Calibrating large-ensemble European climate projections using observational data

Christopher H. O'Reilly[1,2], Daniel J. Befort[1], and Antje Weisheimer[1,2]

[1]Atmospheric, Oceanic and Planetary Physics, Department of Physics, University of Oxford.
[2]NCAS-Climate, Department of Physics, University of Oxford.

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

Abstract.

This study examines methods of calibrating projections of future regional climate 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, is used for the subsequent analysis. An extension to this method - applying it 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 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.

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 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 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). 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 (e.g. Giorgi and Mearns, 2002, 2003; Knutti et al., 2017; Sanderson et al., 2015) 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 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 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.

Despite large ensembles providing clearer estimates of forced climate signal and internal variability, it is obvious that that these ensemble 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

5   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

10 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

15 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 - and verified against - CMIP5 model data and also applied to observations. Conclusions follow in section 4.

20 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 Large Ensemble (Kay et al., 2015), hereafter referred to as "LENS", 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

25 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 but uses the MPI Earth System Model and consists of 100 members starting in 1850. 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. We only use 99 members of the MPI-GE that had all of the variables used here available

30 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 data is taken from the HadSLP2 dataset (Allan and Ansell, 2006).

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. Data from the CMIP5 models, MPI-GE ensemble and the observational datasets were regridded to the same grid as the LENS dataset prior to the analysis.

5   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).

10   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, 15   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 lies with respect to the underlying ensemble and is higher when the verification 20   data is further from the centre of the ensemble. As such, a lower CRPS value represents a more skillful probabilistic forecast.

2.3   Ensemble calibration methods

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

25   2.3.1   Uncalibrated ensemble

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 30   calibration methods, which is generally the observational period 1920-2016 in the following analysis.

2.3.2 Variance inflation (VINF)

One calibration method that we will test is "Variance inflation", hereafter referred to as VINF, following Doblas-Reyes et al. (2005). For each uncalibrated ensemble member, X_{uncalib}, VINF 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

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

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

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

The scaling variables α and β are calculated as

$$\alpha = \rho \frac{s_{\text{r}}}{s_{\text{m}}}; \tag{3}$$

$$\beta = \sqrt{1 - \rho^2} \frac{s_{\text{r}}}{s_{\text{uncalib}}}; \tag{4}$$

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 between the ensemble and, 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.

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

$$X_{\text{uncalib}}(t) = X_{\text{m}}(t) + \epsilon_{\text{uncalib}}(t), \qquad \epsilon_{\text{uncalib}}(t) = N[0, s^2(t)]; \tag{5}$$

and the calibrated ensemble is expressed as

$$X_{\text{calib}}(t) = b X_{\text{m}}(t) + \epsilon_{\text{calib}}(t), \qquad \epsilon_{\text{calib}}(t) = N[0, c + d s^2(t)]; \tag{6}$$

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 Gneiting et al. (2007). The coefficients b, c, and d are constrained to be non-negative values. The EMOS technique allows for meaningful differences in spread across the ensemble at different time (i.e. the coefficient d), so is sometimes referred to as

5 "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 ensemble projection. To avoid overfitting to the observations when producing the ensembles and to include some measure of sampling uncertainty in the fitting process, the EMOS method is applied to randomly resampled years from the calibration period, to produce 1000 valid combinations of the coefficients b, c,

10 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)

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 time and is expressed as

$$X_{\text{calib}}(t) = bX_{\text{m}}(t) + \epsilon_{\text{calib}}(t), \qquad \epsilon_{\text{calib}}(t) = N[0, c]. \tag{7}$$

15 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

20 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). The dynamical decomposition splits the variables

25 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 corrensponding month from other years over the reference/observational period (1920-2016). This regression fit

30 yields weights which are then used to computed the associated dynamical surface temperature or precipitation anomaly. Each field can then be separated into a dynamical and residual component. 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. 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 the ensemble, the reference frequently lies outside the ensemble. 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 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 *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 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 period, 1920-2016. The future period is then used to analyse the impact of the calibration by verifying against 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 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). Overall, the calibration periods tended 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 Figure 3. In this example, the raw LENS data for NEUR summer temperature is shown in red , along with the same index from one of the CMIP5 models.

