Earth System Dynamics Discussions

Earth Syst. Dynam. Discuss., author comment AC2
https://doi.org/10.5194/esd-2022-15-AC2, 2022

Reply on RC2

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-AC2, 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 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 will do a better job at describing this in the paper and will make our recommendations more clearly tied to the results in the manuscript.

"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 will be refining the manuscript to connect the specific recommendations more carefully to the results of this study and will better delineate general recommendations.

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 will also be 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 will improve on the inclusion of these citations throughout the manuscript, wherever relevant, and will incorporate additional studies to bolster the references in 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 will provide the basic equations and explanation of the setup of the different model weighting strategies in the supplemental material for the readers and reviewers.

"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 will make a clear distinction in the revision for specific recommendations based on this paper and general recommendations from the authors experiences. We will also do a better job of including 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 will be 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 will be altered to read "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 will be revised to read "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, we will include this.

"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 will enhance the references listed here 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)"

We are not entirely clear about what the reviewer is stating here. However, we believe that the reviewer is referring to the unweighted ensemble mean, which makes sense to us in the context of that sentence. Given our assumption, we agree with the reviewer, and the study points out the recognition of the need to use weighting in the following paragraph. However, we will make it clear in the sentence here that we are referring to unweighted ensemble means.

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

Thank you, we will fix this statement to read 'projections based on model weights'.

"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, and we will change this read '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, we will fix this statement to read '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, we will fix this statement.

"70: What is a common variable?"

We will refine this statement to refer to common climate variables (e.g., precipitation and temperature, Wuebbles et al., 2017)

"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 it more clear."

Yes, we will fix this statement to read 'calculated weights based on metrics in one domain and then applied them to projections for another domain.'

"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. To satisfy this concern, we will include a statement in the text about the connection of temperature and precipitation in a physical sense, and why investigating the weighting based on one variable to estimate projections on another variable is important. For example, IPCC or NCA reports look at the projected changes of one variable, when the weighting might have been based on another variable, which is precisely why we wanted to investigate this question. Furthermore, we will incorporate a brief discussion of the emergent constraints issue in our introduction in the context of our study.

"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?"

We will add labels to Figure 1 to point out the states of New Mexico and Louisiana for the audience along with the other surrounding states.

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

At the time of this writing, CMIP6 downscaled with LOCA was not available. Therefore, we focused on CMIP5 to facilitate the comparisons between a raw global model ensemble and a downscaled global model ensemble. We will clarify this in the text.

"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?"

Average high temperature refers to the climatological average of daily high temperatures over the region. The time period used is the same historical period (1981-2005) as in the analysis. This will be clarified in the text.

"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, however, this is done only to facilitate analysis and calculation of the multi-model weights and ensemble means. The 10km resolution matches the resolution of the observation data used. We will clarify this in the text.

"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 will be provided in the supplemental material.

"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, we will fix this in the text.

"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, the reviewer is right we should find another term to refer to weighting schemes being utilized. We will fix this in the text.

"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 will be 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." We will add similar language at line 166.

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

Yes. The figure caption will be adjusted accordingly.

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

Yes, they are, which is why each y-axis goes up to 1. This will be made explicit in the text.

"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. We will clarify this in the text.

"185: what are 'model combinations'"

The highest weighted models that result from each weighting strategy is listed in Table 1. We will revise the sentence to refer to weighting strategies as opposed to model combinations or weighting combinations to remove any confusion.

"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. Perhaps 'noticeably'.

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

Per the following response, this sentence will be removed in the revision.

"197: what does 'combinations' refer to here?"

Combinations refers to each weighting strategy (i.e., the weighting scheme and domain,

variable, and ensemble used). This will be revised to remove the confusion.

"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 authors believe this is because BMA tends to estimate much higher weights for a few selected models and lower weights for all other models. Whereas the other weighting schemes have a more balanced outcome as to how much weight each model receives. We will incorporate brief comments to speculate on the behavior in the revision.

"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.

"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. We will discuss this more in this paper in a new supplemental section that provides more details on the model weighting strategies.

"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. 2020). As such, this historical period was used to facilitate comparisons. This will be made explicit in the text. 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 will be 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"

We thank the reviewer for pointing us to these works. We will take a closer look at them and incorporate them into our paper.

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

Again, we thank the reviewer for pointing us to these works. We will take a closer look at them and incorporate them into our paper.

"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. We will edit the sentence for clarity.

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

Thank you, we will fix this statement

"342: Who are these 'others'? Please provide references"

Thank you, we will fix this statement by providing proper references.

"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. We will edit the text to clarify this point.

"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.

"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 true, and we will fix this statement.

"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. In line with our previous responses, we will be revising 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 will be revised to read "…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. We will revise to clarify how our recommendations are connected to the results.

"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 makes the model averaging results less coherent for a smaller domain than they would be for a larger domain. The extreme case of this is applying model weights using a single grid cell. This point will be clarified in the revised manuscript.