Earth Syst. Dynam. Discuss.,
https://doi.org/10.5194/esd-2020-23-RC2, 2020

Earth System Dynamics
Discussions

Interactive comment on "Reduced global warming from CMIP6 projections when weighting models by performance and independence" *by* Lukas Brunner et al.

Anonymous Referee #2

Received and published: 23 July 2020

General Comments

Some models are more consistent with historical observations than others. In climate projection, it makes intuitive sense to give more weight to the models that are more consistent with observed climate shifts and less weight to models that are less consistent with observed climate shifts.

But how?

This paper reports on a method of assigning model weights that relies on two distinct distance measures: the distance of models from observations and the distance of models from other models. The method requires the specification of two parameters

that determine how each of these distances are turned into model weights. The method for determining the parameter associated with inter-model distance is poorly explained (see specific comment 8 below). The method for determining the parameter associated with distance from observations is also poorly explained, but for many experiments, involves future-time-pseudo-observations from the future states that are the objective of the prediction (see comment 10 below). In other words, the tuning method appears to render the tests of the method to be of the "in-sample" variety. To weaken the degree of "in-sampleness" an additional test is performed using CMIP5 runs. However, since one expects many of the CMIP6 models to be closely related to the CMIP5 models, there are strong reasons to believe that this test is not truly "out-of-sample" either.

Even with the use of "future-time-pseudo-observations" in the tuning procedure, the improvements from this weighting scheme seem very modest in comparison with, for example, those obtained in Abramowitz and Bishop (2015, J. Climate) – (using a a method that solely required historical observations for the weights). The revised paper should include some attempt to compare/contrast/explain the Abramowitz and Bishop results.

A superficially appealing feature of the method is that it gives more weight to models that are both skillful and statistically independent of other models. However, this independence is just described in terms of inter-model distance and not in terms of the independence of the model error. Is there some unstated proof that increased inter-model distance equates to increased model error independence? (It seems easy to think of counter examples). As demonstrated in Bishop and Abramowitz (2013), it is the independence of the error of the individual models comprising an ensemble forecast (as measured by inter-model forecast error correlation) that increases the predictive power of the ensemble. The revised paper needs to address the issue of the relationship or lack of relationship between inter-model distance and model error independence.

After applying the method to the CMIP6 ensemble members, the authors find reduced

warming relative to the simple sample means of CMIP6 ensembles for the high and low CO2 concentration scenarios considered. However, any confidence in this prediction must be strongly tempered by the "in sample" circular- nature of the testing and tuning procedures used by this method.

My overall recommendation would be that the paper be returned to the authors to address the specific comments below and to include results from experiments in which only historical observations (or model-based-historical-pseudo-observations) were used to determine the weights. This constitutes major revision.

Specific Comments

1. Line 16. Consider explaining what TCR is in the abstract to appeal to a broader audience.

2. Line 31. Do you mean model uncertainty, unknown model climate error, unknown model-climate-sensitivity-to-CO2 error or model climate differences? We know what the model is, and we can determine its climate past, present and future by running it. We can also determine the differences between the climates of different models. Given the limitations of the spatio-temporal distribution of observations, the uncertain thing is the actual climate both past, present, and future, is it not?

3. Line 35. Lorenz, the father of chaos theory, argued that while the accuracy of weather forecasts was limited to a few weeks the climate of a system was not sensitive to specified initial conditions and could be known provided the forcing on the system was known. I guess "climate" in the sense of Lorenz refers to the statistical description of the attractor of the chaotic system. When you refer to "internal variability" do you just mean slow modes of the model's chaotic attractor that might possibly be confused with a change in the mean of the model's attractor if the ensemble size was too small?

4. Line 102: I'm guessing you are referring to Section 3.2 of Brunner et al., 2019. Is that correct? If so, please state this in the text. Your wording suggested that you had

estimated an observation error variance. However, on reading Section 3.2 of Brunner et al., 2019, I'm now guessing that you are referring to how your derived weights change depending on which subset of all observations you use. Are you suggesting that the reason for your weights changing is because the observations have different errors? Can you rule out the possibility that your weighting scheme isn't just over-fitting each individual observational data set? In any case, the revised paper needs to clarify whether in fact you are referring to the size of the change in weights associated with using differing observational data sets. Also, the observed values are known. They are not uncertain. The errors of the observed values are unknown. It is the observational error that is uncertain.

5. Line 145. "We want to . . ." If there was a hypothetical user of the climate projection that only cared about temperature trend and not about year-to-year variability, might you not be doing them a disservice by down-weighting members that have an excellent temperature trend but poor inter-annual variability? Consider changing to "We choose to . . . "

6. Line 147-149. Equations should be added to precisely describe these observation derived quantities – perhaps in an appendix or supplementary material.

7. Line 170. You must state what was used as a proxy for a perfect model. I would think that the derived sigma_D must be related to the ensemble variance of the model states around the time averaged state. That quantity will depend on the model will it not? Please clarify.

8. Line 183. I looked at Section 2.3 of Brunner et al., 2019 for an explanation but Brunner et al. (2019) just directs the reader to Lorenz et al., 2018. Your work needs to be reproducible. When referring to another paper for a key explanation, you must give very specific information about where in the paper the explanation resides (e.g. a section number) to ensure reproducibility. You have not done this.

9. Line 191. The method used to evaluate performance given here seems almost

identical to that given in Abramowitz and Bishop (2015) but no reference is given to this paper or others that may have used this approach before. Such literature is relevant and should be cited.

10. Line 200-205. Here, we learn that sigma_D weights are determined in part from information from a place that is inaccessible in reality: the future. Only model futures are accessible. By line 205 we learn that the model future states (rather than observations) are, in fact, an integral part of choosing the weights. This is a significant departure from many other observation-based methods for improving ensemble forecasts and projections. The use of future time observations in the training causes all of the associated tests to be "in-sample" tests – dramatically reducing their trustworthiness. Since the CMIP5 models belong to the same general class of human produced climate simulators they can barely be considered "out-of-sample". Please comment on the limitations of this approach. In addition, you have not clarified how the method of tuning for future states interacts with the method to determine sigma_D referred to on line 170 (see previous comment).

11. Line 266-280. Here we learn that the method is very prone to creating decreased skill relative to the multi-model unweighted mean. This negative result is in contrast to the positive results found in Abramowitz and Bishop (2015) using the method of Bishop and Abramowitz (2013).

Interactive comment on Earth Syst. Dynam. Discuss., https://doi.org/10.5194/esd-2020-23, 2020.