The model is calibrated over the 97 year reference period (1920-2016) - shown by the black vertical 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 CMIP5 model index. For both the LENS and MPI-GE large ensemble datasets, this analysis was performed for all 39 CMIP5 models. This process was then repeated for both temperature and precipitation, for summer and winter, and over

5 all three European regions.

The verification statistics for the LENS 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-

10 parametric Mann-Whitney U-test (e.g. Wilks, 2011). For the summer temperature over all three regions, all of the calibration methods lower the RMS error of the ensemble projection compared with the uncalibrated ensemble. 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

15 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 ensemble in the out-of-sample future 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 *signficantly* negative impact on the projections.

We also tested the different calibration methods on the MPI-GE, for precipitation and for the different seasons. The verifi-

20 cation measures for the MPI-GE summer temperature are shown in Figure S1. The performance of the calibration methods on the MPI-GE summer temperature is qualitatively similar to that for the LENS (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 data (i.e. Figure 4). For the other regions there is an improvement over the uncalibrated ensemble but

25 this is not significant for any of the calibration methods, or for any of the verification measures. Nonetheless, as in the LENS data, none of the calibration methods displays a signficantly 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 S1). Analysis of the equivalent figures for precipitation and the winter season also show a reasonably consistent performance between the

30 calibration methods (not shown). The similarity of 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 are shown in Figure 5 for the calibrated LENS projections (note that the red and orange data in the first column

35 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 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 precipitation projections. The spread of the

5 uncalibrated LENS 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

10 regions or seasons, indicating that the overall impact of the calibration in to generally improve 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 dataset (Figure 5). Whilst it is only one of the verification measures performed across both the LENS and MPI-GE datasets, it is a concern because it reduces how much confidence we can have in applying the calibration using observations.

15 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

20 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 circulation driven signal 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.

25 An example of the dynamical decomposition, applied to summertime projections for the CEUR region in the LENS dataset, is shown in Figure 6. 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.

30 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. We use the HGR method to do the separate decompositions, and refer to this methods 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 S2) As with the HGR verification, the crosses/circles indicate where the verification of the HGR-decomp calibrated ensemble is 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 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 S2), albeit with a lower level of significance. Specifically, in ten of the twelve total regions/variables verified for the MPI-GE dataset, HGR-decomp calibraton 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 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 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 appears to be slightly "over-fitting" the ensemble to the reference period, resulting in a consistent 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 to combine the single measurements to produce one verification score across all the models. To estimate the uncertainty of these verification measures, we performed a bootstrap resampling over the 39 CMIP5 model projection/verification pairs. Verification results for the 2041-2060 climatologies for both the LENS and MPI-GE are shown in Figure 7.

The HGR-decomp calibration tends to improve the projected 2041-2060 climatology of temperature. 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 and MPI-GE ensembles but all tend to somewhat overconfident. The winter temperature shows less obvious improvement in terms

5   of RMS error of the calibrated projections, but the reliability is significantly improved for all the regions in both ensembles, but again, the calibrated LENS 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 dataset. The calibration has the least influence on the 2041-2060 climatology of precipitation, acting to

10   worsen the RMS error in some 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

15   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. Another reason is that much of the increased reliability in the yearly projections comes 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

20   this internal variability.

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

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. On the basis of this analysis, we will now apply the HGR-decomp calibration method to the large

25   ensembles, targeted at the observational indices of temperature and precipitation.

The calibrated LENS projections for the summer temperature and precipitation are shown in Figure 8. Based on the imperfect model tests we expect the calibrated summer LENS to represent the most accurate and reliable projection from the all of the ensemble/variable/season combinations tested. For the summer temperature projections, calibrating the LENS against the observations until 2016, the rate of warming until 2060 is reduced by varying amounts. There is also a noticeable change in

30   spread, particularly for the NEUR region but also evident in the other regions, indicating that the calibration method is acting to increase the uncertainty in the future projections. For the precipitation, the signal in LENS projection is much weaker with respect to the inter-annual variability. In the projections shown here the calibration has a fairly modest 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, with the ensemble medians of the MPI-GE also plotted in Figure 8 for comparison (the full ensemble projection plots are shown in Figure S3). 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

5 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 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

10 drying trend over the whole observational period in these regions that is not seen very clearly in the observations. Based on the imperfect model tests, however, we have less confidence in the performance of the calibrated ensembles for precipitation.

