Earth System
Dynamics

Calibrating large-ensemble European climate projections using observational data

Christopher H. O'Reilly[1,2], **Daniel J. Befort**[1], and **Antje Weisheimer**[2,3]

[1]Atmospheric, Oceanic and Planetary Physics, Department of Physics, University of Oxford, Oxford, UK
[2]NCAS-Climate, Department of Physics, University of Oxford, Oxford, UK
[3]European Centre for Medium-Range Weather Forecasts (ECMWF), Reading, UK

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 Community Earth System Model (CESM) Large Ensemble and Max Planck Institute (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/representative concentration pathway 8.5 (RCP8.5) simulations from the Coupled Model Intercomparison Project 5 (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. As an extension to the HGR calibration method it is applied to dynamically decomposed data, in which the underlying data are separated into dynamical and residual components (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.

1 Introduction

To make informed assessments of climate impacts and implement relevant adaptation strategies, reliable climate projections are important for policymakers 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 estimat-

ing 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; Merrifield et al., 2020) 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 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. Deser et al., 2020). 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 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.

The paper is organised as follows. The datasets, verification techniques and calibration methods are described in the next section. In Sect. 3, we present results from the different calibration methods, namely, "variance inflation", "ensemble model output statistics" and "homogeneous Gaussian regression". These calibration methods are applied to, and verified against, CMIP5 model data and also applied to observations. Conclusions follow in Sect. 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 Community Earth System Model version 1 (CESM1) Large Ensemble (Kay et al., 2015), hereafter referred to as "CESM1-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 (RCP8.5) over the period 2006–2100. The second large-ensemble dataset is the Max Planck Institute (MPI) Grand Ensemble (Maher et al., 2019), hereafter referred to as "MPI-GE", which is similar to CESM1-LE but uses the MPI Earth System Model and consists of 100 members starting in 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 RCP8.5 data to compare with 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 are taken from the CRU-TS v4.01 gridded surface dataset (Harris et al., 2014). The observational sea-level pressure (SLP) data are 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; Compo et al., 2011). The results were generally very similar regardless of the observational datasets; however, some of the differences from the observational dataset are highlighted in Sect. 3.5.

The data used for out-of-sample verification were 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 RCP8.5 (from 2006) scenarios. The CESM1-LE has a 1° × 1° 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° 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 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.
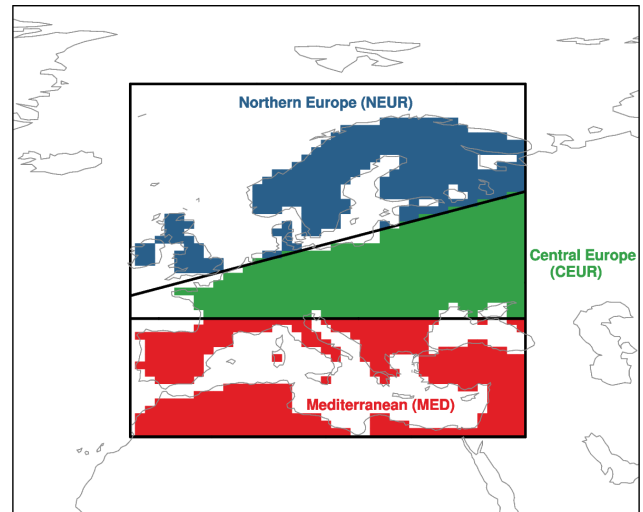

Figure 1. The SREX regions over which area-averaged projections and observations are analysed in this study, following Field et al. (2012).

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 Fig. 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 (RMSE) 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 RMSE 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 1 (e.g. Jolliffe and Stephenson, 2012). A spread / error ratio greater than 1 indicates an underconfident ensemble, whereas a spread / error ratio less than 1 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

are further from the centre of the ensemble. As such, a lower CRPS value represents a more skilful probabilistic forecast.

2.3 Ensemble calibration methods

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

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 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, $X_{\mathrm{ens\text{-}anom}}$, from the uncalibrated ensemble. The uncalibrated ensemble can be expressed in these terms as

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

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:

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

The scaling variables α and β are calculated as

$$\alpha = \rho \frac{s_{\mathrm{r}}}{s_{\mathrm{m}}}, \tag{3}$$

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

where s_{r} is the standard deviation of the reference (or observational) data that are 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 0 and there is no skilful correlation

between the ensemble and the reference dataset, we set ρ to be 0 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

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

and the calibrated ensemble is expressed as

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

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 is arguably the most general method we will test because it allows for meaningful differences in spread across the ensemble at different times (i.e. the coefficient d), so it is sometimes referred to as "nonhomogeneous Gaussian regression" (e.g. Wilks, 2006; Tippett and Barnston, 2008). EMOS represents a simplification over the VINF method because 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 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 Homogeneous Gaussian regression (HGR)

The third calibration method that we test is homogeneous Gaussian regression, hereafter referred to as HGR. The HGR method is a simplified version of EMOS, in which the calibrated variance is constant in time and is expressed as

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

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 0.

