Reviewer 1 comments

In this study, the authors set up a systematic exploration of several combinations of choices in multimodel ensemble weighting schemes, and describe the resulting projections when weighting CMIP5 models and their downscaled and bias-corrected LOCA versions.

We thank the reviewer for taking the time to thoroughly review our manuscript. 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 authors of this study are all directly and indirectly involved with the NCA discussions, led by the United States Global Change Research Program (USGCRP), surrounding downscaling and multi-model ensemble weighting for the Fifth NCA. There are 10-15 people representing multiple agencies of the United States government discussing issues associated with downscaling and model weighting during bi-weekly meetings over the past two years. The effort in this study, and the questions of interest in this study, delivers research on questions of interest to the USGCRP and the broader discussion group. 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 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.

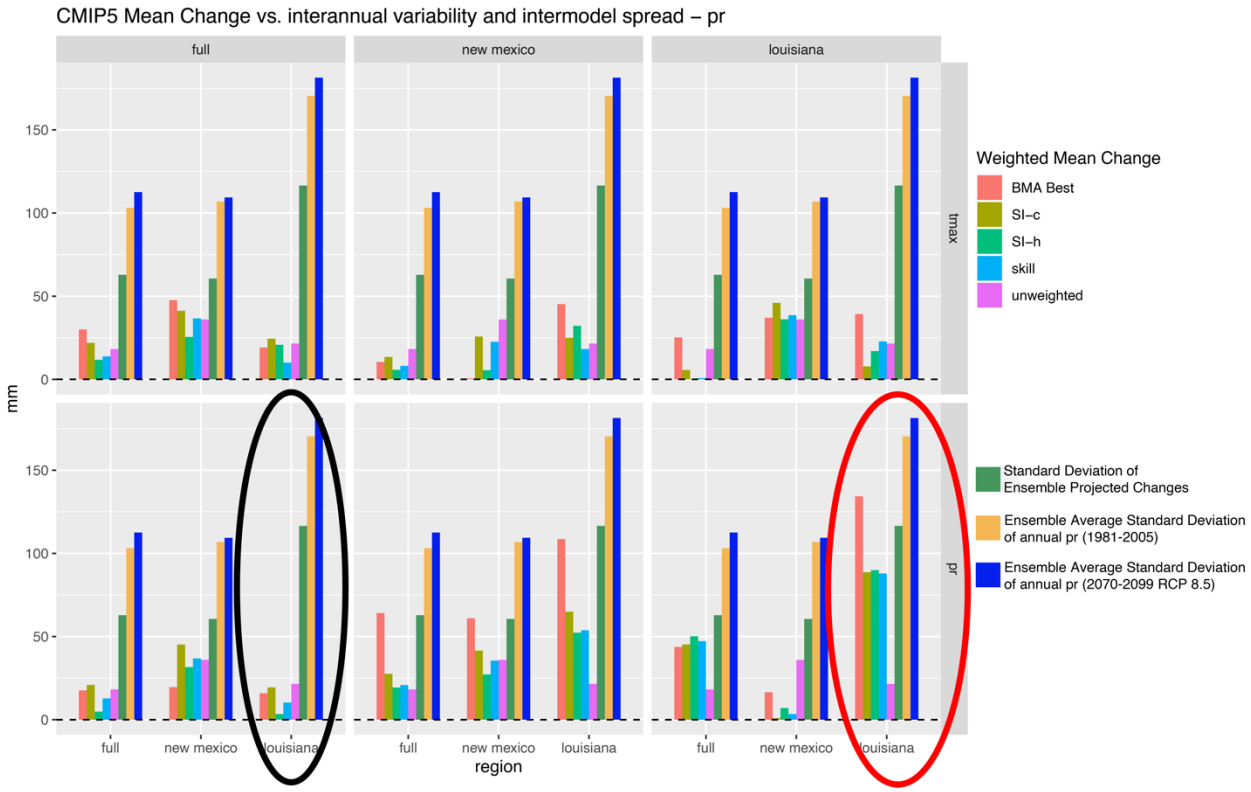
Given the above, we believe that our study is timely given the debate in the community regarding the use of multi-model weighting.

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, the revised manuscript includes an approximate analysis and an additional figure will be placed after Figures 5-6 with additional discussion.

