https://doi.org/10.5194/npg-2021-12
Preprint. Discussion started: 19 March 2021

Nonlinear Processes in Geophysics
Discussions
Open Access
EGU

Improving the Potential Accuracy and Usability of EURO-CORDEX Estimates of Future Rainfall Climate using Mean Squared Error Model Averaging

Stephen Jewson[1], Giuliana Barbato[2], Paola Mercogliano[2], Jaroslav Mysiak[2], Maximiliano Sassi[3]

[1]Independent Researcher, London, UK

[2]Euro-Mediterranean Center on Climate Change (CMCC) Foundation, Via Augusto Imperatore, 16, 73100, Lecce, Italy

[3]Risk Management Solutions Ltd, EC3R 7AG, London, UK

Correspondence to: Stephen Jewson (stephen.jewson@gmail.com)

Abstract. Probabilities of future climate states can be estimated by fitting distributions to the members of an ensemble of climate model projections. The change in the ensemble mean can be used as an estimate of the unknown change in the mean of the distribution of the climate variable being predicted. However, the level of sampling uncertainty around the change in the ensemble mean varies from case to case and in some cases is large. We compare two model averaging methods that take the uncertainty in the change in the ensemble mean into account in the distribution fitting process. They both involve fitting distributions to the ensemble using an uncertainty-adjusted value for the ensemble mean in an attempt to increase predictive skill relative to using the unadjusted ensemble mean. We use the two methods to make projections of future rainfall based on a large dataset of high resolution EURO-CORDEX simulations for different seasons, rainfall variables, RCPs and points in time. Cross-validation within the ensemble using both point and probabilistic validation methods shows that in most cases predictions based on the adjusted ensemble means show higher potential accuracy than those based on the unadjusted ensemble mean. They also perform better than predictions based on conventional Akaike model averaging and statistical testing. The adjustments to the ensemble mean vary continuously between situations that are statistically significant and those that are not. Of the two methods we test, one is very simple, and the other is more complex and involves averaging using a Bayesian posterior. The simpler method performs nearly as well as the more complex method.

1 Introduction

Estimates of future climate are often created using climate projection ensembles. Examples of such ensembles include the CMIP5 project (Taylor, et al., 2012), the CMIP6 project (Eyring, et al., 2016) and the EURO-CORDEX project (Jacob, Petersen, & authors, 2014). If required, distributions can then be fitted to these ensembles to produce probabilistic predictions. The probabilities in these predictions are conditional probabilities and depend on the assumptions behind the climate model projections, such as the choice of RCP (Moss, et al., 2010; Meinshausen, et al., 2011), and the choice of models and model resolution. Converting climate projection ensembles to probabilities in this way is helpful for those applications of climate projections for which the impact models can ingest probabilities more easily than they can ingest individual ensemble members. An example would be the catastrophe models used in the insurance industry, which quantify climate risk using simulated natural catastrophes embedded in many tens of thousands of simulated versions of one year (Kaczmarska, et al. (2018), Sassi, et al. (2019)). Methodologies have been developed by which these catastrophe model ensembles can be adjusted to include climate change, based on probabilities derived from climate projections (Jewson, et al., 2019).

We will consider the case in which distributions are fitted to changes in climate model output, rather than to absolute values. When fitting distributions to changes in climate model output, the change in the ensemble mean can be used as an estimate of the unknown change in the mean of the distribution of the real future climate. Model weights or bias corrections may be included at this point (Sanderson, et al. (2015b), Knutti, et al. (2017), Chen, et al, (2019)). However, because climate model ensembles are finite in size, the ensemble mean suffers from estimation uncertainty. A number of studies have investigated the various uncertainties in climate ensembles, including estimation uncertainty (such as Deser, et al. (2010), Thompson, et al. (2015)) and various methods have been developed for the post-processing of ensembles to help understand these uncertainties and take them into account, addressing issues such as how to break the uncertainty into components (Hawkins and Sutton, (2009), Yip, et al. (2011), Hingray and Said (2014)), how to identify forced signals given the uncertainty (Frankcombe, et al. (2015), Sippel, et al. (2019), Barnes, et al. (2019) and Wills, et al. (2020)), and how quickly signals emerge from the noise given the uncertainty (Hawkins and Sutton (2012), Lehner, et al. (2017)).