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; Guo et al., 2019). 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 analogue 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, 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 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 Fig. S1 in the Supplement (and also in the example calibration schematic in Fig. 3). The regional dynamical and residual time series were calibrated using the above techniques towards the corresponding dynamical and residual time series from the target dataset (i.e. CMIP5 or observations over the period 1920–2016). The calibrated dynamical and residual time series 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 Fig. 2, alongside a synthetic observational index. There is a large spread across the ensemble, with the reference frequently lying close to the ensemble mean. The lower panel of
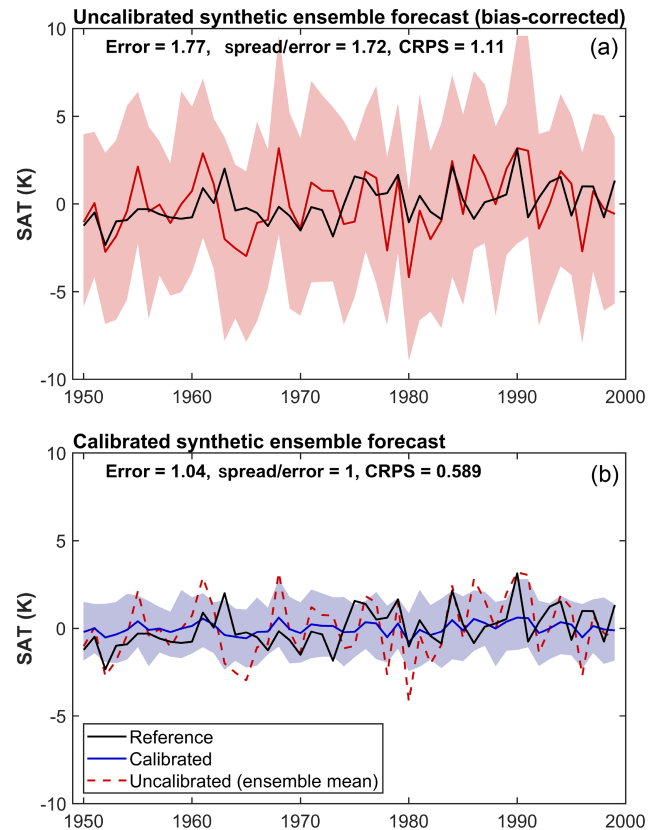

Figure 2. **(a)** Synthetic data for an example (bias-corrected) ensemble temperature evolution are shown for the ensemble mean in red and 90 % ensemble range (shaded), along with the synthetic (observational) reference index in black. **(b)** The synthetic ensemble calibrated using the variance inflation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

Fig. 2 shows the ensemble calibrated towards the reference data using the VINF method. The 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 Fig. 2. Also, it is important to note that the calibrated ensemble is perfectly reliable over the reference period, as indicated by the spread / error ratio 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 Fig. 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 out-of-sample future 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 reference period, 1920–2016. The future period from the CMIP5 realisation is then used to analyse the impact of the calibration by verifying against the 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 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 (some examples are shown in Fig. S8). Overall, the calibration 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 Fig. 3. In this example, the uncalibrated CESM1-LE data for CEUR summer temperature are shown in red, along with the same index from one of the CMIP5 models over the reference period (1920–2016). The model is calibrated 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, are used to verify the uncalibrated and calibrated large-ensemble projections using each individual year in the 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 models, we can therefore calculate verification statistics (i.e. RMSE, spread/error ratio, CRPS). The process is then repeated for each of the 39 CMIP5 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 CESM1-LE summer temperature for the uncalibrated ensemble and the three calibration methods are shown in Fig. 4. The distribution of the verification statistics over the 39 models is shown, with the hor-

izontal 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 RMSE 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 lower for all of the calibration methods in all regions, demonstrating that the calibrations are improving the probabilistic predictions of summer temperature by the CESM1-LE 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 Fig. 4, none of the calibration methods ever has a *significantly* negative impact on the projections.

We also performed the same testing described above for the CESM1-LE on the MPI-GE. The verification measures for the MPI-GE summer temperature are shown in Fig. S2. The performance of the calibration methods on the MPI-GE summer temperature is qualitatively similar to that for the CESM1-LE (shown in Fig. 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 CESM1-LE data (i.e. Fig. 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 CESM1-LE data, none of the calibration methods displays a 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. Figs. 4 and S2). Analysis of the equivalent figures for precipitation (Figs. S3 and S4), as well as for the winter season (not shown), also demonstrates a reasonably consistent performance between the calibration methods. The similarity of the performance of the calibration methods indicates 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
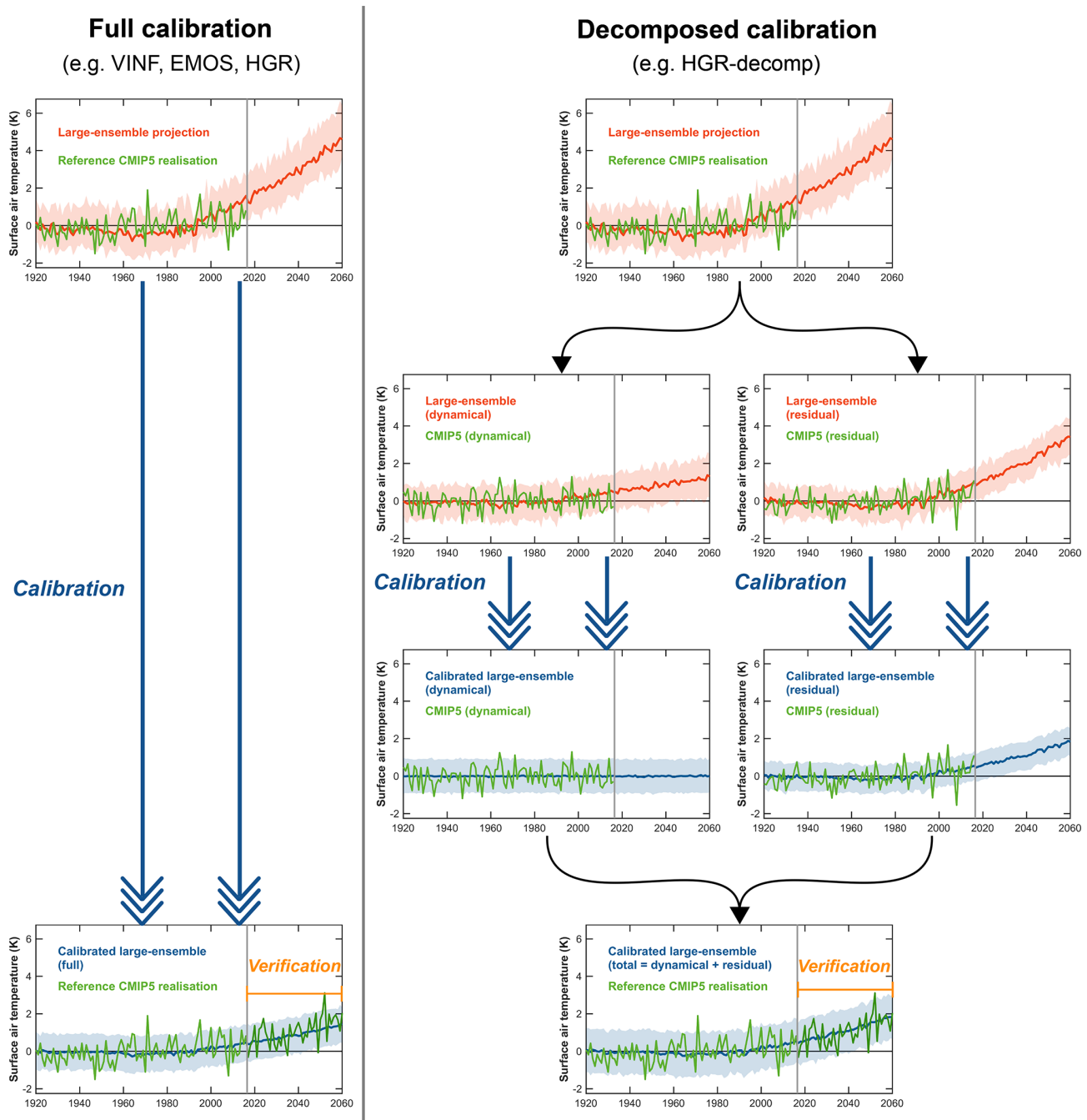
Figure 3. An example showing the steps of the full calibration methods (left-hand column) and calibration of the dynamically decomposed variables (right column). This example is the calibration of the summer (JJA) Central European (CEUR) temperature using the CESM1-LE and one of the CMIP5 models. The shading shows the 5–95 % range 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.