The new Figure 7 (a portion of Figure 7 is included below) includes the absolute value of the ensemble mean changes in comparison with the standard deviation of the projected changes of the ensemble (representing the ensemble mean change), 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. 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 comments discussed here are included and expanded upon in section 3.3 of the revised manuscript.



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 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 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. The following is included in the conclusions of the revised manuscript: "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."

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 (one of several cited in the study – Skahill et al. 2021 – ***https://doi.org/10.3390/cli9090140****). The text has been revised to make these points clearer and incorporate additional references as suggested by Reviewer 2.*

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 have provided 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 have also adjusted the text to provide some discussion and caveats as mentioned above.

Citation: https://doi.org/10.5194/esd-2022-15-RC1

Reviewer 2 Comments

Review of 'To weight or not to weight: assessing sensitivities of climate model weighting to multiple methods, variables, and domains' by Wootten and colleagues

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

The manuscript presents a comparison between what seems to be 2 different climate model weighting schemes in 4 different setups. Weights are based on 2 different variables, 3 different domains, and 2 different model ensembles resulting in 48 different sets of weights. These are applied to the same 2 variables and 3 regions, resulting in 288 sets of weighted ensemble means discussed in the paper. The differences between these setups are visualized, described and discussed. In the second part of the manuscript several recommendations are given regarding how to apply weighting methods in general.

With this manuscript the authors set out to answer a big question as stated in their title: 'to weight or not to weight'? And in more detail (line 77): 'Should model weights be developed separately when investigating different climate variables? Should model weights be estimated separately when investigating different domains?

I would like to propose a somewhat provocative argument about this aim: With the setup suggested here it is impossible to answer these questions. To advocate for (or against) weighting future projections in whatever way one would need to show an added value of the weighted ensemble compared to the unweighted one (for example increased skill by some metric). This is notoriously hard to prove (some would say impossible) as we do not know the ground truth in the future. Approaches have been suggested to circumvent this problem, at least partly. These include out-of-sample validation of weighted ensembles in the historical period where there are still observations available or model-as-truth approaches. None of this is done in this manuscript. The authors merely provide an extensive comparison of the effect of different weighting setups. As far as I can tell, most of the recommendations on weighting provided in the second part of the manuscript are not connected to the results presented (which mainly show relative difference between the different methods employed and as such can not answer the questions posed).

Thanks again for these comments. The aim of this study is not to show that we find the 'best' future climate change projection, but to highlight the different approaches to estimate model weights and the resulting effects on the estimates of projected climate change and provide an extensive comparison of the effect of different weighting setups. 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).

We thank the reviewer for pointing out that our recommendations are disconnected from the results. In this study, we show that, yes, for different variables and domains that different weights need to be estimated. And this ultimately produced different climate change signal. Although our results do not directly provide a way forward for model weighting (which is an extremely difficult problem to solve), this extensive study showcases how different strategies can impact estimated model weights and their respective climate change signals, which makes this a study that has not been done in this magnitude and provides comprehensive guidance for future studies and impact assessments which are considering incorporating model weighting. We have revised significant portions of the paper, particularly in Section 4, to better connect the results of the analysis to our recommendations and findings.

My second main criticism is that the results presented in this work are not really new or surprising. The authors basically show that weights based on different variables and regions differ - but this is what they are designed to do. If weights for different regions and variables were all identical there would be something wrong with the model ensemble or the setup of the weights, right? Finally, the authors give several recommendations but these are more of a general nature and I had a hard time connecting them to the specific results presented. As a matter of fact several of the arguments have been made before and are not connected to any of the work done here (for example the discussion about spatial coherence in line 363f).

We again agree with the reviewer that our recommendations can be better connected to our results. The primary purpose of this study was to provide the extensive and comprehensive comparison of the setups associated with multiple weighting strategies in a manner and extent that, to our knowledge, has not been done before. This is discussed more in our response to the

following comment. We have revised the manuscript to connect the specific recommendations more carefully to the results of this study.

Given the current debate on model weighting in the community, and the general sense of not knowing a path forward, an extensive and comprehensive study like ours is just what the community needs right now. Even though there is no direct path forward that is reported in our study, we provide an extremely large experimental matrix that other authors and scientists can draw on when asking for their own application whether "to weigh or not to weigh". In addition, the conclusions include general recommendations based on the author's experience and some additional specific recommendations more clearly tied to this study have been included.

