily choose one which has also participated in CMIP5 (e.g., CESM1, CanESM, MPI)."

While we agree 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 multi-model 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 will include a discussion of this topic in the conclusions and as a topic of future study.

"A study that can tell me something more than "things look different" and can distinguish differences that are simply noise from differences in the signal estimated by these various weighting schemes, then proceed to tell me which one of these weighting schemes, if any, produces projections closest to the "truth" would be really valuable and a real step forward in this old and somewhat frustrating debate.

And I realize that using the perfect model set-up pre-empties the idea of using LOCA, but I would argue that the loss would be more than balanced by the gain in interpretability of the results. Plus, the bias correction of LOCA makes the value of using performance-based weights rather debatable, and my guess is that the differences that surface in that part of the exercise would turn out to be drowned by internal variability if that was accurately accounted for (given that observations used to bias-correct are also just one realization, heavily affected by internal variability at these grid-point scales)."

The authors recognize that the use of statistical downscaling methods may make it debatable to use ensemble weighting. It was in recognition of this debate that we chose to include the ensemble of LOCA downscaled climate models in our experimental matrix, to try and provide answers to this debate.

"I also would like to raise a point about impact modelling. The authors discuss more than once the relevance of the weighting choice for impact modelers, but I would like to be better convinced of that. My experience of impact model(er)s is that they need climate information that looks like reality (one realization of it, or multiple realization of it) not like a big smooth mean. So I agree that the multimodel mean (weighted or unweighted) might be relevant as a synthetic "bird-eye view" of how climate impact-drivers look in the future, and can inform discussions and produce useful catalogs of maps in documents like IPCC or NCA assessments. However, when it comes to impact modeling, my expectation is that feeding multimodel means to a process or empirical model would be nonphysical. Even a large, global scale impact modeling exercise like ISIMIP (https://www.isimip.org/) has provided individual realizations of multiple models for use in its "children" exercises. I would think that using temperature, precipitation and whatever else is needed that behave like reality as input to the impact model, and only after having produced the impact response worrying about averaging, is even more necessary for regional impact assessments like the ones that the authors are mostly concerned about. If I'm wrong, I will happily stand corrected, but in that case I would like to see citations of current impact modeling studies that use multimodel ensemble means."

What the reviewer describes is precisely why Bayesian Model Averaging (BMA) is utilized. BMA has been proven to be a useful model weighting tool because BMA does not simply provide a single smooth multi-model mean, but instead provides samples from a posterior of model weights in which each sample produces a realistic model average that is not necessarily smoothed out but keeps the internal variability of the system intact. We reported on the mean from the BMA distribution, but in fact the samples can be investigated independently (as in Figure 6 of this study), for both looking at future climate change signals but also for driving impact models. We refer the reviewer to Massoud et al, (2019, 2020a), where BMA was extensively reported on and explained in detail. For example, Massoud et al., 2020a show how climate variability, such as annual and interannual variability of precipitation, is better explained with a BMA model average compared to most individual models, and especially compared to an unweighted model mean, which in comparison washes out any variability in the climate system it represents. In fact, the NCA's 5th assessment report will be utilizing BMA for these reasons, among others, as their chosen tool for applying model averaging. In addition, the study does point to early signs that multi-model ensemble means created with weighting schemes are being used for impact assessments (cited in the study - Skahill et al. 2021 - ***https://doi.org/10.3390/cli9090140****). The text will be modified to make the above points more clear.*

"In conclusion, my assessment of this work is that it represent a very diligent and substantial exercise, informed by thoughtful considerations, but does not help to advance the field until it takes up a better treatment of internal and model variability that could help to determine the significance of the differences resulting from the various weighting schemes, and until it can say something about the usefulness of weighting at all. I tried to suggest ways to do just that. I would be very excited to see the new results, which I hope would not be too difficult to produce, given the efficient machinery that the authors have obviously already in place."

We thank the reviewer for their comments and critiques. As mentioned above, we will provide an analysis of the ensemble means compared to the internal and inter-model variability (e.g., Figure 7 component shown above in this response letter). We will also adjust the text to provide some discussion and caveats as mentioned above.

Please also note the supplement to this comment:
https://esd.copernicus.org/preprints/esd-2022-15/esd-2022-15-AC1-supplement.pdf