seasons for the calibrated CESM1-LE projections are shown in Fig. 5 (note that the red and orange data in the first column are the same as those shown in Fig. 4). The equivalent verification plot for the MPI-GE dataset is shown in Fig. S2, and the results are, generally, qualitatively similar to those shown for the CESM1-LE dataset in Fig. 5. The improvement for the winter temperature in terms of RMSE is not as clear as in the summer season, but the reliability of the ensemble projections is 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.
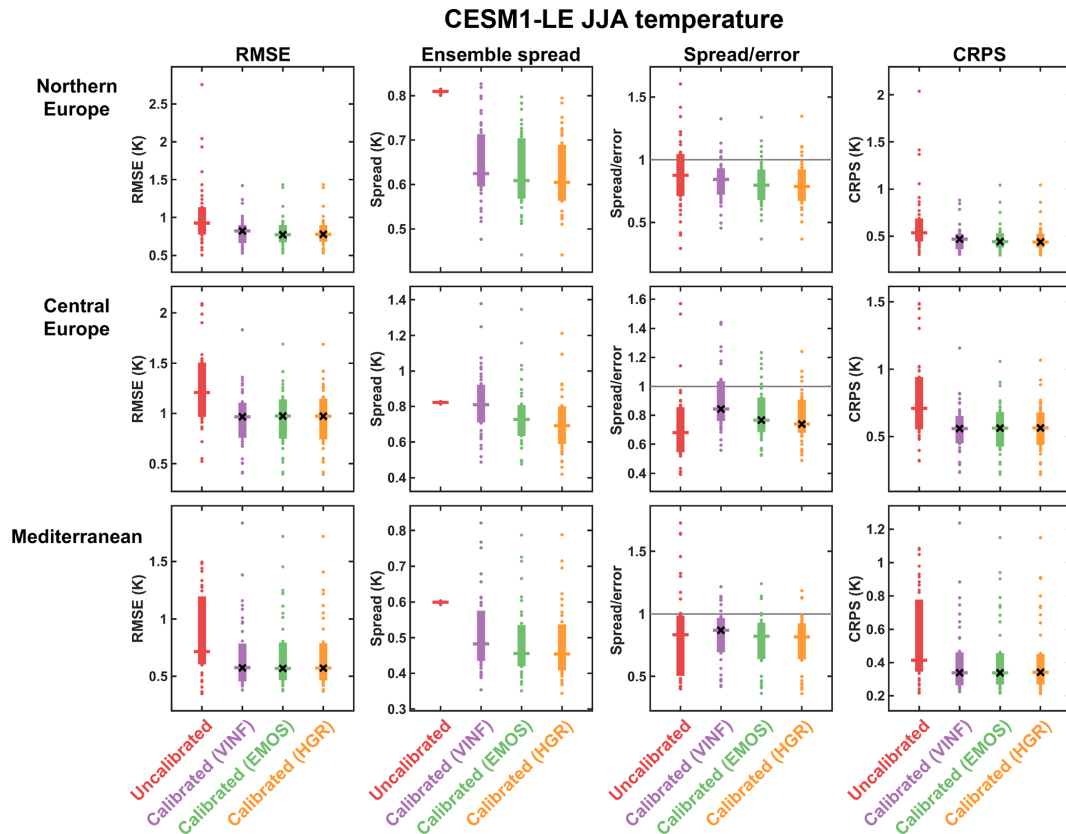
Figure 4. Comparison of calibration methods applied to the CESM1-LE summer temperature projections calibrated to the CMIP5 models over the observational period (1920–2016) and verified using the 44 years in the out-of-sample period (2017–2060). The verification statistics for each of the individual CMIP5 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 RMSE, spread / error ratio 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.

