

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 specific 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 "significantly". 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 "precipitation".

- 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.

Reply to RC2 (comments in blue, reply in black)

General comments

This is an ambitious and novel study aimed at improving climate projections using calibration techniques developed for initialized seasonal prediction. The approaches are tested on two single-model Large Ensembles (LE) using out-of-sample verification methods based on CMIP5 models. The analysis focuses on temperature and precipitation over Europe and takes into account seasonality. Another novel aspect is the application of the calibration method on the dynamical and residual thermodynamic components separately using the technique of “dynamical adjustment”. This yields an improvement in the accuracy of projections of temperature but not precipitation. The study is comprehensive and the methods are scientifically sound. The paper is generally well written, although some clarification is needed in places. I have a number of comments and suggestions as detailed below, but they are mostly minor in scope.

We thank the reviewer for their positive comments and feedback. We agree that there are some aspects of the paper that require clarification and would benefit from further discussion in the revised manuscript – further details follow the specific points below.

Specific comments

1) P2 L24: remove “was applied to”

Agreed, will change in the revised manuscript.

2) P2 L32: Perhaps reference Deser et al. (2020) which provides a broader view of the utility of Large Ensembles with multiple models, and includes a more comprehensive listing of LE experiments to date. Deser, C., F. Lehner, K. B. Rodgers, T. Ault, T. L. Delworth, P. N. DiNezio, A. Fiore, C. Frankignoul, J. C. Fyfe, D. E. Horton, J. E. Kay, R. Knutti, N. S. Lovenduski, J. Marotzke, K. A. McKinnon, S. Minobe, J. Randerson, J. A. Screen, I. R. Simpson and M. Ting, 2020: Insights from earth system model initial-condition large ensembles and future prospects. Nat. Clim. Change, doi: 10.1038/s41558-020-0731-2.

Yes, this is an important reference. I think (or hope) that this was published after submission but is clearly a very relevant and useful reference and will be included in the revised manuscript.

3) P3 L23: Suggest using “CESM1-LE” in place of “LENS” throughout for parallel construction with “MPI-GE”.

Agreed, this is neater and we will make this change in the revised manuscript.

4) P3 L33: Please do some sensitivity tests on the choice of SLP dataset. I know that HadSLP2 generally has lower amplitude variability (and maybe trends) than 20CR or ERA20C.

We will perform some sensitivity tests and include details in the revised manuscript.

5) P4 L19: *“lies” should be “lie”*

Yes, this will be corrected.

6) P4 L20: *“is further” should be “are further”*

Yes, this will be corrected.

7) P5 L17: *“correlation” is mis-spelled and there is some missing text after “ensemble and,*
“

Yes, the spelling and text in this passage will be corrected in the revised manuscript.

8) P6 L4: *“time” should be plural*

Agreed, will change in the revised manuscript.

9) P6 L24: *Add “Guo et al., 2019” to your list of references (this was an application to precipitation) Guo, R., C. Deser, L. Terray and F. Lehner, 2019: Human influence on winter precipitation trends (1921-2015) over North America and Eurasia revealed by dynamical adjustment. Geophys. Res. Lett., 46, doi: 10.1029/2018GL081316.*

Good point - this was an oversight on our part and is a very relevant paper. A reference to this will be included in the revised manuscript.

10) P7 L8: *“is clearly has a” is not grammatical*

Agreed, will change in the revised manuscript.

11) P8 L19: *This sentence is confusing because it sounds like you are only testing the methods on the MPI-GE, but that is not the case. I suggest first discussing the LENS results and then moving on to the MPI results.*

This is a good suggestion – we will edit this passage accordingly in the revised manuscript.

12) P9 L2: *is the lack of improvement in winter because the characteristics of the variability are not distinguishable between LENS and CMIP5?*

That is the case in the MED region but more generally it might be because there is generally less forced change, so that the internal variability component is more important to calibrate. In NEUR for example, this led to a clear improvement in the reliability despite no change in the RMSE (e.g. Figure 5). The text will be edited to clarify this point in the revised manuscript.

13) P9 L 3: “are” should be “is”

Agreed, will change in the revised manuscript.

14) P9 L5: “larger than is appropriate”: please explain what you mean. Does this imply that LENS has more variability than the other CMIP5 models, or a larger forced signal? Relatedly, it would be very nice to see some discussion of the relevance of the so-called “signal-to-noise paradox” in the seasonal-to-interannual prediction literature for climate change projections.

As highlighted by the reviewer, it is not correct to say that the spread is “larger than appropriate” because it just means it is larger than the other CMIP5 models. This will be amended in the revised manuscript. There may certainly be some aspects of the “signal-to-noise paradox” which are relevant and have implications for climate projections and we will try to highlight this in the discussion in the revised manuscript.

15) P9 L10: “in to” should be “is to”

Agreed, will change in the revised manuscript.

16) P9 L18: Change “covarying signal in the reference/observational index” to “covarying signal between the reference and observational indices” for clarity (unless I misunderstand your approach).

Yes, this should certainly have been clearer. The meaning here is to highlight the covarying signal between the reference and the ensemble mean which is being calibrated. This will be clarified in the revised manuscript.

17) P9 L21: “with a circulation driven signal”: do you want to specify whether this can be an “internal” circulation driven signal, or forced, or both?

Agreed, this could be both and is important to make that clear here – will amend in the revised manuscript.

18) P10 L20: “separately” is mis-spelled

Yes, this will be corrected.

19) P10 L21: *“in the ensemble with a signal”*: please clarify your intended meaning; the language is confusing.

Agreed, will change in the revised manuscript.

20) P11 L1: *“of temperature.”*: I would add *“in both seasons and models, but especially summer”*.

Yes, that is a good suggestion, thanks. Will amend this in the revised manuscript.

21) P11 L27: *“from the all of”* ?

This is will be corrected in the revised manuscript.

22) P11 L30-31: *Can you provide a physical explanation for why the calibration method acts to increase the uncertainty in future projections? Does it have to do with differences between the level of variability between observations and the model?*

Yes, it seems to be largely due to the differences in the levels of variability. We will look into this further though and try to elaborate on this in the revised manuscript.

23) P12 L9: *Is the reduced drying mainly dynamical or thermodynamic in origin?*

It seems to be mainly dynamical. In the models there is a stronger dynamical signal over the reference period which doesn't seem to be there in the observations and this is reduced by the calibration. This detail will be added to the revised manuscript.

24) P12 L15 *“far more consistent . . .”*: I think this is an overstatement.

Agreed, the “far” is probably not justified here. This will be amended in the revised manuscript.

25) P12 L22-30: *How do your results relate, if at all, to the trend biases in LENS compared to a synthetic observational Large Ensemble documented in McKinnon and Deser (2018)? McKinnon, K. A and C. Deser, 2018: Internal variability and regional climate trends in an Observational Large Ensemble. J. Climate, 31, 6783–6802, doi: 10.1175/JCLI-D-17-0901.1.*

Very interesting question. Thinking about it, in some sense the calibration is “trying” to account of some of these biases but how is related to the trends is not obvious. Nonetheless, this is an important study and discussion of this will be added to the revised manuscript.

26) P13 L31: suggest adding “in the calibrated ensembles” after “generally smaller”

Agreed, will change in the revised manuscript.

27) P14 L7: “For precipitation, where there is no clear signal over the reference period in the observations”: I am not sure what your evidence is. Guo et al. (2019) found a nice correspondence with dynamically-adjusted precipitation trends from observations and the ensemble-means of LENS and CMIP5 models.

Here the statement is just for the seasons and regions specifically analysed in the paper and when comparing the signal to interannual timescale variability (e.g. black lines in Figure 8). It will be clarified that this is not a general statement in the revised manuscript.

28) P14 L12: add “relative to the internal variability” after “weaker” (i.e., the forced signal doesn’t weaken on smaller scales, just the signal-to-noise weakens).

Agreed, this will be added in the revised manuscript.

29) P14 L21: “is kept” should be “are kept”

This is will be corrected in the revised manuscript.

30) P14 L27: Cite Yeager et al. (2018) for the LENS DPLE. Yeager, S. G., G. Danabasoglu, N. Rosenbloom, W. Strand, S. Bates, G. Meehl, A. Karspeck, K. Lindsay, M. C. Long, H. Teng, and N. S. Lovenduski, 2018: Predicting near-term changes in the Earth System: A large ensemble of initialized decadal prediction simulations using the Community Earth System Model, *Bull. Amer. Meteor. Soc.*, in press, doi: 10.1175/BAMS-D-17-0098.1.

Good point – this reference will be included in the revised manuscript.

31) P14 L28: “merged calibrated climate predictions”: insert “set of” before “climate predictions”?

Agreed, this will be added in the revised manuscript.

32) Caption to Fig. 3: add “summer” before “temperature”

Agreed, will change in the revised manuscript.

33) Title to Fig. 4: It is confusing. Suggest re-wording as: “LENS JJA Temperature” (analogous comment applies to Fig. S1).

Good suggestion – this will be changed here and in Fig S1 in the revised manuscript.

34) Title to Fig. 5: omit the dash after “LENS” for clarity

Agreed, will change in the revised manuscript.

35) Caption to Fig. 5: 2nd sentence: change “Shown” to “Results are shown . . .” . Also, the sentence describing what the black boxes mean is confusing. I would shorten to: “Black boxes indicate where the HGR-decomp method of calibration is significantly better than the HGR method (at the 90% level).”

Thanks for this suggestion. We will change accordingly in the revised manuscript.

36) Caption to Fig. 7, line 3: change “has a” to “is”. In the next line, change “worse that” to “worse than”.

Agreed, will change in the revised manuscript.

37) Caption to Fig. 8: Please state what the various colors and linestyles mean, and what the shading means. Don’t rely on the legend. Indeed, the colors/linestyles in the legend seems to be at odds with that shown in Fig. 7, which had all blue for LENS and all red for MPI. Please make them consistent for clarity.

Good point – the inconsistency in colours is quite stupid really and we will change this in the revised manuscript.

38) Fig. 9: Same comment as above: please use a consistent color scheme as in Fig. 7 (or change Fig. 7 to be consistent with Fig. 9).

As above - we will change this in the revised manuscript.

39) Caption to Fig. 9: Please state the method of calibration in the caption. Is it HGR-decomp?

Yes, it is HGR-decomp. This will be clarified in the revised manuscript.

We thank the reviewer for their incredibly helpful review. There were lots of insightful and helpful comments and we are confident these will help to improve the paper.

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 than 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. VIN^F 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.

Reply to RC4 (comments in blue, reply in black)

Summary

This paper presents a novel study which attempts to create better projections by calibrating large ensembles over a calibration period where we have both observations and large ensemble simulations. This study investigates three methods of calibration and finds that while all methods perform well, no method performs substantially better than the others. They then show improvement by using a dynamical decomposition method. They find that the calibration works much better for temperature than precipitation, and attribute this to the lack of clear forced change in the calibration period for precipitation. For temperature they find improvement for both large ensembles over Europe by using this calibration method and find that it reduces warming as compared to the calibrated ensemble. I recommend publication with a few minor points to be addressed.

We thank the reviewer for the positive comments on our study.

Minor points:

Page 3 line 2 should be 'ensembles'

This will be corrected.

Page 3 lines 27/28 MPI-GE is initialized from different years of a long pre-industrial control run, not in the same way as LENS

This is an important distinction and was an oversight on our part. A description to this effect will be added in the revised manuscript.

Page 4 line 22 should be 'projections'

This will be corrected.

Section 2.3.1 Are you results sensitive to the choice of reference period? For the dynamical decomposition can you explain why and how you use SLP?

No the results are not very sensitive to the reference period. We tested from 30-years up to the full 97-year periods and the verification statistics generally improve with the length of the period, which is why we use the full reference period here. Text describing these tests will be added to the revised manuscript.

The SLP is used to estimate what seasonally anomalies can be attribute to large-scale circulation anomalies (assessed in terms of SLP anomalies). This will be clarified in the revised manuscript, including a schematic illustrating how the dynamical decomposition is applied to produce the calibrated projections.

Page 7 lines 7/8. Please explain what you mean by "The raw ensemble is clearly has a positive bias"

This refers to the observations over the reference period – this will be clarified.

Section 3.3 The explanation at the beginning of the section should be in Section 2.4

Agreed, the description in section 2.4 will be expanded in the revised manuscript and will also include a schematic to visualise how this is used to produce calibrated projections.

Additional studies that may be of interested: only cite if you feel appropriate.