In addition, I find the heavy self-citation, partly ignoring large chunks of other literature, employed in this paper somewhat strange. I would encourage the authors to put their work better into the context of the international scientific literature (for example lines 35, 56-67, specific comments on lines 321, 325, 380). In addition, the authors state at several points that their study is the first to 'assess the sensitivities of the model weights and resulting ensemble means to the combinations of variables, domains, ensemble types (raw or downscaled), and weighting schemes used' (e.g. line 284). This might be so but what is the gain? Again, I am not surprised that the selection of the metrics used to inform the weights has an influence on the weights. If that was not so, weighting would hardly make sense, right?

Our own work is included in this paper because those other studies are highly relevant to this current effort. However, we also refer to over a dozen other studies in the broader literature that specifically address model weighting as well:
Sanderson et al. 2015, 2017; Knutti, 2010; Knutti et al. 2017; Weigel et al. 2008; Pena and Van den Dool, 2008; Min and Hense, 2006; Robertson et al. 2006; Shin et al. 2020; Brunner et al., 2020ab; Kolosu et al. 2021; Skahill et al., 2021. We have incorporated additional studies to bolster our study further.

The number of sets of weights (48) and the number of weighted means produced (288) is in my opinion too excessive. The authors should pick a few representative and/or interesting examples to discuss and move the rest of the results into the supplement. I found it almost impossible to follow the discussion of methods, domains, variables and ensembles that are in turn applied to ensembles and domains.

We thank the reviewer for this recommendation. However, as mentioned in the response to a previous comment, the purpose of this paper is to provide an extensive and comprehensive comparison of the effect of different weighting setups and allow others to assess the question "to weigh or not to weigh" in their own application. As such, the extensive collection of weighting schemes and strategies is critical to include. In addition, the main text includes only a subset of the results, and other results are included in the supplemental materials.

Finally, I would like to urge the authors to provide at least a basic description of the methods which are at the core of this manuscript. As it is, the reader is merely referred to three papers by the authors (Wootten et al. 2020a, Massoud et al. 2020a, 2019) for more information. For a potential reader (or reviewer) it would be quite convenient to have a more self-sustained paper with at least the basic setup of the methods clearly described and only the details requiring reading several more papers.

We agree and the basic equations and explanation of the setup of the different model weighting strategies in the supplemental material for the readers and reviewers. There is also more detail on the weighting schemes added to section 2.4.

Overall this manuscript has several major problems raised above beside the many specific issues outlined below and I do not think that it can be published without a major overhaul. This should include, most importantly, clearer formulated research questions that can be addressed in the manuscript and a clear separation between conclusions based on results and general recommendations based on the authors experience. In addition, a better representation of already existing literature and more focused plots (showing only a subset of cases) would help the manuscript.

We thank the reviewer for the comment. We have made a clear distinction in the revision for specific recommendations based on this paper and general recommendations from the authors experiences. We have also included different referenced papers throughout the text, where relevant. Furthermore, the main manuscript represents only a subset of the figures that are deemed necessary to tell the story of this paper, and all the additional figures and analysis are provided in the supplemental material.

Minor comments

title: the quite narrow focus on parts of the United states should be reflected in the title.

The title now reads "To weight or not to weight: assessing sensitivities of climate model weighting to multiple methods, variables, and domains in the south-central United States"

line 16: At this point I am confused about the terminology. My a priori assumption is that there are different weighting schemes and in addition each scheme might use different variables to calculate the weights. Here they are mixed up so either the authors use another terminology (then they should make it clear) or this should be reformulated.

This sentence now reads "Results suggest that the model weights and the corresponding weighted model means are highly sensitive to the weighting scheme that is applied."

16-21: I am not sure what the authors point is here as this behaviour seems to be totally expected? Is the important point not rather that the metrics (including variable and region) the weights are based on need to be well-justified? With cherry-picked metrics it is probably possible to achieve any kind of weighting, right?

When applying model weighting to future climate projections, it is unclear how the estimated model weights will impact the resulting projected climate change signal. This is the reason for the broad application in our study, and two clearly different examples are listed in the abstract to highlight this difference.