All the projected changes in the 2041-2060 climatological changes from the present day reference period, 1995-2014, are shown in Figure 9. Here, we have plotted both the uncalibrated and (HGR-decomp) calibrated climatological changes for both the LENS and MPI-GE datasets. An interesting feature of these projected changes is that for many of them, the calibrated

15 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 reas-

20 suring 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

25 of the projections (e.g. Figure 7), 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 projection has a wider spread than in the calibrated MPI-GE projection for many of the projected temperature indices. In the imperfect model tests, shown in Figure 7, the calibrated temperature projections for the LENS 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

30 should have more confidence in the broader calibrated LENS projections for future temperature changes.

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 - VINF, EMOS and HGR - are more commonly used for initialised forecasts from

Earth System
Dynamics
Discussions

Open Access
EGU

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 act to scale the ensemble signal and spread so as to optimize the fit over the reference period. The three calibration methods displayed similar performance, all generally improving the out-of-sample projections in comparison to the uncalibrated ensemble. The simplest of the calibration

5   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 vari-

10   ables 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

15   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 in the calibrated ensemble, when subject to the imperfect model tests. The performance of the calibration was substantial for the temperature projections but for

20   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 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

25   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 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

30   certainly a more reliable probabilistic projection of European climate over the next 40 or so years. In both the LENS and MPI-GE datasets the projected increase in European temperatures was generally smaller compared to the uncalibrated ensembles. The calibrated projections are much 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

35   (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

5   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 (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

10   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, 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*).

One novel aspect of this study that is particularly worth emphasising is the imperfect model testing approach. Previous

15   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

20   method, which may result in over-fitting or an over-estimation of the added-values 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 kept separate from the underlying large ensembles throughout, only being used to compare the efficacy of the different methods.

As well as being applied to other datasets and regions, this calibration method can also be applied to initialised decadal forecasts. Previous studies have examined how similar calibration methods can improve multi-year forecasts (e.g. Sansom

25   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 dataset analysed in this study has an initialised counterpart, namely the Decadal Prediction Large Ensemble, and testing how to combine data from these different ensembles to produce a merged calibrated climate predictions would potentially be an exciting extension to the present 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
5 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., 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.

10 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
15 past 50 years: Mechanisms and implications, Journal of Climate, 29, 2237–2258, 2016.

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.

20 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, Geophys-
25 ical 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.

30 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.

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.

Harris, I., Jones, P. D., Osborn, T. J., and Lister, D. H.: Updated high-resolution grids of monthly climatic observations–the CRU TS3. 10
35 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, 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.

5   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 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
10   climate, 20, 843–855, 2007.

Kharin, V. V. and Zwiers, F. W.: Improved seasonal probability forecasts, Journal of Climate, 16, 1684–1701, 2003.

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
15   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 accounting 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
20   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 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 reliabil-
25   ity, Journal of Climate, 29, 3831–3840, 2016.

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
30   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
35   (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.
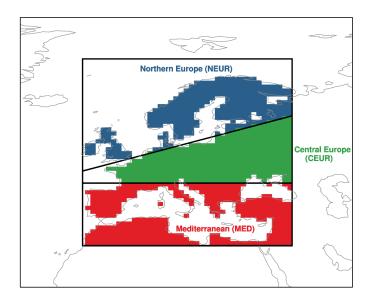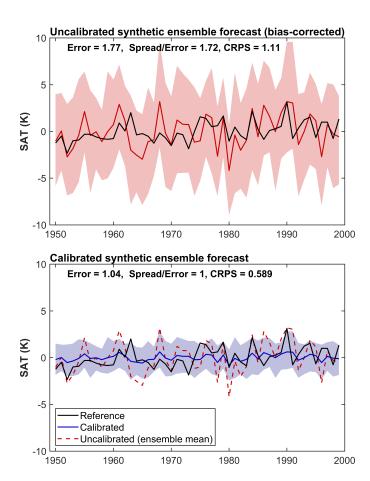
Figure 1. The SREX regions over which area-averaged projections and observations are analysed in this study, following Field et al. (2012).

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.

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.

5 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.

