Earth Syst. Dynam. Discuss., referee comment RC1
https://doi.org/10.5194/esd-2022-15-RC1, 2022

Comment on esd-2022-15

Anonymous Referee #1

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-RC1, 2022

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.

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.

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)?

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).

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).

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.

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.