28: please introduce NCA

Thank you, this now included.

line35: can the authors please cite a broader sample of the literature not limiting it to their own publications (assuming that they are not the only ones publishing on that topic)?

We that the reviewer for noticing this, and we agree. We have included other references here and throughout the manuscript with other works that are relevant to this topic.

39: I would argue that the ensemble mean is not representative for the members (one of the reasons why we need weighting)

The sentence in question is revised to the following: "Large and local scale assessments can make use of the entire ensemble of climate projections (composed of global climate models [GCMs]), or make use of the unweighted ensemble mean."

44: model weights themselves can not have any skill I would argue

Thank you, this statement now reads: "Projections based on model weights derived from historical skill have been shown to have greater accuracy than an arithmetic multi-model mean in many cases, provided that there is enough information to determine a weight for each model."

47: As a matter of fact the idea of independence weighting has not only come up in the last few years and is, e.g., mentioned in Knutti 2010 which is cited by the authors in the line before.

We thank the reviewer for this comment, this now reads 'more recently'

60: ‘performance skill of atmospheric rivers globally’ again, what would be the skill of an atmospheric river? I assume the authors refer to the model skill in simulating atmospheric rivers?

Thank you, this statement now reads ‘performance skill of the models to simulate atmospheric rivers globally.’

69: Knutti et al. 2017 did not base their weights (only) on precipitation as seems to be suggested here

Thank you, this sentence now reads: “Some studies have applied model weighting to a certain variable or to multiple variables and went on to investigate climate change impacts for other variables (e.g. temperature or streamflow) (c.f. Knutti et al., 2017; Massoud et al., 2018).”

70: What is a common variable?

This sentence now reads: *“The National Climate Assessment had previously considered weighting based only on commonly used climate variables (e.g. precipitation and temperature, Wuebbles et al., 2017), but discussions to use additional variables are currently ongoing.”*

73: ‘Other studies have applied model weighting to a specific domain (e.g. globally) and went on to apply the developed weights on a different domain (e.g. North America or Europe) (Massoud et al., 2019).’ This sentence does not seem to make sense. Do the authors mean that they have calculated the weights based on metrics in one domain and then applied them to projections for another domain? Please reformulate this to make in more clear.

The sentence now reads: “Other studies have calculated weights based on metrics in one domain (e.g. globally) and then applied them to projections for another domain (e.g. North America or Europe) (Massoud et al., 2019).”

79: I am not convinced by the relevance of these research questions and their implications. For a weighting method to have skill the weights need to be based on metrics that are physically and statistically connected to the variable that the weights are applied to. See for example the discussion about emergent constraints in Hall et al. (2019; 10.1038/s41558-019-0436-6). In lack of a certain variable in a certain region that is informative for all other variables in all other regions the answer to both questions has to be yes, without any further analysis from a purely skill-based perspective I’d argue. There might be other considerations against it but they depend on the application (and are, hence, independent on the outcome), such as physical and spatial consistency of the weighted distributions.

Thank you for this comment. We believe the reviewer is referring to the physical connection between temperature and precipitation here. This study deliberately constructed the experimental matrix to examine the sensitivity of different model weighting strategies (i.e.

including weighting on temperature and estimating future precipitation or vice versa) precisely to build toward addressing some of the concerns the reviewer mentions above. In addition, the authors of this study are all directly and indirectly involved with the discussions for the Fifth National Climate Assessment (NCA), led by the United States Global Change Research Program (USGCRP), surrounding downscaling and multi-model ensemble. There are 10-15 people representing multiple agencies of the United States government discussing issues associated with downscaling and model weighting during bi-weekly meetings over the past two years. The effort in this study, and the questions of interest in this study, delivers research on questions of interest to the USGCRP and the broader discussion group. This includes questions on the sensitivity of model weighting strategies (such as weighting on temperature and estimating precipitation, or vice versa).

83: is the entire domain the combination of Louisiana and New Mexico or are there additional regions not covered by them? Maybe indicate the sub-domains in figure 1?

Labels for the states used have been added to Figure 1. In addition, the sentence in question is revised to read as the following: "Furthermore, we use two sub-domains, the states of Louisiana and New Mexico, alongside the south-central U.S. study region."