10 Wilks, D. S.: Comparison of ensemble-MOS methods in the Lorenz'96 setting, Meteorological Applications, 13, 243–256, 2006.

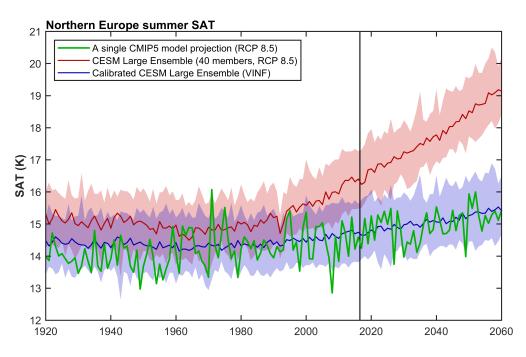Wilks, D. S.: Statistical methods in the atmospheric sciences, vol. 100, Academic press, 2011.

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.
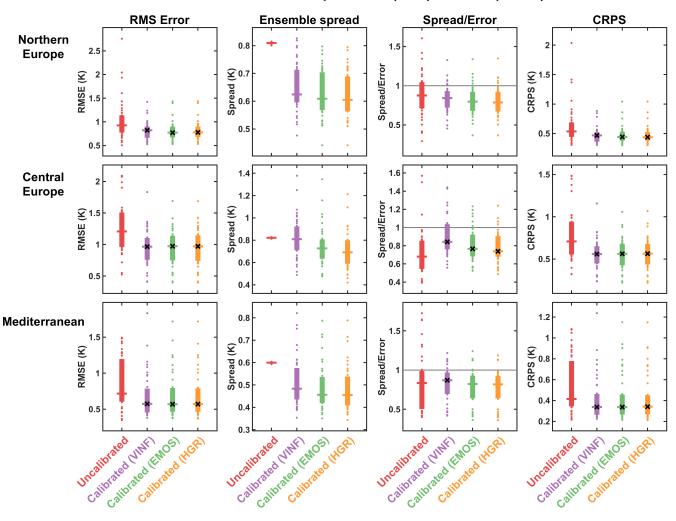
Figure 3. Example of the raw CESM Large Ensemble Northern Europe temperature (red), which is calibrated to fit a single realisation from the CMIP5 model realisation ((green) over the observational period, 1920-2016. The calibrated CESM Large Ensemble using the variance inflation method is shown in blue. The shading indicates the 90% range across the ensemble.

Figure 4. Comparison of calibration methods applied to the LENS 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). 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.
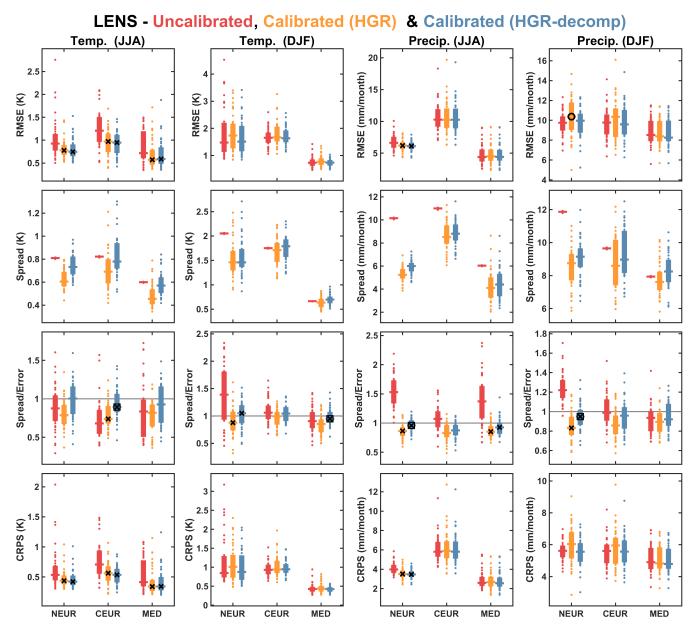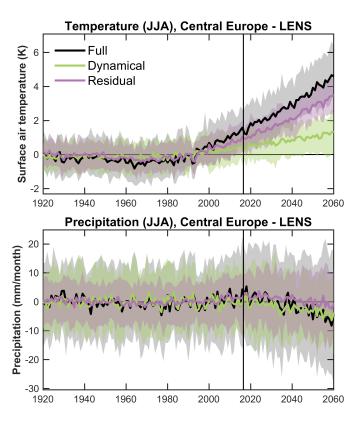
Figure 5. Overview of verification of the HGR and HGR-decomp calibration methods compared with the uncalibrated LENS data in the European regions. Shown for all of the verification measures, for both summer and winter seasons and for temperature and precipitation. The verification statistics for each 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 boxes show where one calibration method is found to be significantly better than the other calibration 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.