The spread of the uncalibrated 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 1; 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 ratio is not significantly negative for any of the variables, regions or seasons, indicating that the 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 RMSE for the NEUR winter precipitation in the CESM1-LE dataset (Fig. 5). Whilst it is only one of the verification measures performed across both the 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 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 Sect. 2.4) to split the model and observations datasets into a dynamical component, associated with
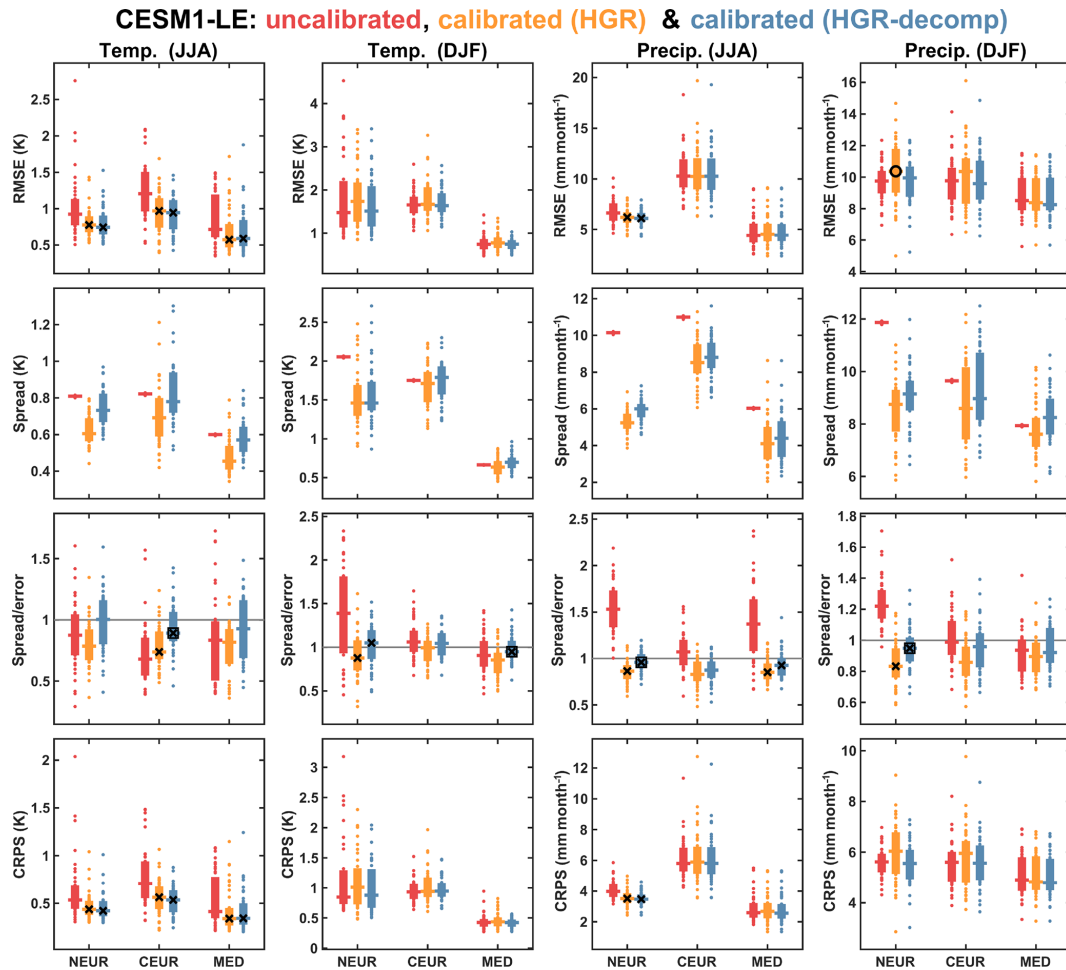
Figure 5. Overview of verification of the HGR and HGR-decomp calibration methods compared with the uncalibrated CESM1-LE data in the European regions. Results are shown for all of the verification measures, for both summer and winter seasons and for temperature and precipitation. The verification statistics 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 RMSE, spread / error ratio 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). Black boxes indicate where the HGR-decomp method of calibration is significantly better than the HGR method (at the 90 % 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.

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 CESM1-LE dataset, is shown in Fig. 3 (and also Fig. S1). In this example, the future temperature response is largely associated with the residual, representing the local thermodynamic response to increased 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) 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 projections is shown in Fig. 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 perform the calibration on the dynamically decomposed data, and we refer to this method as "HGR-decomp" hereafter.

Verification results for the HGR-decomp methods are shown for both temperature and precipitation and for all regions in Fig. 5, alongside the HGR verification results (with the equivalent verification for the MPI-GE shown in Fig. S5) As with the HGR verification, the crosses (circles) indicate where the verification statistics of the HGR-decomp-calibrated ensemble 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 RMSE 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 Fig. 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 ratio verification, where four of the regions/variables are significantly better for the HGR-decomp method applied to the 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 (Fig. S5), albeit with a lower level of significance. Specifically, in 10 of the 12 total regions/variables verified for the MPI-GE dataset, HGR-decomp calibration is found to be more reliable in terms of spread / error ratio than the HGR calibration.