<https://www.earth-syst-dynam-discuss.net/esd-2019-69/>

<https://journals.ametsoc.org/doi/full/10.1175/JCLI-D-16-0905.1>

*Deser, C., F. Lehner, K. B. Rodgers, T. Ault, T. L. Delworth, P. N. DiNezio, A. Fiore, C. Frankignoul, J. C. Fyfe, D. E. Horton, J. E. Kay, R. Knutti, N. S. Lovenduski, J. Marotzke, K. A. McKinnon, S. Minobe, J. Randerson, J. A. Screen, I. R. Simpson and M. Ting, 2020: Insights from earth system model initial-condition large ensembles and future prospects. *Nat. Clim. Change*, doi: 10.1038/s41558-020-0731-2.*

Agreed, the description in section 2.4 will be expanded in the revised manuscript.

We thank the reviewer for their very useful comments and suggestions.

Calibrating large-ensemble European climate projections using observational data

Christopher H. O'Reilly^{1,2}, Daniel J. Befort¹, and Antje Weisheimer^{2,3}

¹Atmospheric, Oceanic and Planetary Physics, Department of Physics, University of Oxford.

²NCAS-Climate, Department of Physics, University of Oxford.

³European Centre for Medium-Range Weather Forecasts (ECMWF).

Correspondence: Christopher H. O'Reilly (christopher.oreilly@physics.ox.ac.uk)

Abstract.

This study examines methods of calibrating projections of future regional climate [for the next 40-50 years](#) using large single model ensembles (the CESM Large Ensemble and MPI Grand Ensemble), applied over Europe. The three calibration methods tested here are more commonly used for initialised forecasts from weeks up to seasonal timescales. The calibration techniques are applied to ensemble climate projections, fitting seasonal ensemble data to observations over a reference period (1920-2016). The calibration methods were tested and verified using an “imperfect model” approach using the historical/RCP 8.5 simulations from the CMIP5 archive. All the calibration methods exhibit a similar performance, generally improving the out-of-sample projections in comparison to the uncalibrated (bias-corrected) ensemble. The calibration methods give results that are largely indistinguishable from one another, so the simplest of these methods, namely Homogeneous Gaussian Regression ([HGR](#)), is used for the subsequent analysis. ~~An extension to this method—applying it~~ [As an extension to the HGR calibration method it is applied](#) to dynamically decomposed data (, in which the underlying data is separated into dynamical and residual components) ~~—is also tested. The verification indicates that this calibration method produces~~ ([HGR-decomp](#)). [Based on the verification results obtained using the imperfect model approach, the HGR-decomp method is found to produce](#) more reliable and accurate projections than the uncalibrated ensemble for future climate over Europe. The calibrated projections for temperature demonstrate a particular improvement, whereas the projections for changes in precipitation generally remain fairly unreliable. When the two large ensembles are calibrated using observational data, the climate projections for Europe are far more consistent between the two ensembles, with both projecting a reduction in warming but a general increase in the uncertainty of the projected changes.

Copyright statement.

20 1 Introduction

To make informed assessments of climate impacts and implement relevant adaptation strategies, reliable climate projections are important for policy-makers and other stakeholders (e.g. Field et al., 2012). There is particular demand for climate projections

on regional-scales for the next 40-50 years, however, such predictions are currently very uncertain (e.g. Stocker et al., 2013; Knutti and Sedláček, 2013). One example of the demand for improved regional climate projections is the EU-funded "European Climate Prediction system" project (EUCP), which aims to produce reliable European climate projections from the present to the middle of the century (Hewitt and Lowe, 2018). In this study, which is a part of the EUCP project, we examine methods of
5 improving the accuracy and reliability of climate projections over the European region.

There are a myriad of factors that contribute to the uncertainty in projections of future regional climate. One large factor is the uncertainty in greenhouse gas emissions and associated future radiative forcing anomalies (e.g. Pachauri et al., 2014). In this study we will focus only on estimating uncertainty of the physical climate system itself in responding to changing greenhouse gas forcing by focusing on a single representative concentration pathway (RCP), following the Coupled Model Intercomparison
10 Project 5 (CMIP5) protocol (Taylor et al., 2012). The majority of analyses of coupled model projections are based upon multi-model ensembles, which combine projections from multiple different coupled ocean-atmosphere climate models. One strength of a multi-model ensemble is that if each of the models has different structural deficiencies and associated errors, then these will not overly influence the ensemble projection. In seasonal forecasting, for example, multi-model products have been found to outperform the individual models in several studies (e.g. Palmer et al., 2004; Hagedorn et al., 2005; Baker et al., 2018).
15 Multi-model ensembles of coupled climate models provide a range of plausible scenarios for the historical and future evolution of the physical climate system. The simplest treatment of these models is to assume that each is equally likely, sometimes referred to as "model democracy" (e.g. Knutti, 2010). However, this approach assumes that models are independent and that they each represent an equally plausible representation of the climate system, neither of which is typically well justified (e.g. Gleckler et al., 2008; Knutti et al., 2013).

Several methods have been developed that go beyond "model democracy" and instead weight models based on their performance in an attempt to improve the representation of uncertainty in multi-model ensembles. One step away from model democracy is to downweight models in the ensemble that are not independent from one another (e.g. Sanderson et al., 2015), as has often found to be applicable in CMIP5-based studies. An additional or alternative approach is to weight models based on their past performance with respect to an observational benchmark, which could be the climatology of one or multiple fields
25 ~~(e.g. Giorgi and Mearns, 2002, 2003; Knutti et al., 2017; Sanderson et al., 2015)~~ (e.g. Giorgi and Mearns, 2002, 2003; Knutti et al., 2017;
or the ability of models to capture past changes (e.g. Kettleborough et al., 2007). In a recent paper, Brunner et al. (2019) applied a model weighting technique ~~was applied to the climate projections from the CMIP5 models~~ to the CMIP5 climate projections over the European region. The model weighting was found to constrain the large spread in the CMIP5 models and reduce the implied uncertainty in the multi-model projections of European climate over the coming decades.

A weakness of multi-model ensembles, however, is that the different externally forced climate response in each of the models can be difficult to isolate from internal variability. This is particularly problematic when each model typically only consists of a few ensemble members or less, as is the case with most models in CMIP5. To overcome the problem of disentangling the forced model response from the internal variability, several modelling groups have performed large single model ensemble simulations, using 40 ensemble members or more ~~(e.g. Kay et al., 2015; Maher et al., 2019)~~. ~~With~~ (e.g. Deser et al., 2020).
35 When dealing with such large ensemble sizes, the ensemble mean provides a good estimate of the externally forced signal and

deviations from this can reasonably be interpreted as the internal variability of the coupled climate system. A further strength of large ensembles is that they can be used to effectively attribute climate variability to changes in large-scale circulation. For example Deser et al. (2016) used a large ensemble to demonstrate that the observed wintertime temperature trends over the second half of the century were due to a combination of forced thermodynamic changes and a dynamically driven temperature trend that was not clearly externally forced. We will use the separation of the large ensembles into forced signals and internal variability, as well as the separation of each into dynamical and thermodynamical components, to examine different methods of calibrating projections of European climate.

Despite large ensembles providing clearer estimates of forced climate signal and internal variability, it is obvious that ~~that these ensemble~~ these ensembles will not perfectly represent observed climate variability, which is also the case with the multi-model ensembles. In this study, we explore the extent to which large ensemble climate projections can be calibrated over the observational period to adjust and potentially improve future projections. The general calibration approach relies upon the large ensemble being clearly separable into a forced signal component and residual internal variability (e.g. Deser et al., 2014). Calibration techniques have previously been applied to output from initialised seasonal forecasts (as well as shorter range forecasts), and have been demonstrated to reduce the forecast error and, perhaps more crucially, improve the reliability of the probabilistic forecasts (Kharin and Zwiers, 2003; Doblas-Reyes et al., 2005; Manzanas et al., 2019). In addition to seasonal timescales, calibration techniques have also been shown to be effective on the output from decadal prediction systems (Sansom et al., 2016; Pasternack et al., 2018). However, these types of ensemble calibration techniques have not previously been applied to ensemble climate model projections. Here we apply ensemble calibration techniques to uninitialised large ensemble climate projections, focusing on European regions, to test whether these ensembles can be calibrated to give reliable probabilistic climate projections for the next 40-50 years. ~~Given that these calibration methods have been shown to be effective when applied to initialised decadal forecasts, if calibration also proves effective for projections beyond 10 years this would present an opportunity to merge the calibrated decadal predictions with calibrated large ensemble climate projections.~~

The paper is organised as follows. The datasets, verification techniques and calibration methods are described in the next section. In section 3, we present results from the different calibration methods, namely, "Variance Inflation", "Ensemble Model Output Statistics", and "Homogeneous Gaussian Regression". These calibration methods ~~applied to~~ are applied to, and verified against ~~CMIP5~~ CMIP5 model data and also applied to observations. Conclusions follow in section 4.

2 Datasets and methods

2.1 Model and observational datasets

In this study we use two different large ensemble coupled climate model datasets. The first is from the CESM(1) Large Ensemble (Kay et al., 2015), hereafter referred to as "LENSCESM1-LE", which consists of 40 members initialised from with random round-off error from a single ensemble member in 1920 and freely evolving thereafter. Each ensemble member is performed with identical external forcing, following the CMIP5 protocol for the 1920-2005 historical period and the representative concentration pathway 8.5 (RCP 8.5) over the period 2006-2100. The second large ensemble dataset is the MPI Grand Ensemble

(Maher et al., 2019), hereafter referred to as "MPI-GE", which is similar to LENS-CESM1-LE but uses the MPI Earth System Model and consists of 100 members starting in 1850-1850, each initialised from a different initial conditions taken from a long pre-industrial control simulation. MPI-GE is integrated through to 2099 using various CMIP5 forcing scenarios but here we use the RCP 8.5 data to compare with LENS-CESM1-LE. We only use 99 members of the MPI-GE that had all of the variables used here available at the time of carrying out the analysis. For both datasets we use data over the period 1920-2060, which is covered by both large ensemble datasets. The near-term (\approx 1-40 years) period is the primary period of interest of the EUCP project.

Observational data for surface air temperature and precipitation is taken from the CRU-TS v4.01 gridded surface dataset (Harris et al., 2014). The observational sea-level pressure (SLP) data is taken from the HadSLP2 dataset (Allan and Ansell, 2006) for the results presented below, however, we tested the sensitivity to the choice of observational SLP dataset by using the 20th Century Reanalysis v3 (20CR; ?). The results were generally very similar regardless of the observational datasets, however, some of the differences between the observational dataset are highlighted in section 3.5.

The data used for out-of-sample verification was taken from the CMIP5 archive (Taylor et al., 2012). We take the first ensemble member for the 39 models that cover the 1920-2060 period for the historical (up to 2005) and RCP 8.5 (from 2006) scenarios. The CESM1-LE has a $1^\circ \times 1^\circ$ degree horizontal resolution in the atmosphere (with 30 vertical levels), which is generally comparable or higher resolution than the models in the CMIP5 ensemble. The MPI-GE has a comparatively low T63 spectral resolution (equivalent to around a 2° degree horizontal resolution), with 40 levels in the vertical. Data from the CMIP5 models, MPI-GE ensemble and the observational datasets were regridded to the same grid as the LENS-CESM1-LE dataset prior to the analysis. Tests on a small subset of the results showed that the results were not sensitive to the regridding procedure.

We analyse the evolution and projections of surface-air temperature (referred to as temperature hereafter) and precipitation over the three European SREX regions (Field et al., 2012). These are the Northern Europe, Central Europe and Mediterranean regions, which are shown in Figure 1 and will be hereafter referred to as NEUR, CEUR and MED, respectively. Our analysis focuses on projections of seasonal mean climate for European summer (defined as the June-July-August average) and winter (defined as the December-January-February average).

2.2 Verification metrics

The impact of the calibration is assessed through a series of verification metrics. The root-mean square error (RMS error) is a simple measure of the accuracy of the ensemble mean prediction. In addition, the spread of the ensemble is also calculated, which is defined as the square root of the mean ensemble variance over the verification period (e.g. Fortin et al., 2014). By calculating the RMS error and spread we are able to estimate the reliability of the ensemble by calculating the spread/error ratio, which for a perfectly reliable ensemble will be equal to one (e.g. Jolliffe and Stephenson, 2012). A spread/error ratio greater than one indicates an underconfident ensemble, whereas a spread/error ratio less than one indicates an overconfident ensemble. The final metric that we will consider is the continuous rank probability score (CRPS), which is a probabilistic measure of forecast accuracy that is based on the cumulative probability distribution (e.g. Hersbach, 2000; Wilks, 2011; Bröcker, 2012).

The CRPS measures where the verification data point lies with respect to the underlying ensemble and is higher when the verification data is-are further from the centre of the ensemble. As such, a lower CRPS value represents a more skillful probabilistic forecast.

