Earth Syst. Dynam. Discuss.,
https://doi.org/10.5194/esd-2020-6-RC1, 2020

Earth System Dynamics
Discussions
Open Access
EGU

Interactive comment on "Calibrating large-ensemble European climate projections using observational data" *by* Christopher H. O'Reilly et al.

Francisco J. Doblas-Reyes (Referee)

francisco.doblas-reyes@bsc.es

Received and published: 12 April 2020

This is a valuable manuscript that aims to apply ideas common in weather and climate prediction into the post-processing of climate projections, in particular with the use of large ensembles. The authors undertake an ambitious analysis to illustrate the relevance of calibrating the projection ensembles to increase their accuracy and reliability, where reliability is considered from the point of view of the trustworthiness of the probabilities formulated for the ensemble projections. The ideas are solid and clearly laid out, the text is clear, the figures adequate both in number and quality, the study is exhaustive. However, I am concerned by the description of the "out-of-sample with

Printer-friendly version

Discussion paper

imperfect model test". The method is explained in page 7 and an example is given in figure 3, but it is hard to understand how the results displayed in figure 4 are obtained. As a result, Figure 4 is a bit hard to interpret. It will benefit from a more detailed caption and better referencing in the main text. Also, the wording and the interpretation of the results can be misleading. For instance, it is hard to accept that the results of the methods lead to improvements when the verification is performed without using observations. It is also a pity that the supplementary information does not include the results equivalent to figure 4 but for precipitation. The HGR-decomp method looks promising. However, it would be really useful if the authors could provide a full illustration of how each component is calibrated before the ensemble is reconstructed, that is, to go beyond what is currently shown in figure 6. This is far from obvious and would help to understand how the method works. Figures 8 and 9 show that the mean projected change is weaker in the calibrated with respect to the uncalibrated large ensembles, particularly for precipitation. This is an important statement, although it comes with a widening of the uncertainty intervals. I wonder how these results compare to other post-processing exercises (like model selection or model weighting) performed with other ensembles in the same areas and period. I consider the manuscript needs major revisions, not that much from the technical or conceptual point of view, but more for the need to clarify some details in the text. Some minor comments follow: - p. 2, l. 24: "applied" appears twice in the sentence. - p. 3, l. 1: "that" appears twice. - p. 3, l. 14-15: This is an interesting idea, although the reader might benefit from more details about how this merging could work and why it's a relevant issue. - p. 4, l. 3-4: To what measure is the regridding affecting the results? Is LENS the ensemble with the coarser resolution? Has the regridding to a different grid been tested? - p. 5, l. 17: Correct "corrlation". Also, the sentence is incomplete. - p. 6, l. 9: Can you say a bit more about the resampling done. For instance, is it performed with or without replacement? - p. 6, l. 13: Use "constant in time". - p. 6, l. 30: Use "to compute". - p. 7, l. 8: Remove "is". - p. 8, l. 18: Correct "signficantly". This mistake appears in other parts of the text. - p. 10, l. 24: How can the reader see the overfit of the HGR method when compared to

the HGR-decomp method? - p. 11, l. 1: This is an example of my main concern with this manuscript. The text mentions an improvement for the projected climate over the period 2041-2060. However, it's hard for me to accept that there is an improvement when no comparison with the observations (which obviously do not exist yet) is made. - p. 11, l. 17: Change "it it calibrated". - p. 11, l. 29-31: It is hard to see any changes in spread in figure 8. - p. 11, l. 32-33: I would not say that the impact of the calibration on the precipitation projections is "fairly modest". - p. 12, l. 6: Correct "preciptation". - p. 13, l. 20-27: This argument seems a bit hard to follow to me. How can we determine if a third calibrated ensemble outperforms or not the former two in terms of future projections? - The figure 4 caption mentions a 44-year verification period starting in 1917, which seems wrong. Also, in the caption the sentence "For the calibrated RMS Error, spread/error and CRPS values, the black crosses indicate where the calibration represents a significant improvement over the uncalibrated (but bias-corrected) ensemble at the 90% significance level" misses to explain what is actually tested: the median of the distribution of calibrated scores, all the scores in a single sample or anything else. Finally, what does the range of values for the uncalibrated ensemble represent? If they haven't been calibrated, do they represent the scores against the CMIP5 single models?

Interactive comment on Earth Syst. Dynam. Discuss., https://doi.org/10.5194/esd-2020-6, 2020.

Printer-friendly version

Discussion paper