Overall, the HGR-decomp method is found to be an improvement over the HGR method, and it 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 ratio. The spread / error ratio 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 separately has the effect of increasing the overall spread, likely because the method avoids fitting a forced thermodynamical or dynamical signal in the ensemble 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 ratio of the dynamical components of the ensemble is 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 1 and are less reliable than for the HGR-decomp method. In this sense, the HGR method appears to be slightly overfitting the ensemble to the reference period, resulting in a consistently overconfident 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 Figs. 4 and 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 perform a bootstrap resampling over the 39 CMIP5 model projection–verification pairs. Verification results for the 2041–2060 climatologies for both the CESM1-LE and MPI-GE are shown in Fig. 6.

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. RMSE) and reliability (i.e. spread / error ratio) of the out-of-sample verification. The calibrated summer temperature projections are more reliable in all three European regions in the CESM1-LE and MPI-GE ensembles but all tend to somewhat overconfident. The winter temperature shows less obvious improvement in terms of RMSE of the calibrated projections, but the reliability is significantly improved for all the regions in both ensembles, but, again, the calibrated CESM1-LE data are 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 RMSE, but the reliability is more mixed, with the calibration actually worsening the reliability in the MED region for the CESM1-LE dataset. The calibration has the least influence on the 2041–2060 climatology of precipitation, acting to worsen the RMSE in some instances but also to modestly improve the reliability.

Overall, the verification of the projected 2041–2060 climatologies in the imperfect model tests indicates 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 does not generally seem to perform as well as the calibration for the yearly pro-
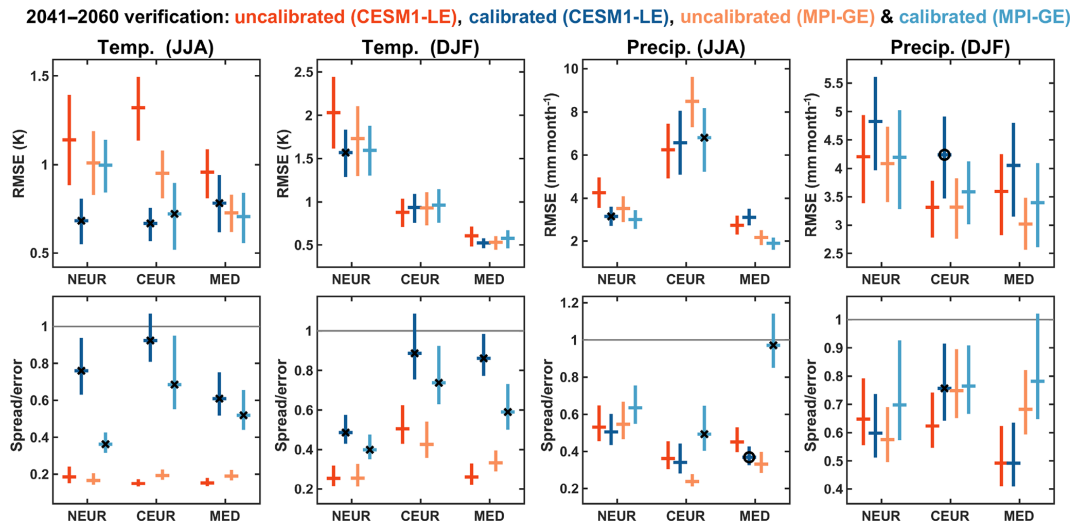
Figure 6. Verification of the 2041–2060 mean projections calculated relative to the out-of-sample CMIP5 models for both the CESM1-LE and MPI-GE datasets. The horizontal lines show the mean across all models and the vertical lines show the 90 % confidence intervals, calculated by randomly resampling across the CMIP5 models with replacement 1000 times. The black crosses indicate where the calibrated ensemble is significantly better than the equivalent uncalibrated ensemble; the black circles indicate where the calibrated ensemble is significantly worse that the uncalibrated ensemble.

jections examined in the 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, as the forced signal in the ensemble diverges from the observations to which it is calibrated. This is demonstrated clearly when the verification is applied to different future periods (specifically 2021–2040, 2041–2060 and 2061–2080; see Fig. S6). We find that the accuracy and reliability clearly deteriorate as the target period 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 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

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, us-

ing the observational indices of temperature and precipitation to calibrate against.