Earth System
Dynamics
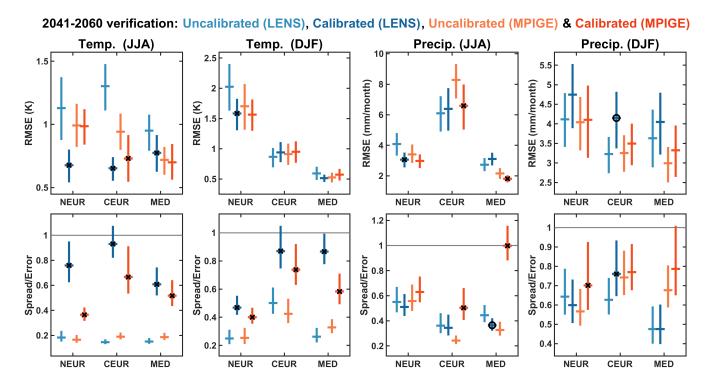Discussions

Open Access

EGU

Figure 6. 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.
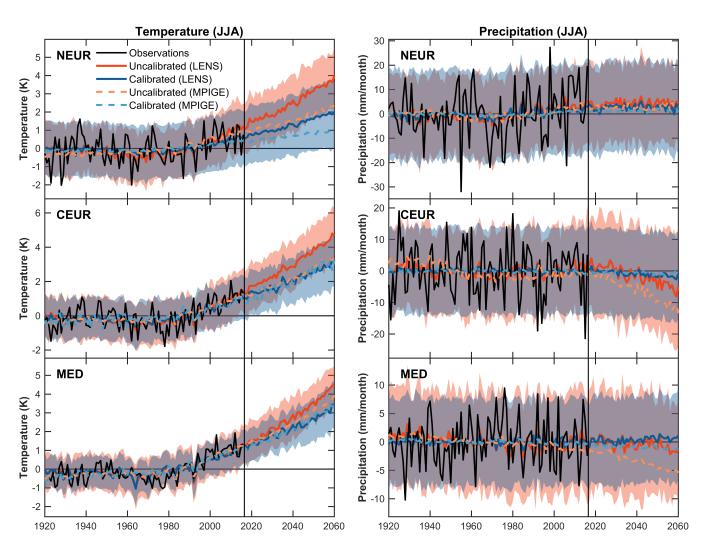
Figure 7. Verification of the 2041-2060 mean projections calculated relative to the out-of-sample CMIP5 models for both the LENS and MPIGE datasets. The horizontal lines show the mean across all models and the vertical lines show the 90% confidence intervals, calculated by randomly resampling across the CMIP5 models with replacement. The black crosses indicate where the calibrated ensemble has a significantly better than the equivalent uncalibrated ensemble; the black circles indicate where the calibrated ensemble is significantly worse that the uncalibrated ensemble.
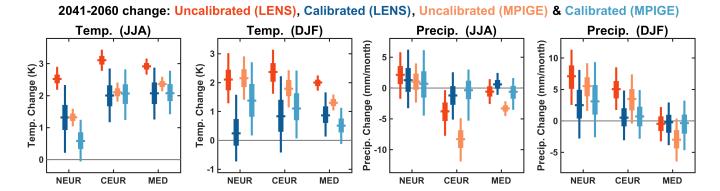
Figure 8. Uncalibrated and calibrated (HGR-decomp) LENS projections, where here the calibrated projections have been calibrated against the observations over the period 1920-2016. The lines show the ensemble medians for the uncalibrated and calibrated ensembles for both the LENS and MPI-GE datasets. The shading shows the 90% range of the LENS 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.

Figure 9. 2041-2060 mean calculated relative to 1995-2014 climatology for both LENS and MPIGE, calibrated 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.