2.3 Ensemble calibration methods

- 5 We will assess the effectiveness of calibrating ensemble climate projection-projections using a series of different calibration techniques, which are outlined in this section. The calibrations are performed seperately-separately for each region and season, on annually-resolved indices.

2.3.1 Uncalibrated ensemble

- 10 The benchmark for the calibration methods is the uncalibrated ensemble. Here we use the term uncalibrated ensemble to refer to an ensemble that has been bias corrected by removing the mean value over a particular reference period. Of course, this is not strictly an uncalibrated ensemble but this is the most common way that climate projections are presented in the literature (e.g. Hawkins and Sutton, 2016). In the analysis that follows the reference period is always the same as for the corresponding calibration methods, which is generally the observational period 1920-2016 in the following analysis.

2.3.2 Variance inflation (VINI)

- 15 One calibration method that we will test is "Variance inflation", hereafter referred to as VINI, following Doblas-Reyes et al. (2005). For each uncalibrated ensemble member, X_{uncalib} , VINI adjusts the ensemble mean signal, X_m , and anomaly with respect to the ensemble mean, $X_{\text{ens-anom}}$, from the uncalibrated ensemble. The uncalibrated ensemble can be expressed in these terms as

$$X_{\text{uncalib}}(t, e) = X_m(t) + X_{\text{ens-anom}}(t, e). \quad (1)$$

- 20 Here t and e indicate dependence on time and ensemble member, respectively. The VINI method produces a calibrated ensemble, X_{calib} , through the following scaling

$$X_{\text{calib}}(t, e) = \alpha X_m(t) + \beta X_{\text{ens-anom}}(t, e). \quad (2)$$

The scaling variables α and β are calculated as

$$\alpha = \rho \frac{s_r}{s_m}; \quad (3)$$

25

$$\beta = \sqrt{1 - \rho^2} \frac{s_r}{s_{\text{uncalib}}}; \quad (4)$$

where s_r is the standard deviation of the reference (or observational) data that is being calibrated towards, s_m is the standard deviation of the ensemble mean, s_{uncalib} is the square-root of the mean variance of the uncalibrated ensemble members, and ρ

is the correlation between the ensemble mean signal and the reference dataset over the calibration period. Where ρ is less than zero and there is no skillful ~~corrlation~~-correlation between the ensemble and the reference dataset, we set ρ to be zero in the calibration. VINF scales the signal and ensemble spread but maintains the underlying correlation and ensemble distribution, rather than fitting a parametric distribution as in the following methods.

5 2.3.3 Ensemble Model Output Statistics (EMOS)

The next calibration method is the "Ensemble Model Output Statistics" approach, hereafter referred to as EMOS (Gneiting et al., 2005). The EMOS method has widely been applied to the output of ensemble prediction systems for medium-range and seasonal forecasts. EMOS involves fitting a parametric distribution to the underlying data, such that the uncalibrated is expressed as

$$10 \quad X_{\text{uncalib}}(t) = X_m(t) + \epsilon_{\text{uncalib}}(t), \quad \epsilon_{\text{uncalib}}(t) = N[0, s^2(t)]; \quad (5)$$

and the calibrated ensemble is expressed as

$$X_{\text{calib}}(t) = bX_m(t) + \epsilon_{\text{calib}}(t), \quad \epsilon_{\text{calib}}(t) = N[0, c + ds^2(t)]; \quad (6)$$

where $N[\mu, \sigma^2]$ is a Gaussian distribution with mean μ and variance σ^2 and s^2 is the time-dependent variance across the ensemble. The coefficients b, c , and d are found using numerical methods to minimise the CRPS over the calibration period

15 Gneiting et al. (2007). The coefficients b, c , and d are constrained to be non-negative values. The EMOS technique is arguably the most general method we will test because it allows for meaningful differences in spread across the ensemble at different ~~time-times~~ (i.e. the coefficient d), so is sometimes referred to as "Nonhomogenous Gaussian Regression" (e.g. Wilks, 2006; Tippett and Barnston, 2008). EMOS represents a simplification over the VINF method because the the ensemble distribution is parameterised as Gaussian. In this study, the EMOS technique is used to produce 1000 sampled ensemble members in the
 20 ensemble projection. To avoid overfitting to the observations when producing the ensembles and to include some measure of sampling uncertainty in the parameter fitting process, the EMOS method is applied to randomly resampled years (with replacement) from the calibration period, to produce 1000 valid combinations of the coefficients b, c , and d . These combinations are used to produce the 1000 sampled ensemble members used to produce the calibrated projection.

2.3.4 Homogenous Gaussian Regression (HGR)

25 The third calibration method that we test is "Homogenous Gaussian Regression", hereafter referred to as HGR. The HGR method is a simplified version of EMOS, in which the calibrated variance is constant ~~is-in~~ time and is expressed as

$$X_{\text{calib}}(t) = bX_m(t) + \epsilon_{\text{calib}}(t), \quad \epsilon_{\text{calib}}(t) = N[0, c]. \quad (7)$$

Effectively, this method assumes that there is no information in the time variation of the ensemble spread. The coefficients b and c are found as in EMOS and are constrained to be greater than or equal to zero.

2.4 Dynamical decomposition of climate anomalies

In this study we will test calibrating the full variables as well as calibrating dynamically decomposed variables. The dynamical decomposition aims to express variables - surface air temperature and precipitation in this case - as a dynamical and residual component. The rationale for testing this on the calibration methods is that they may be fitting a thermodynamic signal in the ensemble to something that is dynamically driven in the reference (or observational) data and therefore conflating different mechanisms. Dynamical decomposition has previously been used to understand observed large-scale climate variability on decadal timescales where there is a contribution from the thermodynamic climate change signal and large-scale circulation anomalies (e.g. Cattiaux et al., 2010; Wallace et al., 2012; Deser et al., 2016) (e.g. Cattiaux et al., 2010; Wallace et al., 2012; Deser et al., 2016; Guo et al., 2016). The dynamical decomposition splits the variables at each grid-point over Europe into $FULL = DYNAMICAL + RESIDUAL$.

The dynamical component was calculated for all model ensemble members, CMIP5 models and observations following the analog method of Deser et al. (2016). The method here is exactly the same as that used in O'Reilly et al. (2017), which provides full details. In this method, sea-level pressure (SLP) anomaly fields for each month are fitted using other SLP anomaly fields from the corresponding month from other years over the reference/observational period (1920-2016). This regression fit yields weights which are then used to ~~computed~~ compute the associated dynamical surface temperature or precipitation anomaly.

Each field can then be separated into a dynamical and residual component. An example of the dynamical decomposition of the CESM1-LE projection into dynamical and residual components is shown in Figure S1 (and also in the example calibration schematic in Figure 3). The regional dynamical and residual timeseries were calibrated using the above techniques towards the corresponding dynamical and residual timeseries from the target dataset (i.e. CMIP5 or observations over the period 1920-2016). The calibrated dynamical and residual timeseries are then combined to give a full calibrated ensemble projection, further detail is provided in the following section. Results from the calibrated dynamical decomposition are shown later in the paper for the HGR method and referred to as HGR-decomp.

3 Results

3.1 An example ensemble calibration

Before we begin our analysis, it is useful to motivate our approach by briefly describing an example calibration. A synthetic, randomly-generated 100 member ensemble is shown in Figure 2, alongside a synthetic observational index. ~~The raw ensemble is clearly has a positive bias and as a result has a large error. Despite the large spread of~~ There is a large spread across the ensemble, with the reference frequently ~~lies outside the ensemble lying close to the ensemble mean.~~

The lower panel of Figure 2 shows the ensemble calibrated towards the reference data using the VINF method. ~~The first step is a simple bias correction towards the reference mean, then the VINF~~ VINF method scales the ensemble mean and spread to make the ensemble reliable in a probabilistic sense. The improvement of the calibrated ensemble is clear from the reduction in error and CRPS, which is also shown in Figure 2. Also, it is important to note that calibrated ensemble is perfectly reliable over the reference period,

as indicated by the spread/error being equal to 1 after calibration. The EMOS and HGR methods would have yielded almost identical results for this synthetic ensemble.

It is clear from the example shown in Figure 2 that it is trivial to calibrate an ensemble to known data such that it is perfectly reliable. Of more interest here is whether calibrating to observed data can improve the accuracy and reliability of a prediction
5 *outside* of the reference period used for the calibration.

3.2 Comparing calibration techniques using an "imperfect model" test

Our aim in this study is to test how calibrating large ensemble projections using observations will influence the accuracy and reliability of the projections. The common period of the large ensembles and observations used in this study is 1920-2016, so we can in principle calibrate the ensembles using this period. However, we cannot test how effective this calibration is in the
10 future, out-of-sample period. To examine the performance of the calibration we employ an "imperfect model" test, using 39 CMIP5 models. In this test, the large ensemble dataset is calibrated to each of these 39 models over the observational-reference period, 1920-2016. The future period from the CMIP5 realisation is then used to analyse the impact of the calibration by verifying against calibrated large-ensemble in the out-of-sample period 2017-2060. We refer to this as an imperfect model test because, in this approach, the large ensemble calibration is tested mostly on simulations from different climate models. This is
15 a strength of the imperfect model test, as the observations can, in some sense, be considered an out of sample test. In addition to the 1920-2016 calibration period, we also tested the calibration over shorter periods (~~not shown some examples are shown in~~ Figure S8 of the Supplementary Material). Overall, the calibration ~~periods tended was found~~ to perform better over the longer periods, so in this study we focus on the results of the calibration on the longest available common period (i.e. 1920-2016).

An example of calibrating a large ensemble projection to a CMIP5 model index is shown in the left hand column in Figure 3.
20 In this example, the ~~raw LENS data for NEUR-uncalibrated CESM1-LE data for CEUR~~ summer temperature is shown in red, along with the same index from one of the CMIP5 models over the reference period (1920-2016). The model is calibrated ~~over~~ towards data from the CMIP5 model realisation over the reference period. Following the calibration step, the CMIP5 data from
the future period (2017-2060), which was withheld prior to the calibration, is used to verify the uncalibrated and calibrated
large-ensemble projections using each individual year in the ~~97-year reference period(1920-2016)- shown by the black vertical~~
25 ~~line in Figure 3 – and the calibrated ensemble is shown in blue. Over the 2017-2060 period, the calibrated large ensemble is~~
~~then verified against the out-of-sample verification period. The verification is performed on 44 pairs of probabilistic predictions~~
~~and validation data points from this future period (2017-2060). From each of the~~ CMIP5 model index. For both the LENS and
MPI-GE large ensemble datasets, this analysis was performed for all models, we can therefore calculate verification statistics
(i.e. RMSE, Spread/Error, CRPS). The process is then repeated for each of the 39 CMIP5 models. This process was then
30 ~~repeated for models and the distribution of these verification statistics is presented in the results that follow. We performed this~~
~~analysis for each of the calibration methods using both the CESM1-LE and MPI-GE datasets. The analysis for both temperature~~
~~and precipitation, for summer and winter, and over all three European regions~~ is presented and discussed below.

The verification statistics for the LENS-CESM1-LE summer temperature for the uncalibrated ensemble and the three calibration methods are shown in Figure 4. The distribution of the verification statistics over the 39 models is shown, with the

horizontal lines indicating the median of the distribution. The black crosses indicate where the verification of the calibrated ensemble is significantly better than the verification of the uncalibrated ensemble at the 90% confidence level, calculated using the non-parametric Mann-Whitney U-test (e.g. Wilks, 2011). For the summer temperature over all three regions, all of the calibration methods significantly lower the RMS error of the ensemble projection compared with the uncalibrated ensemble.

5 The calibration methods generally perform similarly, acting to typically reduce the spread of the uncalibrated ensemble and narrowing the range of the spread/error ratios in the verification compared to the uncalibrated ensemble. There is significant improvement in reliability, indicated by the spread/error relationship, for the CEUR region with all three calibration methods. The CRPS is significantly lower for all of the calibration methods in all regions, demonstrating that the calibrations are improving the probabilistic predictions of summer temperature by the LENS-CESM1-LE ensemble in the out-of-sample future

10 period. An important point to note is that, despite not being a significant improvement for all the verification metrics shown in Figure 4, none of the calibration methods ever has a significantly significantly negative impact on the projections.