The calibrated CESM1-LE projections for the summer temperature and precipitation are shown in Fig. 7. Based on the imperfect model tests we expect the calibrated summer CESM1-LE to represent the most accurate and reliable projection out of all of the ensemble–variable–season combinations tested. For the summer temperature projections, calibrating the CESM1-LE against the observations over the reference period, the rate of warming until 2060 is reduced by varying amounts. There are also small changes in spread, perhaps most notable for the NEUR region, with the calibration method acting to increase the uncertainty in the future projections. For the precipitation, the signal in CESM1-LE projection is much weaker with respect to the inter-annual variability. In the projections shown here the calibration has a 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 CESM1-LE, with the ensemble medians of the MPI-GE also plotted in Fig. 7 for comparison (the full ensemble projection plots are shown in Fig. S7). The warming in the NEUR region is reduced over the 2017–2060 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 time-capturing the warming variability seen during the observational period. In the projections of summer precipitation in the MPI-GE dataset, there
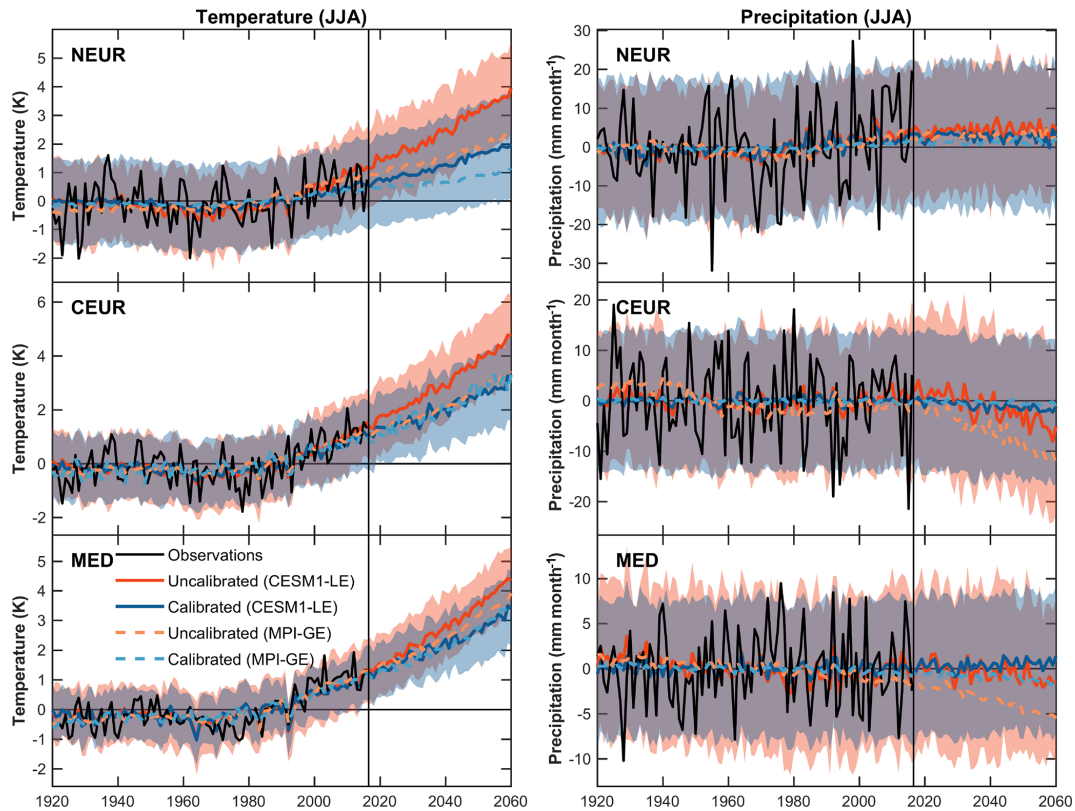
Figure 7. Uncalibrated and calibrated (HGR-decomp) CESM1-LE projections, where here the calibrated projections have been calibrated against the observations over the period 1920–2016. The lines show the ensemble medians for the uncalibrated and calibrated ensembles for both the CESM1-LE (solid) and MPI-GE (dashed) datasets. The shading shows the 5–95 % range of the 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.

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, 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 (Fig. 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 in the Supplement), and the decompositions are also substantially different (Table S2). This indicates that

for this season and region the 20CR data are not capturing what seems to be a clearer dynamical signal in the HadSLP2 dataset and, as a result, are 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 Eq. (7) are plotted in Fig. 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.

The projected changes in the 2041–2060 climatologies – compared with the present-day reference period, 1995–2014 – are shown in Fig. 8. Here, we have plotted both the uncalibrated and (HGR-decomp) calibrated climatological changes for both the CESM1-LE and MPI-GE datasets. An interesting feature of these projected changes is that for many of
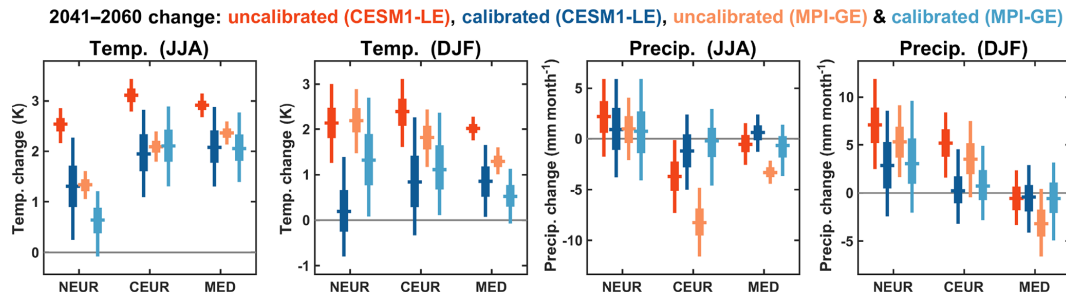
2041–2060 change: uncalibrated (CESM1-LE), calibrated (CESM1-LE), uncalibrated (MPI-GE) & calibrated (MPI-GE)

Figure 8. The 2041–2060 mean calculated relative to the 1995–2014 climatology for both CESM1-LE and 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.

them the calibrated ensembles are 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 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, which are themselves often very different.

Another feature of the calibration's 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. Fig. 6), 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 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 Fig. 6, the calibrated temperature projections for the CESM1-LE dataset are consistently more reliable (in terms of spread / error ratio) 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 CESM1-LE projections for future temperature changes.

In a recent paper, Brunner et al. (2020) 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 RCP8.5 scenario. The HGR-decomp method gener-

ally 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 overlaps. 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, Fig. 2 of Brunner et al., 2020). 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 – VINF, 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 methods act to scale the ensemble signal and spread so as to optimize the fit over the reference period. The three calibration methods 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 ensem-

ble. 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 overfitting 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 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 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 be expected to depend on the raw ensemble projection.

We then proceeded to apply the 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 40–50 years. In both the CESM1-LE and MPI-GE datasets the projected increase in European temperatures is generally smaller in the calibrated ensembles compared to the uncalibrated ensembles. The calibrated projections are 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) are projected to be about 2 °C for the 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., 2020).

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 overfitting or an overestimation of the added value of the constraint. The imperfect model approach we have used in this study is less susceptible to this type of overfitting as the data used for the verification 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 of 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. Doblas-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 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.

