Earth System Dynamics Discussions

Earth Syst. Dynam. Discuss., referee comment RC2
https://doi.org/10.5194/esd-2022-15-RC2, 2022

Comment on esd-2022-15

Anonymous Referee #2

Referee 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-RC2, 2022

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

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 visualised, 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).

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).

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?

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.

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.

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.

Minor comments

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

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.

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?

28: please introduce NCA

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)?

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

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

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

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?

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

70: What is a common variable?

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.

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.

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?

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

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 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?

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?

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.

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

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?

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

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?

figure 3: ‘grey dots’ do the authors mean the red dots?

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

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.

185: what are ‘model combinations’

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?

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.

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

197: what does ‘combinations’ refer to here?

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?

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?

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

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

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

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?

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.

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

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

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.

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

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

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?

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

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

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).

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

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