We also ~~tested the different calibration methods on the~~ performed the same testing described above for the CESM1-LE on the MPI-GE, ~~for precipitation and for the different seasons~~. The verification measures for the MPI-GE summer temperature are shown in Figure S1S2. The performance of the calibration methods on the MPI-GE summer temperature is qualitatively

15 similar to that for the LENS-CESM1-LE (shown in Figure 4). The calibration methods in general improve the out-of-sample verification statistics, resulting in a more accurate and reliable projection over the three European regions compared to the uncalibrated ensemble. Again, there is a particularly notable improvement for the CEUR region, as with the LENS-CESM1-LE data (i.e. Figure 4). For the other regions there is an improvement over the uncalibrated ensemble but this is not significant for any of the calibration methods, or for any of the verification measures. Nonetheless, as in the LENS-CESM1-LE data, none of

20 the calibration methods displays a significantly significantly negative impact on the projections.

The comparison of the ensemble calibration methods for the summer temperature suggests that there is no significant difference between the performance of the different methods, for both of the large ensembles (i.e. Figure 4 and S1S2). Analysis of the equivalent figures for precipitation ~~and the winter season also show~~ (Figures S3 S4), as well as for the the winter season (not shown), also demonstrate a reasonably consistent performance between the calibration methods (not shown). The similarity of

25 the performance of the calibration methods indicate that the extra information included in the VINF and EMOS calibrations, compared with the HGR calibration, is not important to the performance. Therefore, we will focus on the simplest method of the three, HGR, for the analysis that follows.

The out-of-sample verification results for the HGR method for temperature and precipitation for both summer and winter seasons for the calibrated CESM1-LE projections are shown in Figure 5 ~~for the calibrated LENS projections~~ (note that the red and orange data in the first column are the same as those shown in Figure 4). The equivalent verification plot for the MPI-GE dataset is shown in Figure S2, and the results are, generally, qualitatively similar to those shown for the LENS-CESM1-LE dataset in Figure 5. The improvement for the winter temperature in terms of RMS error is not as clear as in the summer season but the reliability of the ensemble projections are improved significantly over the NEUR region. There is also some improvement for the precipitation projections in some regions, particularly in terms of the spread/error ratios of the regional

35 precipitation projections. The spread of the uncalibrated LENS-CESM1-LE data seems to be larger than is appropriate for the

targeted indices, particularly for precipitation, which is evident in the general reduction in spread in the calibrated ensemble. The spread/error ratios of the calibrated ensembles are consistently close to one, this is a particularly notable improvement for the uncalibrated ensembles over the NEUR region, which are generally underconfident prior to calibration. For some other regions, there is a smaller improvement or no noticeable difference. Crucially, the influence of the calibration on the spread/error is not significantly negative for any of the the variables regions or seasons, indicating that the ~~overall impact of the calibration in to generally improve~~ calibration generally improves the reliability of the projections. The only verification statistic where the calibrated ensemble performs significantly worse than the uncalibrated ensemble is the RMS error for the NEUR winter precipitation in the LENS-CESM1-LE dataset (Figure 5). Whilst it is only one of the verification measures performed across both the LENS-CESM1-LE and MPI-GE datasets, it is a concern because it reduces how much confidence we can have in applying the calibration using observations.

3.3 Examining calibration using dynamically decomposed variables

One potential problem with the calibration methods examined in the previous section is that they are calibrated towards a single (observational) index. The implicit assumption with this calibration approach is that the forced signal in the large ensembles is scaled based on the co-varying signal in the reference/observational index. However, we might expect the forced climate change signal to be largely thermodynamic in nature rather than being driven by changes in large-scale circulation, particularly for temperature. It is possible therefore that when fitting the calibration of the ensemble to the reference, there is an incorrect conflation of, for example, the forced thermodynamic response with ~~a~~ an circulation driven signal associated with internal variability in the reference index. To account for this potential shortcoming in the calibration method we used a dynamical decomposition method (as outlined above in section 2.4) to split the model and observations datasets into a *DYNAMICAL* component, associated with large-scale circulation anomalies, and a *RESIDUAL* component, which can often be interpreted as a thermodynamic component.

An example of the dynamical decomposition, applied to summertime projections for the CEUR region in the LENS-CESM1-LE dataset, is shown in Figure 6-3 (and also Figure S1). In this example, the future temperature response is largely associated with the residual, representing the local thermodynamic response to increase greenhouse gas concentrations. There is also some dynamical contribution to the signal but this also contributes to the uncertainty in the overall ensemble projection. In contrast, there is a much weaker signal in future precipitation changes, and the modest drying signal that is projected seems to be mostly due to dynamical changes.

We will now examine how calibrating the dynamically decomposed parts of the ensemble projection (e.g. the *DYNAMICAL* and *RESIDUAL* components ~~in Figure 6~~) separately, against the respective decomposed parts of the reference indices, before recombining affects the ensemble calibration performance. A demonstration of this process applied to one of the CESM1-LE projection is shown in Figure 3. The large ensemble projection and reference dataset are both separated into dynamical and residual components. These are then calibrated separately, which in this particular example reduces the dynamical signal substantially. Next, the calibrated decomposed projections are recombined to produce the total calibrated projection. This total calibrated projection is then used to calculate verification statistics, in the same way as for the full calibration techniques

examined previously. We use the HGR method to ~~do the separate decompositions~~ perform the calibration on the dynamically decomposed data, and refer to this ~~methods~~ method as "HGR-decomp" hereafter.

Verification results for the HGR-decomp methods are shown for both temperature and precipitation and for all regions in Figure 5, alongside the HGR verification results (with the equivalent verification for the MPI-GE shown in Figure S2S5). As with the HGR verification, the crosses/circles indicate where the verification statistics of the HGR-decomp calibrated ensemble ~~is~~ are significantly better/worse than the uncalibrated ensemble. The HGR-decomp calibration generally performs better than the uncalibrated ensemble, and for none of the verification measures does the HGR-decomp calibration perform significantly worse than the uncalibrated ensemble. This is in contrast with the HGR calibration method, for which there is a significant increase in the RMS error for the wintertime precipitation in the NEUR region.

To formally compare the HGR-decomp and HGR, we assessed the significance of the difference in the verification measures of the two methods. In Figure 5, the black boxes indicate where either of the calibration methods is found to be significantly better than the other, at the 90% level (based on a Mann-Whitney U-test). The only statistically significant differences are seen for the spread/error verification, where four of the regions/variables are significantly better for the HGR-decomp method applied to the LENS-CESM1-LE dataset. In contrast, none of the verification measures for any of the regions/variables are significantly worse for the HGR-decomp method. The HGR-decomp method also performs better for the calibrated MPI-GE indices (Figure S2S5), albeit with a lower level of significance. Specifically, in ten of the twelve total regions/variables verified for the MPI-GE dataset, HGR-decomp calibration is found to be more reliable in terms of spread/error than the HGR calibration.

Overall, the HGR-decomp method is found to be an improvement over the HGR method, and very clearly outperforms the uncalibrated ensembles. The improvement of the HGR-decomp method over the HGR method is clearest in the reliability of the projection, as measured in terms of spread/error. The spread/error is consistently higher in the HGR-decomp calibration, primarily due to the spread, which is consistently larger in the HGR-decomp calibrated ensemble. Calibrating on the dynamical and residual components ~~seperately~~ separately has the effect of increasing the overall spread, likely because the method avoids fitting a forced thermodynamical or dynamical signal in the ensemble ~~with a signal towards a forced~~ or internal variability of a different origin in the reference index. Examining the verification of the HGR calibrated *DYNAMICAL* and *RESIDUAL* components separately reveals that the spread/error of the *DYNAMICAL* components of the ensemble are particularly well calibrated (not shown). In comparison with the HGR-decomp method, the HGR method generally has a lower spread, which in many cases results in projections that have a spread/error ratio lower than one and are less reliable than for the HGR-decomp method. In this sense, the HGR method appears to be slightly "over-fitting" the ensemble to the reference period, resulting in a ~~consistent~~ consistently over-confident ensemble projection.

3.4 Examining the impact of calibration on projections of future climatologies

To assess how the calibration influences the projections of average European climate during the mid-21st century period, we will examine projections of the mean 2041-2060 climate. Until this point we have focused on verifying the yearly projections of each season over the out-of-sample period 2017-2060, which gives a verification measure for each CMIP5 model (e.g. as shown in Figures 4 & 5). We also need to verify the out-of-sample projections for the 2041-2060 means. However, since

there is only a single verification point for the climatology in each of the CMIP5 models, we instead need to combine the single measurements to produce one verification score across all the models. To estimate the uncertainty of these verification measures, we ~~performed~~ perform a bootstrap resampling over the 39 CMIP5 model projection/verification pairs. Verification results for the 2041-2060 climatologies for both the ~~LENS-CESM1-LE~~ and MPI-GE are shown in Figure ~~7-6~~.

5 The HGR-decomp calibration tends to improve the projected 2041-2060 climatology of temperature in both seasons and ensembles, but especially summer. This is a particular improvement during the summer, in both the accuracy (i.e. RMS error) and reliability (i.e. spread/error) of the out-of-sample verification. The calibrated summer temperature projections are more reliable in all three European regions in the ~~LENS-CESM1-LE~~ and MPI-GE ensembles but all tend to somewhat overconfident. The winter temperature shows less obvious improvement in terms of RMS error of the calibrated projections, but the reliability
10 is significantly improved for all the regions in both ensembles, but again, the calibrated ~~LENS-CESM1-LE~~ data is slightly more reliable. There is less improvement for precipitation projections than seen for the temperature projections. For the summer precipitation, there are modest but significant improvements in some regions in terms of the RMS error but the reliability is more mixed, with the calibration actually worsening the reliability in the MED region for the ~~LENS-CESM1-LE~~ dataset. The calibration has the least influence on the 2041-2060 climatology of precipitation, acting to worsen the RMS error in some
15 instances but also to modestly improve the reliability.

Overall, the verification of the projected 2041-2060 climatologies in the imperfect model tests indicate that the HGR-decomp calibration acts to generally improve the accuracy and reliability of the projections. The calibrated temperature projections perform better than the calibrated precipitation projections. It is notable however, that the out-of-sample verification for the 2041-2060 climatologies do not generally seem to perform as well as the calibration for the yearly projections examined in the
20 previous sections. There are several possible factors contributing to this. The first is that when we examine the performance of the calibration on the yearly projections, the beginning of the 2017-2060 verification is found to be more accurate and reliable than the latter period (~~not shown~~), as the forced signal in the ensemble diverges from the observations to which it ~~it calibrated~~.
is calibrated. This is demonstrated clearly when the verification is applied to different future periods (specifically 2021-2040, 2041-2060 and 2061-2080; see Figure S6). We find that the accuracy and reliability clearly deteriorate as the target period
25 moves further into the future, indicating that the HGR-decomp calibration method is less appropriate for periods further into the future. Another reason is that much of the increased reliability in the yearly projections ~~comes stems~~ from calibrating the (unpredictable) internal variability in the ensemble to the target index, but in the 20-year climatology there is a much smaller contribution of this internal variability.

3.5 Calibrating large ensembles to observations and assessing the impact on future climate projections

30 The imperfect model tests in the previous sections demonstrate that the calibration methods generally act to improve future projections in a out-of-sample verification. In particular, the HGR-decomp method is a categorical improvement over the uncalibrated ensembles in the imperfect model analysis using the CMIP5 ensemble, as described in the previous sections. On the basis of this analysis, we will now apply the HGR-decomp calibration method to the large ensembles, ~~targeted at using~~ the observational indices of temperature and precipitation to calibrate against.

The calibrated LENS-CESM1-LE projections for the summer temperature and precipitation are shown in Figure 8-7. Based on the imperfect model tests we expect the calibrated summer LENS-CESM1-LE to represent the most accurate and reliable projection ~~from the out of~~ all of the ensemble/variable/season combinations tested. For the summer temperature projections, calibrating the LENS-CESM1-LE against the observations ~~until 2016~~ over the reference period, the rate of warming until 2060 is reduced by varying amounts. There ~~is also a noticeable change~~ are also small changes in spread, ~~particularly perhaps most notable~~ for the NEUR region ~~but also evident in the other regions, indicating that~~, with the calibration method ~~is~~ acting to increase the uncertainty in the future projections. For the precipitation, the signal in LENS-CESM1-LE projection is much weaker with respect to the inter-annual variability. In the projections shown here the calibration has a ~~fairly modest~~ notable impact on the future projections, acting to weaken the drying projected in the CEUR region and adjusting the ensemble uncertainty in all the regions.