In this article, we explore some of the implications of estimation uncertainty around the ensemble mean in more detail. Ensemble mean estimation uncertainty varies by season, variable, projection, time and location. In the worst cases, the uncertainty may be larger than the change in the ensemble mean itself, and this makes the change in the ensemble mean, and distributions that have been fitted to the ensemble, potentially misleading. In these large uncertainty cases the change in the ensemble mean is dominated by the randomness of internal variability from the individual ensemble members, and it would be unfortunate if this randomness was allowed to influence adaptation decisions. A standard approach used to manage varying uncertainty in the change in the ensemble mean is to consider statistical significance of the changes (e.g., see shading of regions of statistical significance in climate reports such as the EEA report (European Environment Agency, 2017) or the IPCC 2014 report (Pachauri & Meyer, 2014)). Statistical significance testing involves calculating the signal-to-noise ratio (SNR) of the change in the ensemble mean, where the signal is the ensemble mean change, and the noise is the standard error of the ensemble mean. The SNR is then compared with a threshold value. If the SNR is greater than the threshold then the signal is declared statistically significant (Wilks, 2011).

Use of statistical significance to filter climate projections in this way is appropriate for scientific discovery. However, it is perhaps less appropriate in some practical applications of climate model projections. There are a number of reasons for this, which can be illustrated by considering the shortcomings of a system which applies statistical testing and sets regions of non-significant change in the ensemble mean to zero. The first problem with such a system is that analysis of the properties of predictions made using statistical testing show that they have poor predictive skill. This is not surprising, since statistical testing was never designed as a methodology for creating predictions. The second problem is that statistical testing creates abrupt jumps of the climate change signal in space, between significant and non-significant regions, and between different RCPs and time points. These jumps are artefacts of the use of a method with a threshold, rather than an aspect of the climate change signal itself. This may lead to situations in which one location is reported to be affected by climate change, and an adjacent location not, simply because the significance level has shifted from e.g., 95.1% to 94.9%. From a practical perspective this may undermine the credibility of climate predictions in the perception of users, to whom no reasonable physical explanation can be given for such features of the projections. Finally, the almost universal use of a threshold p-value of 95% creates a skew towards avoiding false positives (type I errors) at the expense of false negatives (type II errors). Depending on the application, this may not be appropriate. Reducing false negatives in this way is particularly a problem for risk modelling, since risk models should attempt to capture all possibilities in some way, even if low significance.

How, then, should those who wish to make practical application of climate model ensembles deal with the issue of varying uncertainty in the change in the ensemble mean, as captured by spatial variation of the SNR, in cases where the uncertainty is large but statistical testing is not appropriate? We describe and compare two model averaging procedures as possible answers to this question. The procedures work by using a bias-variance trade-off argument to reduce the change captured by the ensemble mean when it is uncertain. They are based on standard statistical ideas related to parameter shrinkage as a way of improving prediction performance when fitting distributions (see, e.g., Copas (1983)). We will call the methods Mean-squared-error Model Averaging (MMA). The averaging in MMA consists of averaging of the usual estimate for the mean with an alternative estimate of the change in the mean of zero. The averaging weights in MMA depend on the SNR and are designed to minimise the predictive root-mean-squared-error (PRMSE) of the adjusted ensemble mean. One of the two MMA procedures we describe uses a simple plug-in estimator, and we refer to this method as Simple MMA (SMMA). The other procedure involves integration over a Bayesian posterior, and we refer to this method as Bayesian MMA (BMMA). In regions where the SNR is large these methods make no material difference. In regions where the SNR is small, the changes in the ensemble mean are reduced by MMA, in accordance with the theory we present below, in such a way as to increase the accuracy of the predictions. The MMA methods can be considered as continuous analogues of statistical testing, in which rather than setting the change in the ensemble mean to either 100% or 0% of the original value, we allow a continuous reduction that can take any value between 100% and 0% depending on the SNR. As a result, both methods avoid the abrupt jumps created by statistical testing.

We illustrate and test the SMMA and BMMA methods using a large dataset of high-resolution EURO-CORDEX ensemble projections of rainfall over Europe. We consider for four seasons, three rainfall variables, two RCPs and three future time periods, giving 72 cases in all. In section 2 we describe the EURO-CORDEX data we will use. In section 3 we describe the MMA procedures, and present some results based on simulated data which elucidate the situations in which MMA is likely to perform well versus other methods, for both point and probabilistic predictions. In section 4 we present results for one of the 72 cases in detail. We use cross-validation within the ensemble to evaluate the potential prediction skill of MMA, again for both point and probabilistic predictions, and compare with the skill from using the unadjusted ensemble mean, statistical testing and a conventional model averaging scheme (small-sample Akaike Information Criterion model averaging (AICc) (Burnham & Anderson, 2010)). In section 5 we present aggregate results for all 72 cases using the same methods. In section 6 we summarize and conclude.