Author contributions. COR, DJB and AW conceived the study and planned the research. COR conducted the analysis. COR, DJB and AW discussed the results and wrote the paper.

Competing interests. The authors declare that they have no conflict of interest.

Review statement. This paper was edited by Yun Liu and reviewed by Francisco J. Doblas-Reyes, Philip Sansom, and two anonymous referees.

References

Allan, R. and Ansell, T.: A new globally complete monthly historical gridded mean sea level pressure dataset (HadSLP2): 1850–2004, J. 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, Geophys. Res. Lett., 45, 7808–7817, 2018.

Bröcker, J.: Evaluating raw ensembles with the continuous ranked probability score, Q. J. Roy. Meteor. Soc., 138, 1611–1617, 2012.

Brunner, L., Lorenz, R., Zumwald, M., and Knutti, R.: Quantifying uncertainty in European climate projections using combined performance-independence weighting, Environ. Res. Lett., 14, 124010, https://doi.org/10.1088/1748-9326/ab492f, 2019.

Brunner, L., McSweeney, C., Ballinger, A. P., Befort, D. J., Benassi, M., Booth, B., Coppola, E., de Vries, H., Harris, G., Hegerl, G. C., Knutti, R., Lenderink, G., Lowe, J., Nogherotto, R., O'Reilly, C., Qasmi, S., Ribes, A., Stocchi, P., and Undorf, S.: Comparing methods to constrain future European climate projections using a consistent framework, J. Climate, 33, 8671–8692, https://doi.org/10.1175/JCLI-D-19-0953.1, 2020.

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, Geophys. Res. Lett., 37, L20704, https://doi.org/10.1029/2010GL044613, 2010.

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., Gleason, B. E., Vose, R. S., Rutledge, G., Bessemoulin, P., Brönnimann, S., Brunet, M., Crouthamel, R. I., Grant, A. N., Groisman, P. Y., Jones, P. D., Kruk, M. C., Kruger, A. C., Marshall, G. J., Maugeri, M., Mok, H. Y., Nordli, Ø., Ross, T. F., Trigo, R. M., Wang, X. L., Woodruff, S. D., and Worley, S. J.: The twentieth century reanalysis project, Q. J. Roy. Meteorol. Soc., 137, 1–28, https://doi.org/10.1002/qj.776, 2011.

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, J. 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, J. Climate, 29, 2237–2258, 2016.

Deser, C., Lehner, F., Rodgers, K. B., Ault, T., Delworth, T. L., DiNezio, P. N., Fiore, A., Frankignoul, C., Fyfe, J. C., Horton, D. E., Kay, J. E., Knutti, R., Lovenduski, N. S., Marotzke, J., McKinnon, K. A., Minobe, S., Randerson, J., Screen, J. A., Simpson, I. R., and Ting, M.: Insights from Earth system model initial-condition large ensembles and future prospects, Nat. Clim. Change, 10, 277–286, https://doi.org/10.1038/s41558-020-0731-2, 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, Nat. Commun., 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, 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, Cambridge, 2012.

Fortin, V., Abaza, M., Anctil, F., and Turcotte, R.: Why should ensemble spread match the RMSE of the ensemble mean?, J. Hydrometeorol., 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, J. 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. Res. Lett., 30, 1629, https://doi.org/10.1029/2003GL017130, 2003.

Gleckler, P. J., Taylor, K. E., and Doutriaux, C.: Performance metrics for climate models, J. Geophys. Res.-Atmos., 113, D06104, https://doi.org/10.1029/2007JD008972, 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, Mon. Weather Rev., 133, 1098–1118, 2005.

Gneiting, T., Balabdaoui, F., and Raftery, A. E.: Probabilistic forecasts, calibration and sharpness, J. Roy. Stat. Soc. B, 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, Geophys. Res. Lett., 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, 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 Dataset, Int. J. Climatol., 34, 623–642, 2014.

Hawkins, E. and Sutton, R.: Connecting climate model projections of global temperature change with the real world, B. Am. Meteorol. Soc., 97, 963–980, 2016.

Hersbach, H.: Decomposition of the continuous ranked probability score for ensemble prediction systems, Weather Forecast., 15, 559–570, 2000.

Hewitt, C. D. and Lowe, J. A.: Toward a European Climate Prediction System, B. Am. Meteorol. Soc., 99, 1997–2001, 2018.

Jolliffe, I. T. and Stephenson, D. B.: Forecast verification: a practitioner's guide in atmospheric science, John Wiley and Sons, UK, 2012.

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., Arblaster, J. M., Bates, S., Danabasoglu, G., Edwards, J., Holland, M., Kushner, P., Lamarque, J.-F.,Lawrence, D., Lindsay, K., Middleton, A., Munoz, E., Neale, R., Oleson, K., Polvani, L., and Vertenstein, M.: The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability, B. Am. Meteorol. Soc., 96, 1333–1349, https://doi.org/10.1175/BAMS-D-13-00255.1, 2015.

Kettleborough, J., Booth, B., Stott, P., and Allen, M.: Estimates of uncertainty in predictions of global mean surface temperature, J. Climate, 20, 843–855, 2007.

Kharin, V. V. and Zwiers, F. W.: Improved seasonal probability forecasts, J. 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, Nat. Clim. Change, 3, 369–373, https://doi.org/10.1038/nclimate1716, 2013.

Knutti, R., Masson, D., and Gettelman, A.: Climate model genealogy: Generation CMIP5 and how we got there, Geophys. Res. Lett., 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, Geophys. Res. Lett., 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., Li, H., Manzini, E., Notz, D., Putrasahan, D., Boysen, L., Claussen, M., Ilyina, T., Olonscheck, D., Raddatz, T., Stevens, B., and Marotzke, J.: The Max Planck Institute grand ensemble-enabling the exploration of climate system variability, J. Adv. Model. Earth Syst., 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, Clim. Dynam., 53, 1287–1305, https://doi.org/10.1007/s00382-019-04640-4, 2019.