The calibrated summer projections from the MPI-GE are fairly similar to the LENS-CESM1-LE, with the ensemble medians of the MPI-GE also plotted in Figure 8-7 for comparison (the full ensemble projection plots are shown in Figure S3S7). The warming in the NEUR region is reduced over the 2017-2060 ~~projection~~ period and the uncertainty is increased markedly. The calibration technique makes smaller adjustments to the summer temperature projections in the CEUR and MED regions, which may be because the uncalibrated ensemble already does a reasonable job of capturing the warming variability seen during the observational period. In the projections of summer ~~precipitation~~ precipitation in the MPI-GE dataset, there is a fairly strong future drying signal in both the CEUR and MED regions that is greatly reduced by the calibration. Interestingly a similar result is seen in the time-slice experiments of Matsueda et al. (2016) when calibrated using the results of seasonal hindcast experiments, which tends to reduce the drying in the MED region. In the calibration shown here, this seems to be because the MPI-GE has a drying trend over the whole observational period in these regions ~~that is not seen very~~, which is dynamical in origin and is not seen clearly in the observations. Based on the imperfect model tests, however, we have less confidence in the performance of the calibrated ensembles for precipitation.

We also tested the observational calibration using data decomposed using the 20CR SLP, rather than the HadSLP2 SLP data (Figure S9). The results are generally insensitive to the choice of SLP dataset. One exception is the MED summer temperatures, for which the calibration amplifies future warming. In this instance the DYNAMICAL component of the decomposition when using 20CR SLP accounts for substantially less of the observed variance than in HadSLP2 (Table S1) and the decompositions are also substantially different (Table S2). This indicates that for this season and region the 20CR data is not capturing what seems to be a clearer dynamical signal in the HadSLP2 dataset and, as a result, is perhaps less dependable. On the whole though, the results are largely insensitive to the choice of SLP dataset.

To consider whether the imperfect model testing is really a useful indication of the performance of the observational calibration, it is of interest to compare the fit parameters of the HGR-decomp calibration. The parameters b and c from equation 7 are plotted in Figure S10. The observed scaling parameters, b and c , generally lie within the range of values used to calibrate CMIP5 models in the previous sections. In some cases the parameters lie outside the CMIP5 model ensemble but this is not systematic, so there is no clear reason to expect the efficacy of the calibration to be very different when applied to the observational data.

~~All the~~ The projected changes in the 2041-2060 ~~climatological changes from climatologies, compared with~~ the present day 1995-2014 reference period, ~~1995-2014~~, are shown in Figure 9-8. Here, we have plotted both the uncalibrated and (HGR-decomp) calibrated climatological changes for both the LENS-CESM1-LE and MPI-GE datasets. An interesting feature of these projected changes is that for many of them, the calibrated ensembles are ~~far~~ more consistent with one another than their uncalibrated counterparts. This is perhaps most clear for the summer temperature changes in all the European regions, particularly NEUR and CEUR, in which there is a difference of over 1°K in the mean changes of the uncalibrated projections and with no overlap in the probability distributions. After the calibration is applied, the projected mean changes are ~~far~~ closer to one another, with considerable overlap in their probability distributions. The calibration acts to make the projections more consistent for most of the variables and regions, which is reassuring as this implies that the observations are having a strong impact on the initial uncalibrated ensembles that are themselves often very different.

Another feature of the calibrations influence on the future climatologies is that it fairly consistently acts to increase the uncertainty of the projections, with respect to the uncalibrated ensembles. This is most clear for the projections of future temperature over Europe, where the imperfect model tests indicate that the calibration has a large impact on the reliability of the projections (e.g. Figure 76), suggesting that the broader calibrated distribution is reasonable and is likely to be a better future projection. It is interesting to note that the calibrated LENS-CESM1-LE projection has a wider spread than in the calibrated MPI-GE projection for many of the projected temperature indices, which may be related to particular trend biases in the CESM1-LE (e.g. McKinnon and Deser, 2018). In the imperfect model tests, shown in Figure 76, the calibrated temperature projections for the LENS-CESM1-LE dataset are consistently more reliable (in terms of spread/error) than for the MPI-GE dataset. The calibrated MPI-GE projections were more underconfident in the out-of-sample verification, indicating that we should have more confidence in the broader calibrated LENS-CESM1-LE projections for future temperature changes.

In a recent paper, Brunner et al. compared several different methods of model weighting and constraining climate projections for the European summer season using multi-model ensembles over the same 2041-2060 period under the RCP 8.5 scenario. The HGR-decomp method generally predicts lower levels of warming for European summer than the CMIP5-based model weighting/constraining methods but much of the distributions of the projected changes overlap. It is notable that the HGR-decomp method can project changes that are outside of the uncalibrated distribution, which is clearly not the case for the model weighting/constraining methods (see, for example, Figure 2 of (Brunner et al.). This feature in particular sets the HGR-decomp method apart from these other techniques, whether this is for better or worse though, is not clear.

4 Conclusions

In this study we have examined methods of calibrating regional climate projections from large single model ensembles. The three calibration methods tested here - VINP, EMOS and HGR - are more commonly used for initialised forecasts from weeks up to seasonal timescales. Here we applied these calibration techniques to ensemble climate projections, fitting seasonal ensemble data to observations over a reference period (1920-2016). The calibration ~~techniques~~ methods act to scale the ensemble signal and spread so as to optimize the fit over the reference period. The three calibration methods ~~displayed~~ display similar

performance, all generally improving the out-of-sample projections in comparison to the uncalibrated ensemble. The simplest of the calibration methods, HGR, includes no variability of the ensemble spread and effectively discards any information that may be contained in the year-to-year variability of the spread in the raw ensemble. Based on the performance of the HGR method, we can conclude that the information in the year-to-year changes in the ensemble spread is not important enough, at least in the large ensembles examined in this study, to have a meaningful influence on the ensemble calibration.

We also tested calibrating the variables after they had been subjected to a dynamical decomposition. In this method, all variables were separated into *DYNAMICAL* and *RESIDUAL* components using information of the large-scale circulation, calibrated separately using the corresponding reference indices and then recombined to produce the final calibrated ensemble. The results from the out-of-sample verification of the HGR-decomp calibrations demonstrate a small but noticeable improvement over the HGR method, particularly in terms of the reliability. The HGR calibrated ensembles have a tendency to be overconfident for their future projections and this seems to be due to an apparent over-fitting to variability in the reference period, which is found to be alleviated to some extent by calibrating the *DYNAMICAL* and *RESIDUAL* components of the ensemble separately. Therefore, the HGR-decomp calibration was chosen as the best method to apply to the observational reference data.

The HGR-decomp calibration method was also found to improve the projections of 20-year climatologies during the mid-21st century (i.e. 2041-2060). The accuracy and reliability of the projections ~~improved~~improve in the calibrated ensemble, when subject to the imperfect model tests. The performance of the calibration was substantial for the temperature projections but for precipitation the improvement is much more modest, or even absent in some instances. Whilst both datasets demonstrate an improvement due to calibrations, it is interesting that the LENS-CESM1-LE dataset seems to perform better than the MPI-GE, particularly in terms of the reliability of the future projections. Perhaps it is not too surprising that one ensemble would be found to be better than another when subjected to this type of calibration. For example, if we had a third ensemble that we knew was a much worse representation of the climate system, we might expect the calibration to improve the projection in this ensemble but we would not expect this calibrated ensemble to outperform the other large ensembles. In this sense, the calibration approach taken here is clearly not a panacea for all ensemble projections and ultimately, the accuracy and reliability of the calibrated ensemble projection would expected to depend on the raw ensemble projection.

We ~~applied then proceeded to apply~~ HGR-decomp method to each of the large ensembles using the observations as a reference, over the period 1920-2016. Based on the imperfect model testing we expect that the calibrated ensemble projection provides a more accurate and certainly a more reliable probabilistic projection of European climate over the next ~~40 or so~~40-50 years. In both the LENS-CESM1-LE and MPI-GE datasets the projected increase in European temperatures ~~was generally smaller~~is generally smaller in the calibrated ensembles compared to the uncalibrated ensembles. The calibrated projections are ~~much~~notably more consistent with one another than the respective calibrated projections, indicating that the calibration with observations is having a consistent and substantial influence on the future projections of European climate. For the example of European temperatures, the best estimates for the summer temperature change for the period 2041-2060 (from 1995-2014) is projected to be about 2°C for CEUR and MED regions and 1.3°C for the NEUR region. Each of these is associated with a substantial ensemble spread, however, reflecting the increased uncertainty (or larger ensemble spread) added by the calibration to provide a more reliable projection.

The overall effectiveness of the calibration seems to stem from some key characteristics of the ensemble and reference datasets. The calibration performs well where there is a reasonably strong signal in the ensemble that is also present to some extent in the reference data, as is the case for the temperature indices. In these instances, the signal is scaled and an ensemble spread is added to represent the appropriate estimate of internal variability, much of which is associated with large-scale circulation variability. For precipitation, where there is no clear signal over the reference period in the observations [for the specific regions and seasons analysed here](#) (and in many of the CMIP5 models), any future changes projected are difficult to scale over this reference period. In effect, the calibration then adds value by correcting (mostly by inflating) the ensemble spread. This calibration method could therefore reasonably be applied to many other regions and variables where there is an emerging forced signal in response to external forcing. The calibration can also be applied to smaller spatial scales but as the scales become smaller, the forced signal generally becomes weaker [relative to the internal variability](#), so the calibration will tend to become somewhat less effective. Nonetheless, the calibration has also demonstrated some utility for temperature projections on 2.5° grid-boxes (as included in [Brunner et al., submitted to J. Climate](#) [Brunner et al.](#)).

One novel aspect of this study that is particularly worth emphasising is the imperfect model testing approach. Previous studies have typically used multi-model ensembles to constrain future projections and some in particular have used a “leave-one-out” perfect model approach to examine the effectiveness of these methods (e.g. Knutti et al., 2017; Brunner et al., 2019). However, this leave-one-out approach is often used to tune particular parameters in the methodology, such as the performance weighting parameter in Brunner et al. (2019). The use of the leave-one-out approach to tune the method is certainly well justified. However, this does reduce the power of subsequently re-using this approach to verify the accuracy of the constraining method, which may result in over-fitting or an over-estimation of the ~~added-values~~ [added-value](#) of the constraint. The imperfect model approach we have used in this study is less susceptible to this type of over-fitting as the data used for the verification ~~is-are~~ kept separate from the underlying large ensembles throughout, only being used to compare the efficacy of the different methods. [The imperfect model approach to testing is therefore an advantageous approach, regardless the particular calibration or model weighting that is being subjected to the testing.](#)

As well as being applied to other datasets and regions, this calibration method can also be applied to initialised decadal forecasts. [Decadal predictions exhibit skill in some aspects out to 10-years \(e.g. Doblus-Reyes et al., 2013; Smith et al., 2019\) and a recent study has demonstrated that constraining climate projections using initialised decadal predictions can improve the accuracy of projections in some cases \(Befort et al., 2020\), which is an exciting proposition for improving climate prediction. Given that these calibration methods have been shown to be effective when applied to initialised decadal forecasts, if calibration also proves effective for projections beyond 10 years this would present an opportunity to merge the calibrated decadal predictions with calibrated large ensemble climate projections.](#)

Previous studies have examined how similar calibration methods [to those examined in this paper](#) can improve multi-year forecasts (e.g. Sansom et al., 2016; Pasternack et al., 2018). It would be of particular interest to examine how calibrated decadal predictions could be combined or merged with these calibrated projections. The ~~LENS-CESM1-LE~~ dataset analysed in this study has an initialised counterpart, namely the Decadal Prediction Large Ensemble ([Yeager et al., 2018](#)), and testing how to

combine data from these different ensembles to produce a merged calibrated [set of](#) climate predictions would potentially be an exciting extension to the present study.