2 Data and Methodology

The data we use for our study is extracted from the data archive produced by the EURO-CORDEX program (Jacob, Petersen, & authors, 2014; Jacob, Teichmann, & authors, 2020), in which a number of different global climate model simulations were downscaled over Europe using regional models at 0.11-degree resolution (about 12km). We use data from 10 models, each of which is a different combination of a global climate model and a regional climate model. The models are listed in Table 1. Further information on EURO-CORDEX and the models is given in the guidance report (Benestad, et al., 2017).

Model	Driving GCM	GCM Member	RCM
M1	CNRM-CM5	r1i1p1	ALADIN53
M2	IPSL-CM5A-MR	r1i1p1	RCA4
M3	CNRM-CM5	r1i1p1	RCA4
M4	CNRM-CM5	r1i1p1	CCLM4-8-17
M5	EC-EARTH	r12i1p1	CCLM4-8-17
M6	EC-EARTH	r12i1p1	RACMO22E
M7	EC-EARTH	r12i1p1	RCA4
M8	EC-EARTH	r1i1p1	RACMO22E
M9	EC-EARTH	r3i1p1	HIRHAM5
M10	IPSL-CM5A-MR	r1i1p1	WRF331F

Table 1: Models used in this study

We extract data for four meteorological seasons (DJF, MAM, JJA, SON), for three aspects of rainfall: changes in the total rainfall (RTOT), the 95th percentile of daily rainfall (R95) and the 99th percentile of daily rainfall (R99). We say ‘rainfall’ even though in some locations we may be including other kinds of precipitation. We consider two RCPs, RCP4.5 and RCP8.5, and four 30-year time-periods: 1981-2010, which serves as a baseline from which changes are calculated, and the three target periods of 2011-2040, 2041-2070 and 2071-2100. In total this gives 72 different cases (four seasons, three variables, two RCPs and three target time periods).

Figure 1 illustrates one of the 72 cases: changes in winter (DJF) values for RTOT, from RCP4.5, for the years 2011-2040. This example was chosen as the first in the database, rather than for any particular properties it may possess. Figure 1a shows the ensemble mean change $\hat{\mu}_c$ (the mean change calculated from the 10 models in the ensemble) and Fig. 1b shows the standard deviation of the change $\hat{\sigma}_c$ (the standard deviation of the changes calculated from the 10 models in the ensemble). Fig. 1c shows the estimated SNR $\hat{s}$ calculated from the ensemble mean change and the standard deviation of change using the expression $\hat{s} = n^{1/2}|\hat{\mu}_c| / \hat{\sigma}_c$, where the $n^{1/2}$ term in this equation converts the standard deviation of change (a measure of the spread of the changes across the ensemble) to the standard error of the ensemble mean change (a measure of the uncertainty around the ensemble mean change). Finally, Fig. 1d shows the regions in which the changes in the ensemble mean are significant at the 95% level, assuming normally distributed changes. We see that the ensemble mean change varies considerably in space, with notable increases in RTOT in the west of Ireland, the west of Great Britain, and in parts of France, Germany, Spain and Portugal. The standard deviation of change also varies considerably with the largest values over Portugal, parts of Spain and the Alps. The SNR shows that many of the changes in the West of Ireland, and in France and Germany, have particularly high SNRs, while the changes in Southern Europe (Portugal, Spain, Italy and Greece) have lower values. Accordingly, the changes are statistically significant throughout most of Ireland, Great Britain, France, Germany, and Eastern Europe, but are mostly not statistically significant in Southern Europe. The other 71 cases show similar levels of variability of these four fields, but with different spatial patterns.

