Reply to RC1 (comments in blue, reply in black)

General/major comments

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 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.

We agree with the reviewer that it is important to clarify the description of the imperfect model testing. This is central to this study, so we will include an expanded description, including a schematic illustration of the process involved to arrive at the verification statistics presented in the paper. This will result in a clearer presentation of Figure 4 and the related plots. In addition, we will add the equivalent plot for precipitation to the Supplementary Information, as this may be of interest to some readers, as the reviewer rightly highlights.

Regarding the logical step between demonstrating the efficacy of the calibration in the imperfect model tests and extrapolating this when applying to the observations. We of course cannot verify this simply, but one method that might be useful would be to include some analysis of where the parameters of the calibrated observations fits with respect to the CMIP perfect model tests. We will calculate this and include the results in the supplementary material and a discussion detailing this in the revised manuscript.

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.

We agree, this is a very good suggestion. We will add a schematic to fully illustrate the processes involved, particularly as the methods become more convoluted as the paper goes on. Further discussion will also be added to describe the methodologies in a clearer and more practical manner.

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.

Yes, we agree that the reviewer that the paper would benefit from some discussion of these aspects. We will add discussion and some specifics comparisons with the results for European projections of some other multi-model methods to the revised manuscript (some of these are part of a paper that we are co-authors on and is currently in revision for publication in Journal of Climate).

Minor comments

- p. 2, l. 24: "applied" appears twice in the sentence.

Yes, this will be corrected.

- p. 3, l. 1: "that" appears twice.

Yes, this will be corrected.

- 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.

Agreed. We will add further details to this idea in the revised manuscript.

- 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?

We have tested this on a small subset of the results and the regridding only marginally affects the results. The LENS ensemble (performed at 1x1 degree resolution in the atmosphere) is generally comparable or higher atmospheric resolution than the CMIP5 models, with 30 vertical levels. The MPI-GE is performed at a relatively low T63 spectral resolution (equivalent to around 2-degree horizontal resolution), with 40 vertical levels. This information will be added to the revised manuscript.

- p. 5, l. 17: Correct "corrlation". Also, the sentence is incomplete.

Thanks for spotting this – it was a mistake and will be corrected.

- p. 6, l. 9: Can you say a bit more about the resampling done. For instance, is it performed with or without replacement?

The resampling was performed with replacement – this is a relevant detail and will be added to the revised manuscript.

- p. 6, l. 13: Use "constant in time".

Agreed, will change in the revised manuscript.

- p. 6, l. 30: Use "to compute".

Agreed, will change in the revised manuscript.

- p. 7, l. 8: Remove "is".

Agreed, will change in the revised manuscript.

- p. 8, l. 18: Correct "signficantly". This mistake appears in other parts of the text.

Agreed, will change in the revised manuscript and check for other occurrences of this mistake.

- p.10, l. 24: How can the reader see the overfit of the HGR method when compared to the HGR-decomp method?

Here we were interpreting the relatively low spread in the HGR compared with the HGR-decomp as being due to an overfitting to the reference timeseries – resulting in a consistently lower Spread/Error ratio in the HGR. This interpretation and the justification for it will be added to the revised manuscript.

- 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.

As the reviewer suggests, we of course cannot verify this simply, but one method that might be useful would be to include some analysis of where the parameters of the calibrated observations fits with respect to the CMIP perfect model tests. We will calculate this and include the results in the supplementary material and a discussion detailing this in the revised manuscript. In addition, we will edit the text to state more cautiously that the results suggest that this process may result in improved projections but that there are some important caveats.

- p. 11, l. 17: Change "it it calibrated".

Agreed, will change in the revised manuscript.

- p. 11, l. 29-31: It is hard to see any changes in spread in figure 8.

Agreed, will change in the revised manuscript.

- p. 11, l. 32-33: I would not say that the impact of the calibration on the precipitation projections is "fairly modest".

Agreed, that is not a good description. We will amend in the revised manuscript.

- 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?

Agreed, will change in the revised manuscript.

- 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?

Yes, the year here is a typo and will be changed in the revised manuscript. The significance testing was performed on the distribution of the verification scores and was tested using the Mann-Whitney U-test. Further details will be added to the revised manuscript.

The uncalibrated ensemble has only been bias corrected over the reference period (so is not strictly uncalibrated) but this needs to be stated more clearly and will be corrected in the revised manuscript.

We thank the reviewer for their insightful and helpful comments that we hope will help to improve the paper.