Acknowledgements. This study is part of the European Climate Prediction system project (EUCP). The EUCP project is funded by the European Commission through the Horizon 2020 Programme for Research and Innovation: Grant Agreement 776613. [We are thankful for](#)

5 [the comments and suggestions of Francisco Doblas Reyes and three anonymous reviewers that helped us to improve our study.](#)

References

- Allan, R. and Ansell, T.: A new globally complete monthly historical gridded mean sea level pressure dataset (HadSLP2): 1850–2004, *Journal of Climate*, 19, 5816–5842, 2006.
- Baker, L., Shaffrey, L., Sutton, R., Weisheimer, A., and Scaife, A.: An intercomparison of skill and overconfidence/underconfidence of the wintertime North Atlantic Oscillation in multimodel seasonal forecasts, *Geophysical Research Letters*, 45, 7808–7817, 2018.
- Bröcker, J.: Evaluating raw ensembles with the continuous ranked probability score, *Quarterly Journal of the Royal Meteorological Society*, 138, 1611–1617, 2012.
- Brunner, L., McSweeney, C., Ballinger, A. P., Hegerl, G. C., Befort, D. J., O’Reilly, C., Benassi, M., Booth, B., Harris, G., Lowe, J., et al.: Comparing methods to constrain future European climate projections using a consistent framework, *Journal of Climate*.
- Brunner, L., Lorenz, R., Zumwald, M., and Knutti, R.: Quantifying uncertainty in European climate projections using combined performance-independence weighting, *Environmental Research Letters*, 14, 124 010, 2019.
- Cattiaux, J., Vautard, R., Cassou, C., Yiou, P., Masson-Delmotte, V., and Codron, F.: Winter 2010 in Europe: A cold extreme in a warming climate, *Geophysical Research Letters*, 37, 2010.
- Deser, C., Phillips, A. S., Alexander, M. A., and Smoliak, B. V.: Projecting North American climate over the next 50 years: Uncertainty due to internal variability, *Journal of Climate*, 27, 2271–2296, 2014.
- Deser, C., Terray, L., and Phillips, A. S.: Forced and internal components of winter air temperature trends over North America during the past 50 years: Mechanisms and implications, *Journal of Climate*, 29, 2237–2258, 2016.
- Deser, C., Lehner, F., Rodgers, K., Ault, T., Delworth, T., DiNezio, P., Fiore, A., Frankignoul, C., Fyfe, J., Horton, D., et al.: Insights from Earth system model initial-condition large ensembles and future prospects, *Nature Climate Change*, pp. 1–10, 2020.
- Doblas-Reyes, F., Andreu-Burillo, I., Chikamoto, Y., García-Serrano, J., Guemas, V., Kimoto, M., Mochizuki, T., Rodrigues, L., and Van Oldenborgh, G.: Initialized near-term regional climate change prediction, *Nature communications*, 4, 1–9, 2013.
- Doblas-Reyes, F. J., Hagedorn, R., and Palmer, T.: The rationale behind the success of multi-model ensembles in seasonal forecasting—II. Calibration and combination, *Tellus A: Dynamic Meteorology and Oceanography*, 57, 234–252, 2005.
- Field, C. B., Barros, V., Stocker, T. F., and Dahe, Q.: Managing the risks of extreme events and disasters to advance climate change adaptation: special report of the intergovernmental panel on climate change, Cambridge University Press, 2012.
- Fortin, V., Abaza, M., Anctil, F., and Turcotte, R.: Why should ensemble spread match the RMSE of the ensemble mean?, *Journal of Hydrometeorology*, 15, 1708–1713, 2014.
- Giorgi, F. and Mearns, L. O.: Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the “reliability ensemble averaging”(REA) method, *Journal of Climate*, 15, 1141–1158, 2002.
- Giorgi, F. and Mearns, L. O.: Probability of regional climate change based on the Reliability Ensemble Averaging (REA) method, *Geophysical research letters*, 30, 2003.
- Gleckler, P. J., Taylor, K. E., and Doutriaux, C.: Performance metrics for climate models, *Journal of Geophysical Research: Atmospheres*, 113, 2008.
- Gneiting, T., Raftery, A. E., Westveld III, A. H., and Goldman, T.: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation, *Monthly Weather Review*, 133, 1098–1118, 2005.
- Gneiting, T., Balabdaoui, F., and Raftery, A. E.: Probabilistic forecasts, calibration and sharpness, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69, 243–268, 2007.

- Guo, R., Deser, C., Terray, L., and Lehner, F.: Human influence on winter precipitation trends (1921–2015) over North America and Eurasia revealed by dynamical adjustment, *Geophysical Research Letters*, 46, 3426–3434, 2019.
- Hagedorn, R., Doblas-Reyes, F. J., and Palmer, T.: The rationale behind the success of multi-model ensembles in seasonal forecasting—I. Basic concept, *Tellus A: Dynamic Meteorology and Oceanography*, 57, 219–233, 2005.
- 5 Harris, I., Jones, P. D., Osborn, T. J., and Lister, D. H.: Updated high-resolution grids of monthly climatic observations—the CRU TS3. 10 Dataset, *International journal of climatology*, 34, 623–642, 2014.
- Hawkins, E. and Sutton, R.: Connecting climate model projections of global temperature change with the real world, *Bulletin of the American Meteorological Society*, 97, 963–980, 2016.
- Hersbach, H.: Decomposition of the continuous ranked probability score for ensemble prediction systems, *Weather and Forecasting*, 15, 10 559–570, 2000.
- Hewitt, C. D. and Lowe, J. A.: Toward a European Climate Prediction System, *Bulletin of the American Meteorological Society*, 99, 1997–2001, 2018.
- Jolliffe, I. T. and Stephenson, D. B.: *Forecast verification: a practitioner’s guide in atmospheric science*, John Wiley & Sons, 2012.
- Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., Arblaster, J. M., Bates, S., Danabasoglu, G., Edwards, J., et al.: The 15 Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability, *Bulletin of the American Meteorological Society*, 96, 1333–1349, 2015.
- Kettleborough, J., Booth, B., Stott, P., and Allen, M.: Estimates of uncertainty in predictions of global mean surface temperature, *Journal of climate*, 20, 843–855, 2007.
- Kharin, V. V. and Zwiers, F. W.: Improved seasonal probability forecasts, *Journal of Climate*, 16, 1684–1701, 2003.
- 20 Knutti, R.: The end of model democracy?, *Climatic Change*, 102, 395–404, 2010.
- Knutti, R. and Sedláček, J.: Robustness and uncertainties in the new CMIP5 climate model projections, *Nature Climate Change*, 3, 369, 2013.
- Knutti, R., Masson, D., and Gettelman, A.: Climate model genealogy: Generation CMIP5 and how we got there, *Geophysical Research Letters*, 40, 1194–1199, 2013.
- Knutti, R., Sedláček, J., Sanderson, B. M., Lorenz, R., Fischer, E. M., and Eyring, V.: A climate model projection weighting scheme account- 25 ing for performance and interdependence, *Geophysical Research Letters*, 44, 1909–1918, 2017.
- Maher, N., Milinski, S., Suarez-Gutierrez, L., Botzet, M., Kornblueh, L., Takano, Y., Kröger, J., Ghosh, R., Hedemann, C., Li, C., et al.: The Max Planck Institute grand ensemble-enabling the exploration of climate system variability, *Journal of Advances in Modeling Earth Systems*, 11, 2050–2069, 2019.
- Manzanas, R., Gutiérrez, J., Bhend, J., Hemri, S., Doblas-Reyes, F., Torralba, V., Penabad, E., and Brookshaw, A.: Bias adjustment and 30 ensemble recalibration methods for seasonal forecasting: a comprehensive intercomparison using the C3S dataset, *Climate Dynamics*, pp. 1–19, 2019.
- Matsueda, M., Weisheimer, A., and Palmer, T.: Calibrating climate change time-slice projections with estimates of seasonal forecast reliability, *Journal of Climate*, 29, 3831–3840, 2016.
- McKinnon, K. A. and Deser, C.: Internal variability and regional climate trends in an observational large ensemble, *Journal of Climate*, 31, 35 6783–6802, 2018.
- Merrifield, A. L., Brunner, L., Lorenz, R., and Knutti, R.: A weighting scheme to incorporate large ensembles in multi-model ensemble projections, *Earth System Dynamics Discussions*, pp. 1–30, 2019.

- O'Reilly, C. H., Woollings, T., and Zanna, L.: The dynamical influence of the Atlantic Multidecadal Oscillation on continental climate, *Journal of Climate*, 30, 7213–7230, 2017.
- Pachauri, R. K., Allen, M. R., Barros, V. R., Broome, J., Cramer, W., Christ, R., Church, J. A., Clarke, L., Dahe, Q., Dasgupta, P., et al.: Climate change 2014: synthesis report. Contribution of Working Groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change, *Ipcc*, 2014.
- Palmer, T. N., Alessandri, A., Andersen, U., Cantelaube, P., Davey, M., Delécluse, P., Déqué, M., Diez, E., Doblas-Reyes, F. J., Feddersen, H., et al.: Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER), *Bulletin of the American Meteorological Society*, 85, 853–872, 2004.
- Pasternack, A., Bhend, J., Liniger, M. A., Rust, H. W., Müller, W. A., and Ulbrich, U.: Parametric decadal climate forecast recalibration (DeFoReSt 1.0), *Geoscientific Model Development*, 2018.
- Sanderson, B. M., Knutti, R., and Caldwell, P.: A representative democracy to reduce interdependency in a multimodel ensemble, *Journal of Climate*, 28, 5171–5194, 2015.
- Sansom, P. G., Ferro, C. A., Stephenson, D. B., Goddard, L., and Mason, S. J.: Best practices for postprocessing ensemble climate forecasts. Part I: Selecting appropriate recalibration methods, *Journal of Climate*, 29, 7247–7264, 2016.
- Smith, D., Eade, R., Scaife, A. A., Caron, L.-P., Danabasoglu, G., DelSole, T., Delworth, T., Doblas-Reyes, F., Dunstone, N., Hermanson, L., et al.: Robust skill of decadal climate predictions, *npj Climate and Atmospheric Science*, 2, 1–10, 2019.
- Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V., Midgley, P. M., et al.: Climate change 2013: The physical science basis, 2013.
- Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An overview of CMIP5 and the experiment design, *Bulletin of the American Meteorological Society*, 93, 485–498, 2012.
- Tippett, M. K. and Barnston, A. G.: Skill of multimodel ENSO probability forecasts, *Monthly Weather Review*, 136, 3933–3946, 2008.
- Wallace, J. M., Fu, Q., Smoliak, B. V., Lin, P., and Johanson, C. M.: Simulated versus observed patterns of warming over the extratropical Northern Hemisphere continents during the cold season, *Proceedings of the National Academy of Sciences*, 109, 14 337–14 342, 2012.
- Wilks, D. S.: Comparison of ensemble-MOS methods in the Lorenz'96 setting, *Meteorological Applications*, 13, 243–256, 2006.
- Wilks, D. S.: *Statistical methods in the atmospheric sciences*, vol. 100, Academic press, 2011.
- Yeager, S., Danabasoglu, G., Rosenbloom, N., Strand, W., Bates, S., Meehl, G., Karspeck, A., Lindsay, K., Long, M., Teng, H., et al.: Predicting near-term changes in the Earth System: A large ensemble of initialized decadal prediction simulations using the Community Earth System Model, *Bulletin of the American Meteorological Society*, 99, 1867–1886, 2018.

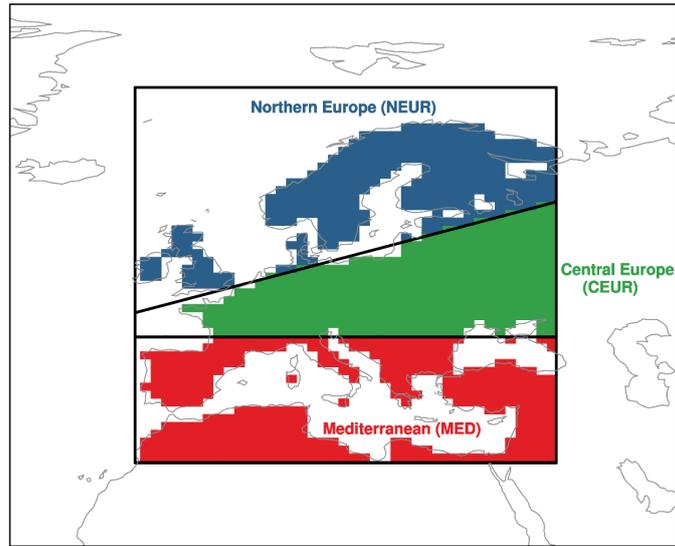


Figure 1. The SREX regions over which area-averaged projections and observations are analysed in this study, following Field et al. (2012).

Example of the LENS projection of summertime Central European temperature (top) and precipitation (bottom) decomposed from the full anomalies into dynamical and residual components. The lines show the ensemble medians and the shading shows the 90% range of the ensemble.