84: can the authors motivate why they use CMIP5 instead of the newer CMIP6?

The following has been added to Section 2.2: "CMIP5 GCMs are used in this study because LOCA downscaling with CMIP6 was not available at the time of this writing."

line 106/figure 1: I am not familiar with the term 'high temperature' is this the same as 'maximum temperature' which is (in my opinion) a frequently used term? And what is annual high temperature? Is it the maximum over different annual mean temperatures or the maximum of the maximum daily temperature or something else entirely? Over which time period?

The sentence in question now reads: "Average annual precipitation in the southeast portion of the domain can be eight times higher than drier western locations and average daily high temperatures can reach 40°C (Figure 1)." The caption for Figure 1 has also been modified to match.

115: Just so that I understand correctly, also the CMIP5 models are interpolated to 10km – corresponding to a resolution much finer than the native one?

Yes, this sentence now reads as: "To facilitate analysis, the data for each ensemble member and the gridded observations are interpolated from their native resolution to a common 10 km grid using a bi-linear interpolation similar to that described in Wootten et al. (2020b)."

section 2.4: The authors aim to provide a comparison of different weighting schemes but here these weighting schemes are not introduced at all requiring the reader to read several other papers to get any information at all about them. Please provide at least the basic properties and differences between the schemes investigated in this study.

We agree, and the basic equations and explanation of the setup of the different model weighting strategies are provided in the supplemental material. More detail on the weighting schemes is also included in section 2.4.

141: 'The Skill strategy utilizes each model's skill in representing the historical simulations' I assume the authors mean 'historical observations' here?

The reviewer is correct, 'historical simulations' has been corrected to 'historical observations.'

150: If the authors write 'weighting schemes are applied' here they mean that weights are calculated is that correct? I find this confusing since they also write 'applied' for the process of calculating a weighted mean of the future projections. Could the authors try to find a less ambiguous language throughout the manuscript?

Yes, we added more clarification of the difference between weighting scheme and weighting strategy throughout. The first portion of section 2.5 now reads: "Each weighting scheme (Skill, SI-h, SI-c, and BMA) is applied to both ensembles (CMIP5 and LOCA) and three domains (south-central U.S., Louisiana, New Mexico) to fill out an experimental matrix of weights, representing a collection of weighting strategies. As a result, for each weighting scheme (skill, SI-h, SI-c, and BMA) and ensemble (CMIP5 and LOCA), there are six sets of weights produced (i.e. 3 regions and 2 variables). One example of a weighting strategy would be the BMA weighting scheme used on the CMIP5 ensemble trained on tmax for the entire domain. Another weighting strategy example would be a skill-based weighting scheme used on the LOCA ensemble trained on precipitation in Louisiana."

156: '(ensemble choice x weighting methods choice x variable choice x domain choice = 2 x 2 x 3 x 4 = 48).' This is mixed up please correct

Yes, ensemble choice x variable choice x domain choice x weighting methods choice = 2 x 2 x 3 x 4 = 48. This is corrected in the text.

166: I am not sure I understand why the weights are applied to the sub-domains separately. The resulting maps should identical to the corresponding region in the full region, correct?

The reviewer is correct that some will be identical. The study makes this point on lines 235-238: "However this maximum number of ensemble means resulting from the experiment contains several duplicates. For example, when using the same set of weights, the resulting ensemble mean in a subdomain will be the same as the resulting ensemble mean from the same portion of the full domain. As such, the actual number of ensemble means is smaller than 288." The following has been added at the end of the section to clarify: "However, we also note that there will be several duplicates in the experiment. For example, when using the same weighting strategy, the resulting ensemble mean in a subdomain will be the same as the resulting ensemble mean in the same portion of the full domain."

figure 3: 'grey dots' do the authors mean the red dots?

Yes. The figure caption has been adjusted accordingly.

figure 3: as a general question: should weights not be normalized in order to be comparable across the different cases?

Yes, the weights are normalized, which is why each y-axis goes up to 1. This has been clarified in section 2.4.

182: 'One observation seen in these weighting combinations is that the weighting schemes themselves are all sensitive to the ensemble, variable, and domain for which they are derived.' I do not agree with this statement in this general form, could the authors provide a bit more detail? To give just two examples: the bcc model gets consistently low weights for all cases and the low weight of NorESM1 (among many other models) is not sensitive to variable and domains but only to the ensemble.

