Reply to RC3 (comments in blue, reply in black)

General comments

This study takes methods of forecast calibration normally used for initialized seasonal forecasts and applies them to half-century scale regional climate projections. Three similar recalibration methods are applied to two single model large ensembles of RCP 8.5 simulations. The recalibration methods are tested using an imperfect model approach where CMIP5 models are used in place of observations to allow out-of-sample evaluation of the skill of the recalibrated projections. The imperfect model testing indicates that recalibration generally produces more reliable projections for future climate in Europe, and only rarely produces significantly less reliable projections. Results are qualitatively similar for both large ensembles. An important aspect of this study is the separate recalibration of dynamically decomposed components of the forecasts, which tends to produce more reliable projections that recalibrating the complete forecasts.

I congratulate the authors on presenting a fascinating idea. The manuscript is generally very clear, and I have little criticism of the imperfect model validation methodology which is very thorough. The idea proposed is can uninitialized mid-term climate projections be recalibrated to be more useful for adaptation and impact assessment using techniques from seasonal/decadal forecasting? The answer is almost certainly yes, as this study demonstrates, but with some important caveats that warrant further discussion without detracting from the novelty and potential utility of the idea.

We thank the reviewer for their positive comments and feedback. We agree that there are some important caveats and that these would benefit from further discussion in the revised manuscript – further details follow the specific points below.

The main concern is conceptual. The three recalibration methods tested are very similar, effectively differing only in their treatment of the ensemble spread. They were conceived for application to seasonal forecasts where uncertainty in the forcing and the thermodynamic response to forcing (i.e., climate change) are negligible. On decadal time scales this assumption may still be a reasonable approximation, but on longer time scales this is not the case, as is clearly visible in Figure 3 by the divergence between the CESM ensemble and the CMIP5 model. The recalibration methods used were not intended to correct for differences in forcing or response to forcing. Therefore, unless the difference over time is approximately constant (which it isn't), or can be corrected by a linear scaling of the signal (Figure 3 suggests not), then the recalibration methods tested are likely to be inadequate to the task. I do not doubt the performance improvements shown in the results, bias correction, signal scaling and correcting the ensemble spread will all improve the imperfect model predictions, but I doubt whether the projections are truly reliable.

Yes, the reviewer makes a very valid point. As we go to longer lead-times (i.e. further into the future) the errors are expected to get larger, as the error in the scaling will be amplified and the contribution of internal variability reduced. We focused on this mid-century timescale because that is the focus of our current project, however, it is important to assess how the effectiveness of the calibration changes with lead-time. We will calculate the verification over some additional future periods (e.g. 2061-2080) to examine this and include the

discussion of these in the revised manuscript (though the results may end up in the supplementary material as the paper is already fairly lengthy).

This makes the dynamical decomposition aspect of the paper all the more interesting and important. The idea appears to be to decompose the forecasts into forced and unforced components, then recalibrate each component separately using the same recalibration method. This makes a lot of sense and goes a long way to addressing my concerns above (it is still questionable whether the recalibrations employed are suitable for the forced component, however this is pardonable given the novelty of the approach). In my view, the decomposition step is critical to making the whole approach credible and needs to be introduced and motivated in the introduction, some further details of the both the decomposition itself and how the components are recombined (Figure 6) included in methodology, and possibly some additional reflection in the conclusions.

The reviewer is right to suggest that the description/presentation of the decomposition method should have been clearer – and this is also reflected in comments by the other reviewers. In the revised manuscript we will include an expanded motivation and description of the method, as well as a schematic showing the specific steps involved in the decomposition and calibration. We agree that this is an important part of our study that was perhaps not illuminated as it might have been in the previous version of the paper.

Specific points:

Page 6, Lines 5-6: Arguably, EMOS is the most general of the three methods. VINF is optimal in mean square error, making it equivalent to EMOS with c=0 when EMOS is optimized on the log score rather than CRPS. Similarly, HGR is equivalent to EMOS with d=0, on the log score.

Yes, that's a good point that EMOS is the most general. We thank the reviewer for making this point – and for suggesting the other comparisons between the methods. These details will be added to this section of the revised manuscript.

Page 6, Lines 7-10: Was a block sampling strategy used to account for trends and periodic features such as ENSO? If not this would represent a great deal of work to repeat, so I do not insist it is done, but more details would be helpful.

Thanks for the suggestion, yes more details would be helpful. The bootstrap resampling was to account for uncertainty in the fit parameters of the calibrations, not to specifically account for periodic features such as ENSO, however, it's likely that the resampling method does implicitly account for some. More details of the method will be added to the revised manuscript.

Page 6, Line 30: computed -> compute

Yes, this will be corrected.

Page 7, Line 8: the raw ensemble is clearly has -> the raw ensemble clearly has

Yes, this will be corrected.

Page 7, Lines 7-9: In apparent contradiction to the text, there is no visible positive bias in the upper panel of Figure 2, and the reference never lies outside of the ensemble.

Agreed - this is a mistake and will be corrected (this comment was in reference to a previous version of this figure that has since been replaced but we should have caught this).

Page 7, Lines 27-29: It would be useful to have some of these results available in the supplementary material. It seems likely that there will be systematic differences depending on the calibration period, given the relative lack of signal in most models until around 1990, the inability of most CMIP5 models to reproduce the so-called hiatus period, and the fact that the forcing after 2005 will differ from the observations. Longer calibration periods will down-weight the information contained in these key periods.

Good point - we did do some sensitivity tests and will include some examples of these in the supplementary material when we revise the manuscript.

Page 11, Lines 15-17: Given my primary concern above, and my comment on Page 7, it would also be useful to have some of these results available in supplementary material, and a little more discussion given.

Again, this is a fair point and something we will address. The verification statistics over the different period are likely important and we will provide more of these in the supplementary material of the revised manuscript, along with some discussion of these results in the main text.

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