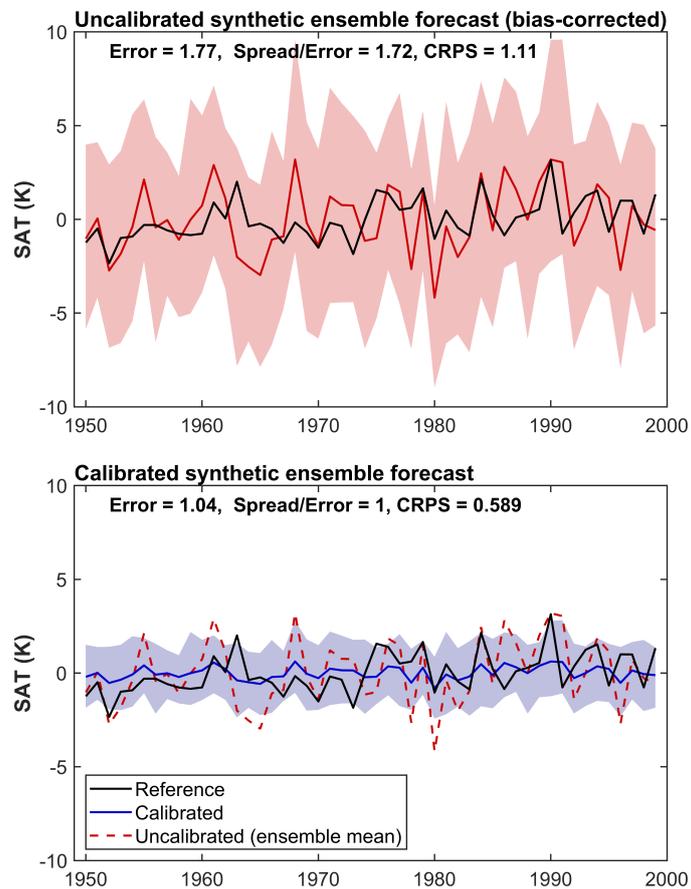


Figure 2. (Top) Synthetic data for an example (bias-corrected) ensemble temperature evolution is shown for the ensemble mean in red and 90% ensemble range (shaded), along with the synthetic (observational) reference index in black. (Bottom) The synthetic ensemble calibrated using the variance inflation method to match the reference dataset shown in blue with the raw ensemble mean shown in dashed red. The RMS Error, spread/error and CRPS calculated for the raw ensemble and the calibrated ensemble are all shown.

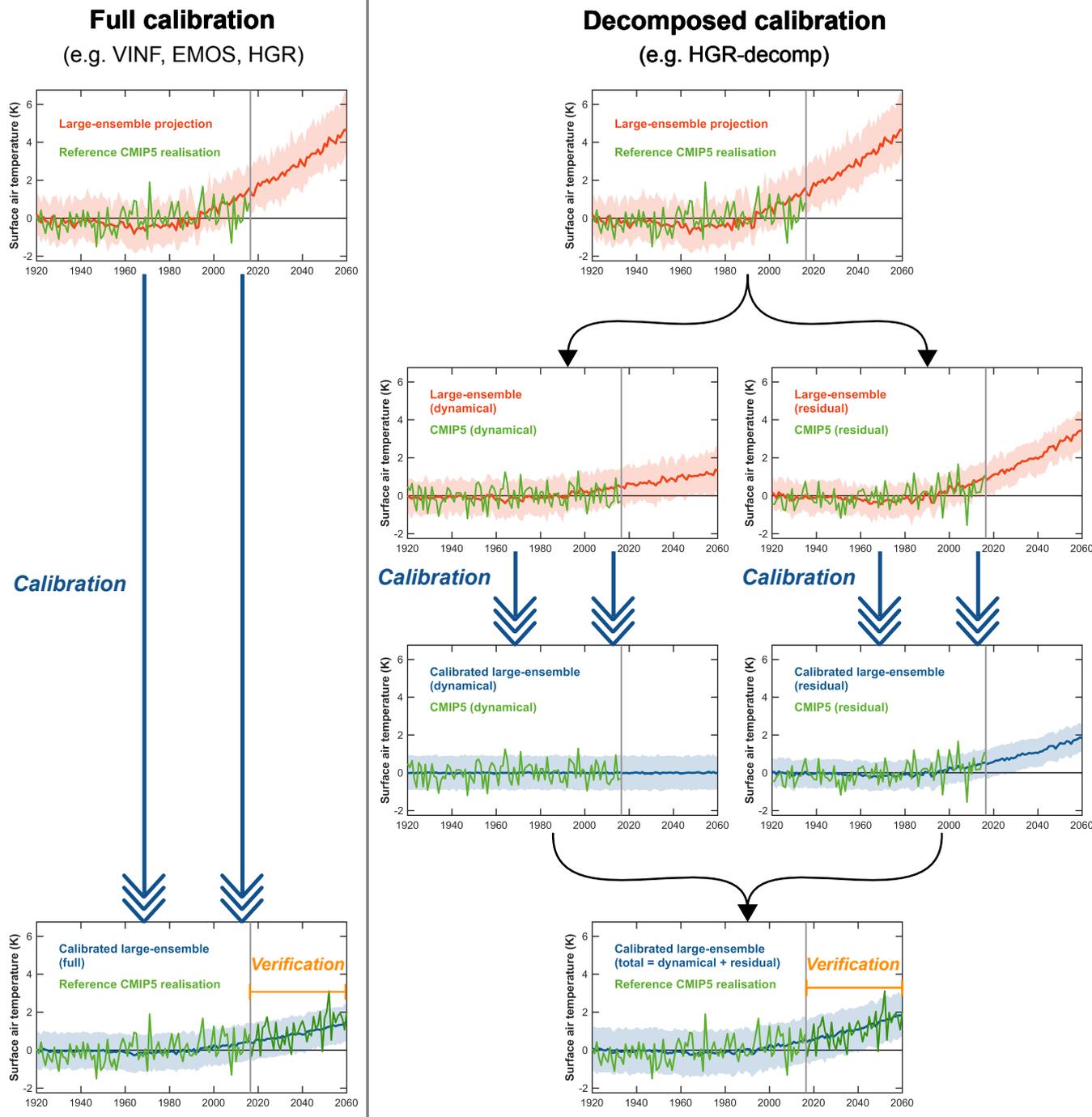


Figure 3. ~~Example~~ An example showing the steps of the ~~raw CESM Large Ensemble Northern Europe temperature full calibration methods~~ (red left hand column) ~~, which~~ and calibration of the dynamically decomposed variables (right column). This example is ~~calibrated to fit a single realisation from the CMIP5 model realisation calibration of the summer (JJA) Central European (green CEUR) over the observational period, 1920-2016.~~ The calibrated CESM Large Ensemble temperature using the variance inflation method is shown in ~~blue~~ CESM1-LE and one of the CMIP5 models. The shading indicates ~~shows~~ the 90-95% range across ~~of~~ the CESM1-LE ensemble. The effectiveness of the calibration is assessed by verifying over data from the period 2017-2060, which is withheld during the calibration step.

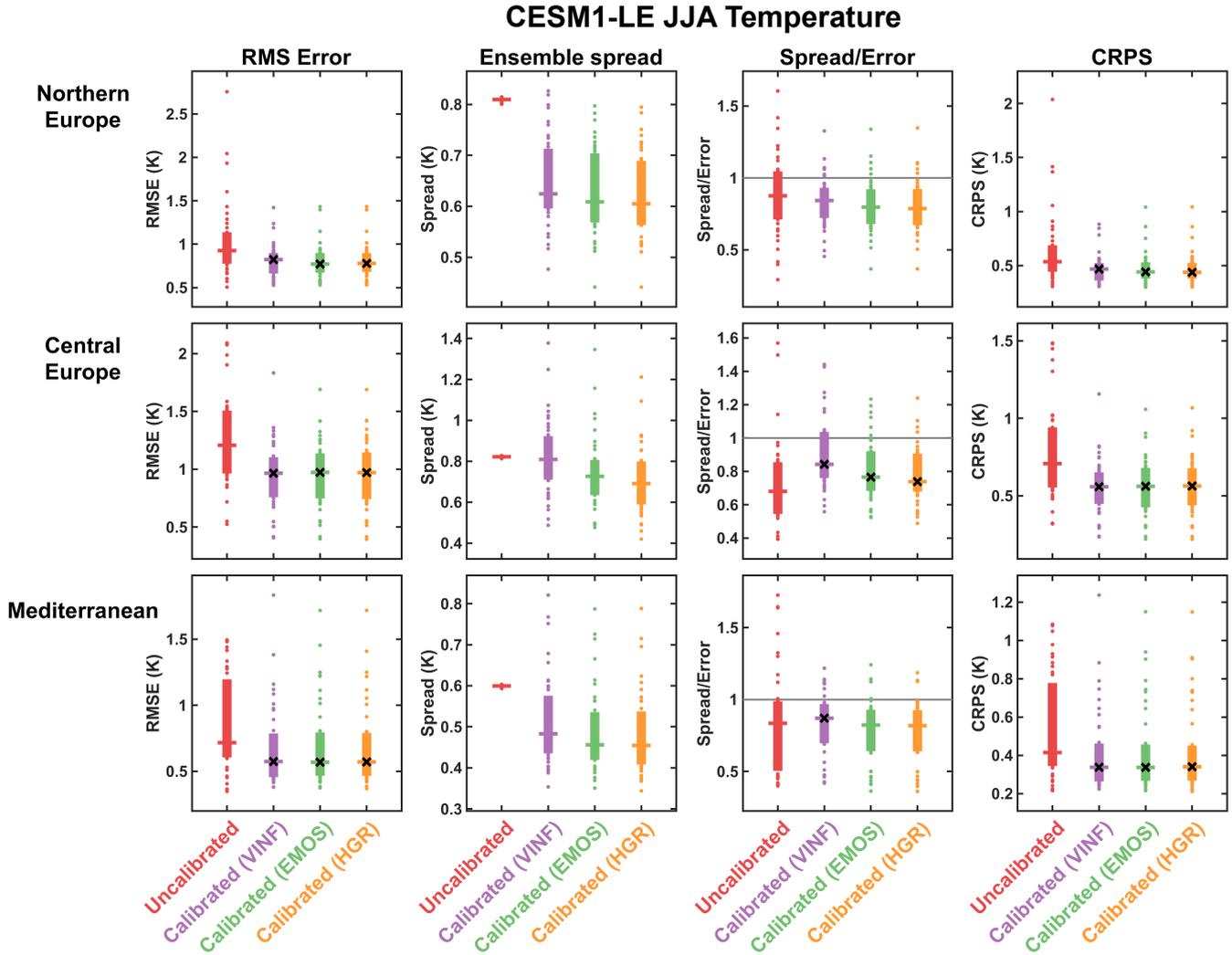


Figure 4. Comparison of calibration methods applied to the [LENS-CESM1-LE](#) summer temperature projections calibrated to the CMIP5 models over the observational period (1920-2016) and verified using the 44 years in the out-of-sample period ([1917-2060](#)[2017-2060](#)). The verification statistics for each of the individual CMIP5 models are shown in dots, the interquartile range of this distribution is shown by the solid bars and the median is indicated by the horizontal lines. 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. The significance levels were calculated using the non-parametric Mann-Whitney U-test, [applied to the distributions of the verification scores from the 39 CMIP5 models](#).

CESM1-LE: Uncalibrated, Calibrated (HGR) & Calibrated (HGR-decomp)