Matsueda, M., Weisheimer, A., and Palmer, T.: Calibrating climate change time-slice projections with estimates of seasonal forecast reliability, J. Climate, 29, 3831–3840, 2016.

McKinnon, K. A. and Deser, C.: Internal variability and regional climate trends in an observational large ensemble, J. Climate, 31, 6783–6802, 2018.

Merrifield, A. L., Brunner, L., Lorenz, R., Medhaug, I., and Knutti, R.: An investigation of weighting schemes suitable for incorporating large ensembles into multi-model ensembles, Earth Syst. Dynam., 11, 807–834, https://doi.org/10.5194/esd-11-807-2020, 2020.

O'Reilly, C. H., Woollings, T., and Zanna, L.: The dynamical influence of the Atlantic Multidecadal Oscillation on continental climate, J. Climate, 30, 7213–7230, 2017.

Pachauri, R. K., Allen, M. R., Barros, V. R., Broome, J., Cramer, W., Christ, R., Church, J. ., Clarke, L., Dahe, Q., Dasgupta, P., Dubash, N. K., Edenhofer, O., Elgizouli, I., Field, C. B., Forster, P., Friedlingstein, P., Fuglestvedt, J., Gomez-Echeverri, L., Hallegatte, S., Hegerl, G., Howden, M., Jiang, K., Jimenez Cisneroz, B., Kattsov, V., Lee, H., Mach, K. J., Marotzke, J., Mastrandrea, M. D., Meyer, L., Minx, J., Mulugetta, Y., O'Brien, K., Oppenheimer, M., Pereira, J. J., Pichs-Madruga, R., Plattner, G.-K., Pörtner, H.-O., Power, S. B., Preston, B., Ravindranath, N. H., Reisinger, A., Riahi, K., Rusticucci, M., Scholes, R., Seyboth, K., Sokona, Y., Stavins, R., Stocker, T. F., Tschakert, P., van Vuuren, D., and van Ypserle, J.-P.: 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, UN, Geneva, 2014.

Palmer, T. N., Alessandri, A., Andersen, U., Cantelaube, P., Davey, M., Delécluse, P., Déqué, M., Diez, E., Doblas-Reyes, F. J., Feddersen, H., Graham, R., Gualdi, S., Guérémy, J.-F., Hagedorn, R., Hoshen, M., Keenlyside, N., Latif, M., Lazar, A., Maisonnave, E., Marletto, V., Morse, A. P., Orfila, B., Rogel, P., Terres, J.-M., and Thomson, M. C.: Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER), B. Am. Meteorol. Soc., 85, 853–872, https://doi.org/10.1175/BAMS-85-6-853, 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), Geosci. Model Dev., 11, 351–368, https://doi.org/10.5194/gmd-11-351-2018, 2018.

Sanderson, B. M., Knutti, R., and Caldwell, P.: A representative democracy to reduce interdependency in a multimodel ensemble, J. 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, J. 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., Kharin, V., Kimoto, M., Merryfield, W. J., Mochizuki, T., Müller, W. A., Pohlmann, H., Yeager, S., and Yang, X.: Robust skill of decadal climate predictions, npj Clim. Atmos. Sci., 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., Alexander, L. V., Allen, S. K., Bindoff, N. L., Breon, F.-M., Church, J. A., Cubasch, U., Emori, S., Forster, P., Friedlingstein, P., Gillett, N., Gregory, J. M., Hartmann, D. L., Jansen, E., Kirtman, B., Knutti, R., Kumar Kanikicharla, K., Lemke, P., Marotzke, J., Masson-Delmotte, V., Meehl, G. A., Mokhov, I. I., Piao, S., Plattner, G.-K., Dahe, Q., Ramaswamy, V., Randall, D., Rhein, M., Rojas, M., Sabine, C., Shindell, D., Stocker, T. F., Talley, L. D., Vaughan, D. G., Xie, S.-P., Allen, M. R., Boucher, O., Chambers, D., Hesselbjerg Christensen, J., Ciais, P., Clark, P. U., Collins, M., Comiso, J. C., Vasconcellos de Menezes, V., Feely, R. A., Fichefet, T., Fiore, A. M., Flato, G., Fuglestvedt, J., Hegerl, G., Hezel, P. J., Johnson, G. C., Kaser, G., Kattsov, V., Kennedy, J., Klein Tank, A. M. G., Le Quere, C., Myhre, G., Osborn, T., Payne, A. J., Perlwitz, J., Power, S., Prather, M., Rintoul, S. R., Rogelj, J., Rusticucci, M., Schulz, M., Sedlacek, J., Stott, P. A., Sutton, R., Thorne, P. W., and Wuebbles, D.: Climate change 2013: The physical science basis, UN, Geneva, 2013.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An overview of CMIP5 and the experiment design, B. Am. Meteorol. Soc., 93, 485–498, 2012.

Tippett, M. K. and Barnston, A. G.: Skill of multimodel ENSO probability forecasts, Mon. Weather Rev., 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, P. Natl. Acad. Sci. USA, 109, 14337–14342, 2012.

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

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

Yeager, S., Danabasoglu, G., Rosenbloom, N., Strand, W., Bates, S., Meehl, G., Karspeck, A., Lindsay, K., Long, M., Teng, H., and Lovenduski, N. S.: Predicting near-term: changes in the Earth System: A large ensemble of initialized decadal prediction simulations using the Community Earth System Model, B. Am. Meteorol. Soc., 99, 1867–1886, 2018.