This is true for models that get low weights consistently, but the comment in our study refers to which models that might have higher weights in some strategies but lower weights in others, and these are in effect the models that provide information to the future projections. This sentence now reads: "One observation is that the weighting schemes themselves are all sensitive to the ensemble, variable, and domain for which they are derived in terms of which GCMs are given the highest weight."

185: what are 'model combinations'

The highest weighted models that result from each weighting strategy is listed in Table 1. This sentence is revised to read: "From Table 1, no model appears in the top three for all weighting strategies."

189: is this surprising given that (from what I understand) BMA is a structurally different method while the other three are variants of the same method?

Yes, that is true. We will include a statement to point this out.

193: I would tend to say the colour is red not orange. How is significance established for this case or is this just a qualitative statement? Then maybe use a different wording.

Agreed, we should use a word other than significantly. This is changed to 'noticeably'.

195: what are differences 'within each combination of ensemble, variable, and domain'?

This sentence was deleted in response to the following comment.

197: what does 'combinations' refer to here?

Combinations refers to each weighting strategy (i.e., the weighting scheme and domain, variable, and ensemble used). We have revised this to refer to them explicitly to weighting strategies.

206 'Similar to the CMIP5 ensemble in Figure 3, the BMA weights tend to be larger for the highest weighted models in the LOCA ensemble compared to those derived with the Skill, SI-h, and SI-c schemes' Can the authors speculate on the reason for this behaviour?

The following is included in the text to speculate on this behavior: "We speculate that the reason for this is because the Skill, SI-h, and SI-c strategies involve the 'skill' of each model when estimating weights, and since the LOCA downscaled ensemble is bias corrected, most models have similar skill and therefore similar weights."

212: 'the weights for the LOCA ensemble [tmax, Louisiana] generally range from 0.025 to 0.05' Do the authors mean 0.25-0.5? Otherwise it is impossible to see this in the figure 4. The authors might want to explain the notable exception from this. How is 'BMA best' calculated from the 100 iterations of BMA? How is a case like MIROC with a median of about 0.25 but a best of close to 0 possible?

This means that, generally, most model weights for the LOCA/tmax/Louisiana strategy are between 0.025 and 0.05. For the second question, the MIROC model has a distribution of weights from BMA that includes lots of high values, but when sampling the combination that

produces the 'best' simulation (i.e. lowest bias compared to historical observation), the sampled combination of model weights just happens to be very low for MIROC. The following text is added at Lines 332-335: "The BMA best combination is the single set of model weights from the BMA posterior that creates a weighted model average that has the best fit to the observations. Although all the samples of model weights from the BMA posterior have an improved fit compared to the original ensemble mean and provide a range of model weights as shown in the BMA distributions in Figures 3 and 4, the BMA best combination is considered the best of all these samples."

223: what is 'co-dependence between models in an ensemble'? Does 'Skill' account for dependence at all as seems to be suggested here?

Co-dependence means when two models provide similar information to an ensemble average. The 'Skill' method does not consider co-dependence, and we will remove this strategy from this sentence. The 'SI-h' and 'SI-c' methods consider co-dependence in the historic and future simulations respectively. The BMA method down-weights models that provide similar information. See supplementary section of Massoud et al., 2020a that discusses this concept in more detail. More detail on this has also been added to section 2.4 of the manuscript.

225: 'BMA tends to be the most sensitive' could this somehow be quantified?

This is observed visually in Figures 3 and 4 in particular, where the magnitude and variability of the weights is much larger for BMA between the different variables, domains, and ensembles, then for the other three weighting schemes. We will clarify this in the text.

239: So why not just not use the sub-domains at all?

The results are only identical for a small sample of these different tests, but they are still unique for most examples.

figure 5: is there are particular reason for selecting a base period of 25 years and a future period of 30 years? What do the boxes, whiskers represent?

There are several other projects in the study region that use 1981-2005 as the historical period (including Wootten et al. 2020b). As such, this historical period was used to facilitate comparisons. This is explicitly stated in Section 2.2. Second question, the boxplots in Figure 5 represent the inter-model spread for both variables for both ensembles. The left-hand boxplots represent this for the historical period, and the right-hand side represents the projected changes from both ensembles.