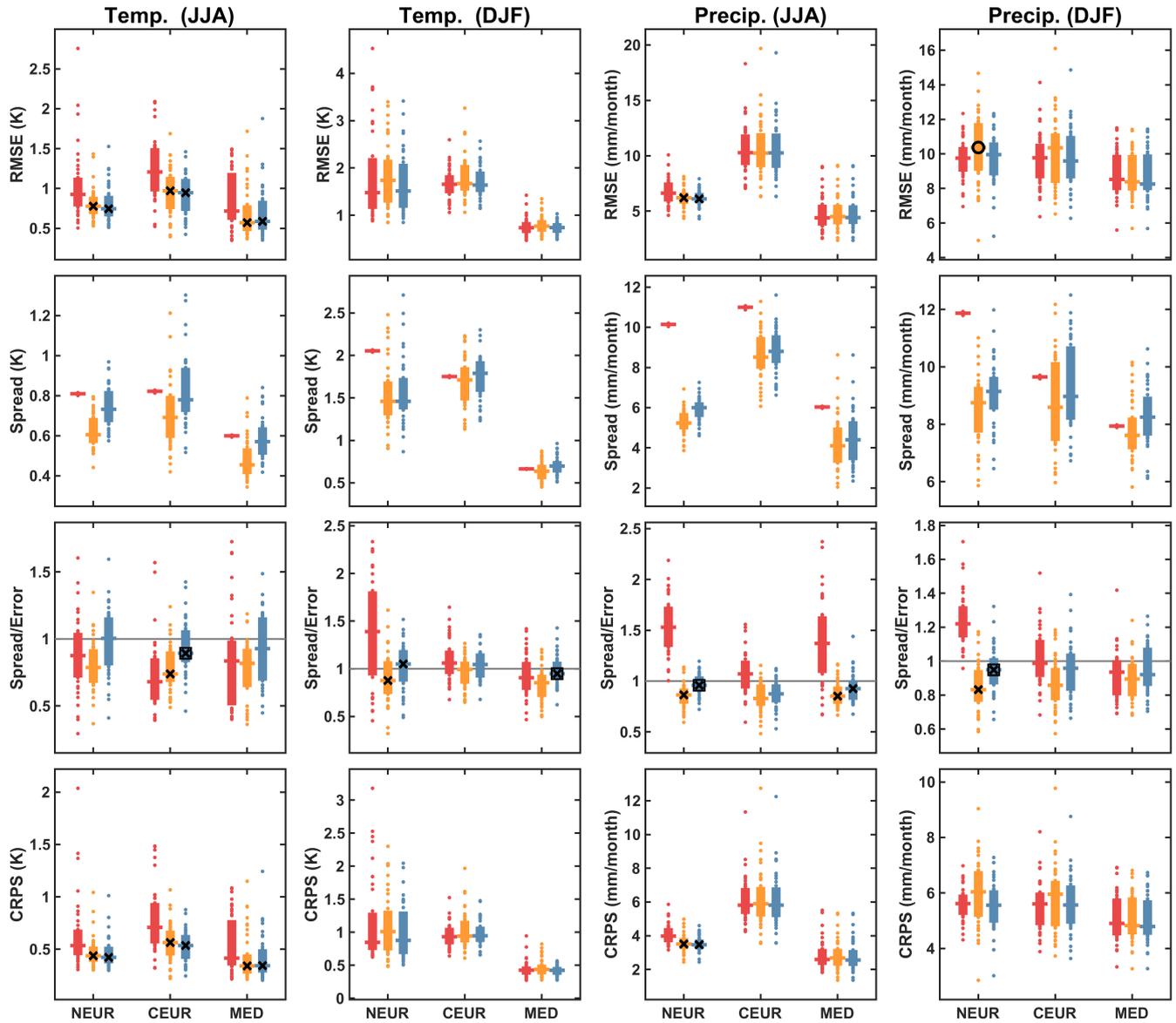


Figure 5. Overview of verification of the HGR and HGR-decomp calibration methods compared with the uncalibrated [LENS-CESM1-LE](#) data in the European regions. [Shown-Results are shown](#) for all of the verification measures, for both summer and winter seasons and for temperature and precipitation. The verification statistics for each of the individual CMIP5 models are shown in dots, the interquartile range of this distribution is shown by the solid bars and the median is indicated by the horizontal lines. 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. Black circles indicate where the calibration is significantly worse than the uncalibrated ensemble (at the 90% level). [The-black-Black boxes show-indicate](#) where [one-calibration-the HGR-decomp](#) method of calibration is found to be significantly better than the [other-calibration-HGR](#) method for the same variable, season and region (at the 90% significance-level). The significance levels were calculated using the non-parametric Mann-Whitney U-test, [applied to the distributions of the verification scores from the 39 CMIP5 models.](#)

2041-2060 verification: **Uncalibrated (CESM1-LE)**, **Calibrated (CESM1-LE)**, **Uncalibrated (MPI-GE)** & **Calibrated (MPI-GE)**

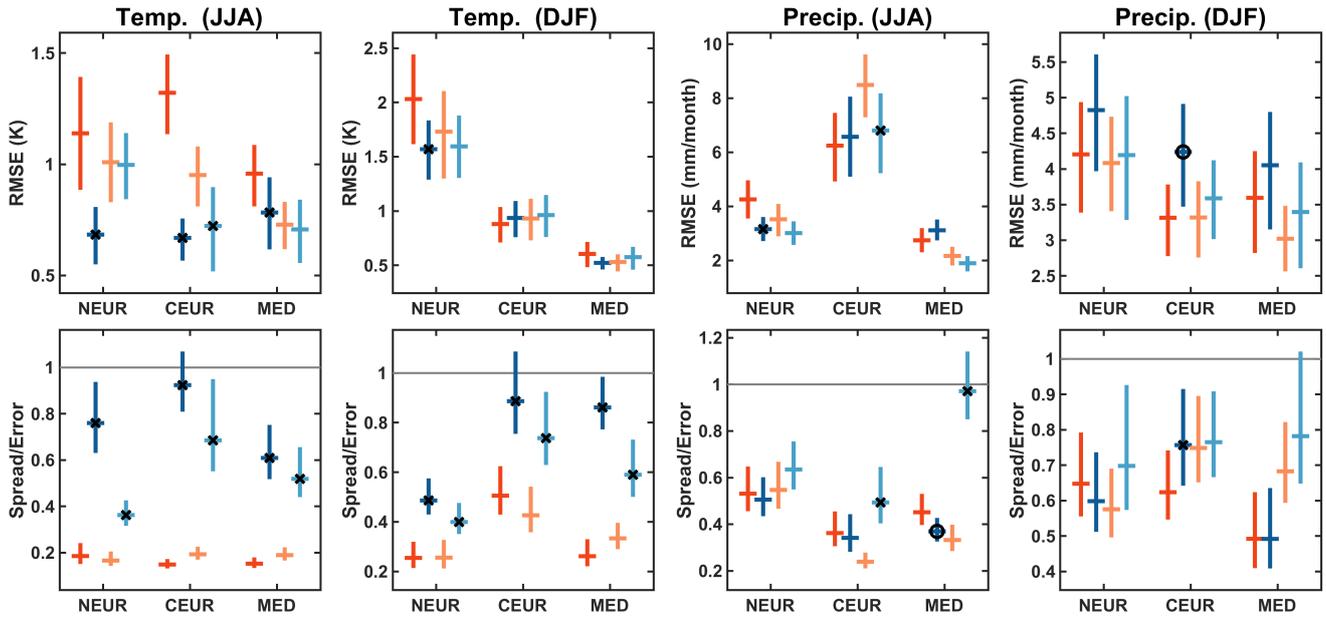


Figure 6. Verification of the 2041-2060 mean projections calculated relative to the out-of-sample CMIP5 models for both the **LENS** **CESM1-LE** and **MPIGE-MPI-GE** datasets. The horizontal lines show the mean across all models and the vertical lines show the 90% confidence intervals, calculated by randomly resampling across the CMIP5 models with replacement 1000 times. The black crosses indicate where the calibrated ensemble has-a-is significantly better than the equivalent uncalibrated ensemble; the black circles indicate where the calibrated ensemble is significantly worse than the uncalibrated ensemble.

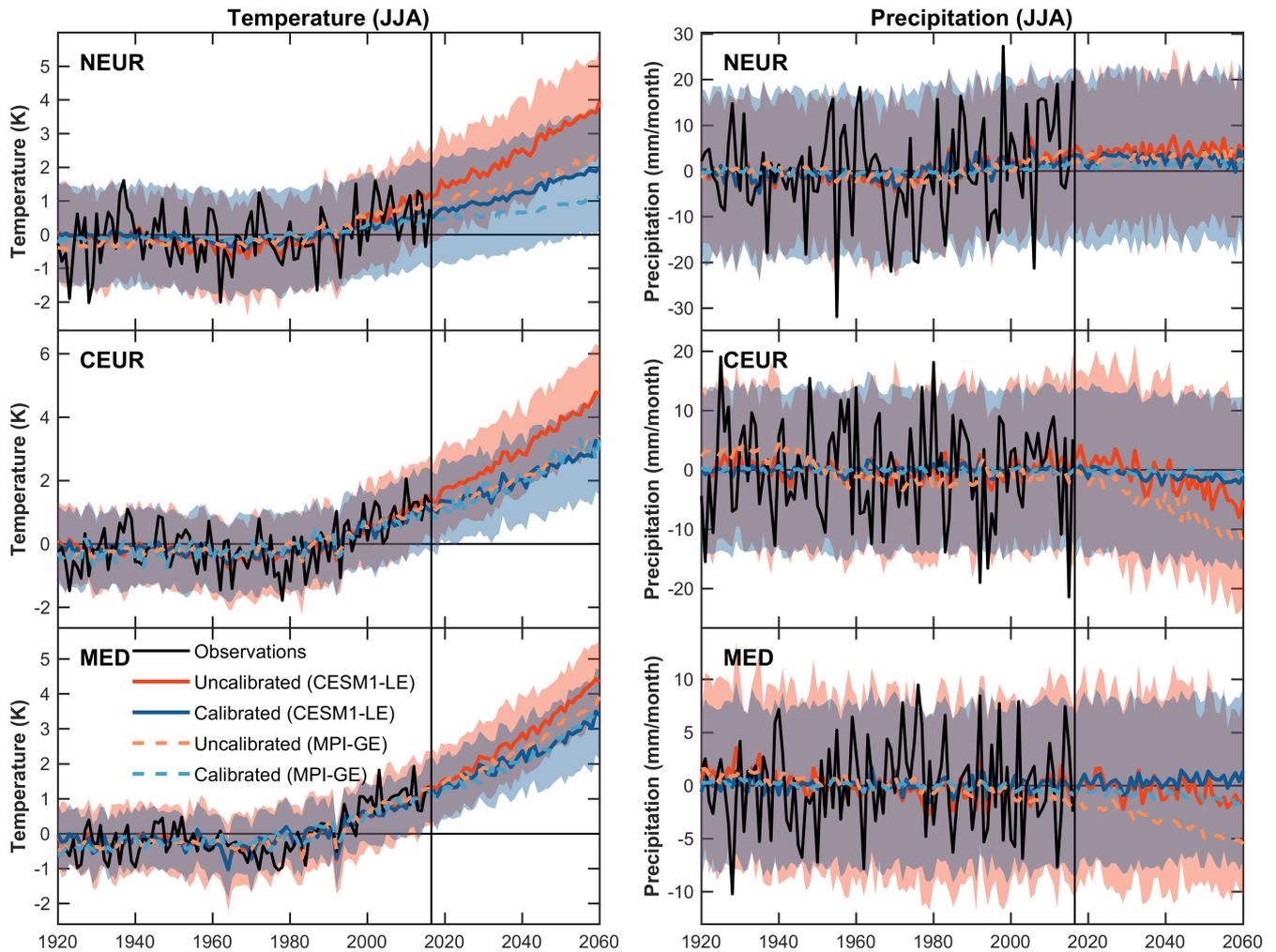


Figure 7. Uncalibrated and calibrated (HGR-decomp) [LENS-CESM1-LE](#) projections, where here the calibrated projections have been calibrated against the observations over the period 1920-2016. The lines show the ensemble medians for the uncalibrated and calibrated ensembles for both the [LENS-CESM1-LE \(solid\)](#) and MPI-GE ([dashed](#)) datasets. The shading shows the 95-95% range of the [LENS-CESM1-LE](#) ensemble. Based on the verification out-of-sample tests using the CMIP5 models the calibrated ensemble is expected to be more reliable than the uncalibrated ensemble, particularly for temperatures.

2041-2060 change: **Uncalibrated (CESM1-LE)**, **Calibrated (CESM1-LE)**, **Uncalibrated (MPI-GE)** & **Calibrated (MPI-GE)**

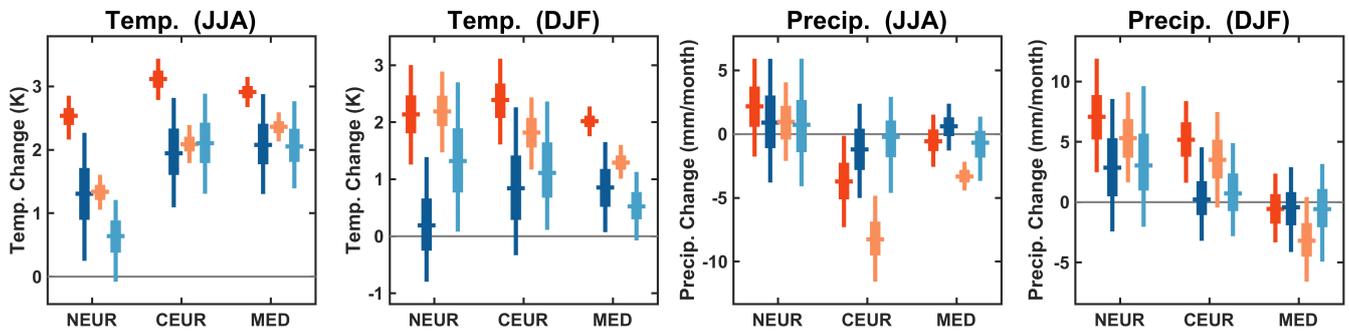


Figure 8. 2041-2060 mean calculated relative to 1995-2014 climatology for both [LENS-CESM1-LE](#) and [MPI-GE-MPI-GE](#), calibrated [using the HGR-decomp method](#) to the observations over the period 1920-2016. The vertical lines show the 90% range of the ensemble, thick boxes show the interquartile range and horizontal lines show the ensemble median.