271-281: I am not sure I understand why this paragraph is here? Should the reader look at and understand all the figures listed here? Or is this just an outlook? The authors might want to consider dropping it.

This section is an outline of the following sections to help guide the reader. It also points out which extra analyses and figures are included in the supplemental material. This section is left in to guide the reader.

321: Maybe the authors could give some examples of the literature that does exist? To give just a few examples (there are more): 10.1029/2019GL083053, 10.1088/1748-9326/ab492f, 10.3389/frwa.2021.713537, 10.1029/2020JD033033

In response to the major comments of both reviewers, Section 4, the Discussion section, was significantly restructured. The sentence where this comment was made was deleted. The references the reviewer provided are referenced in the introduction in Section 1.

325: Again, there are counter-examples that might be good to mention here: 10.5194/acp-20-9961-2020, 10.3389/frwa.2021.713537

In response to the major comments of both reviewers, Section 4, the Discussion section was significantly restructured. The sentence mentioned in this comment was removed as a result. The references the reviewer provided are incorporated in Section 4.4 of the revised manuscript.

327: 'Third, for situations where projections are provided to impact models, does this type of study need to be repeated using impact model results' I don't think I understand this question.

Thank you for this comment. In response to the major comments of the reviewers, the manuscript was restructured, and these research questions were stated more clearly in the introduction. This question is revised to the following: "Should a sensitivity analysis with multi-model weighting strategies be repeated using impact model results?"

334: This is not correct so generally, see references above.

As mentioned above, Section 4 was significantly restructured. The sentence mentioned in this comment was revised to the following and appears at the start of Section 4.3: "At the time of writing, discussion surrounding the use of weighted multi-model ensembles has been traditionally limited to climate model developers and the production of national or international climate assessments, but is beginning to be used in impact assessments."

342: Who are these ‘others’? Please provide references

This statement now reads “Based on expert discussions surrounding downscaling and model weighting, the NCA is now considering weighting based on model climate sensitivity as opposed to traditional model weighting approaches.”

349: Why does a unweighed mean over-favor certain models? I would assume that by definition in an unweighted case all models are treated equally.

You are right, what is meant here that certain models are provided higher weights than they should be receiving. This sentence now reads the following: “An unweighted mean will allow models with large biases and co-dependencies regardless of the domain or variable of interest larger influence in either climate models or impact assessments.”

354: applying multiple methods as suggested here might lead to contradictory results, can the authors say something about what a user that tries to get a single answer should do in such a case?

Thank you for this question. That is the point we are reaching in our study, there is no ‘single answer’ and if the user wants a true accounting of the uncertainty to the question at hand, then the user should use many strategies if it is feasible to do so. This point is emphasize in the revised manuscript.

380: ‘Climate model evaluations and national assessments typically focus on the continental United States or North America.’ There are assessments also for other continents.

Thank you, this is corrected at the beginning of section 4.

394: Is this recommendation somehow connected to the results shown in this manuscript or just the authors opinion?

This reflects a combination of both the results in the manuscript and the authors experience. The manuscript has been revised to delineate connections more carefully between our recommendations and results in the manuscript.

346: ‘a multi-model ensemble of climate projections should incorporate model weighting’ The ensemble itself can not incorporate weighting I’d argue. Weights can only be applied once the ensemble is aggregated along the model dimension (for example by calculating a multi-model mean).

The statement the reviewer is referring to is on line 436-437. The reviewer is correct, the sentence now reads "…efforts using a multi-model ensemble of climate projections should incorporate model weighting."

446 (recommendations): Could the authors connect these recommendations to their results?

Thank you. Section 4 has been restructured to connect the recommendations to the results of analysis. We have also connected the recommendations to our numbered questions in the introduction.

456: how can a domain be small compared to internal variability?

Thank you for this question. What we mean is that the spatially aggregated internal climate variability of a smaller region is much larger than that of a larger domain, which make the model averaging results less coherent for a smaller domain as they would be for a larger domain. The extreme case of this is applying model weights using a single grid cell. This sentence is edited to the following: "First, this study makes use of domains that are fairly small where the spatially aggregated internal climate variability is larger than that of a large domain."

Citation: https://doi.org/10.5194/esd-2022-15-RC2