https://doi.org/10.5194/npg-2021-12
Preprint. Discussion started: 19 March 2021

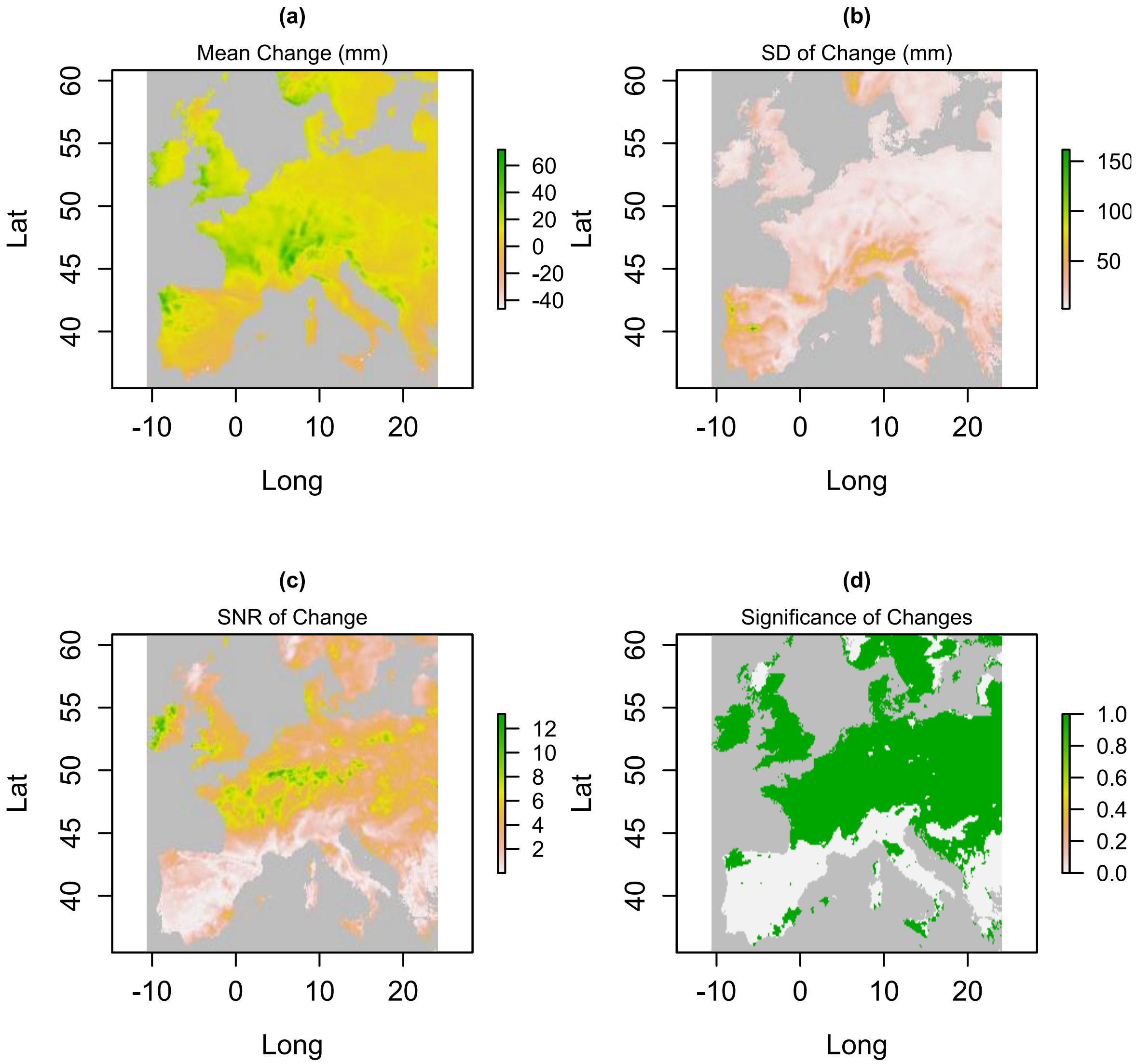


Figure 1: EURO-CORDEX projections for winter, for the change in total precipitation (RTOT) between the period 2011-2040 and the baseline 1981-2010, for RCP4.5. Panel (a) shows the ensemble mean, panel (b) shows the ensemble standard deviation, panel (c) shows the signal-to-noise ratio (SNR) and panel (d) shows the regions in which the changes in the mean are significant at the 95% level (shaded in green).

Figure 2 shows spatial mean values of the SNR (where the spatial mean is over the entire domain shown in Fig. 1) for all 72 cases. Each black circle is a spatial mean value of the SNR for one case, and each of the four panels in Fig. 2 shows the same 72 black circles but divided into sub-categories in different ways. The horizontal lines are the averages over the black circles

https://doi.org/10.5194/npg-2021-12
Preprint. Discussion started: 19 March 2021

in each sub-category. Figure 2a sub-divides by season: we see that there is a clear gradient from winter (DJF), which shows the highest values of the spatial mean SNR, to autumn (SON) which shows the lowest values of spatial mean SNR. Fig. 2b sub-divides by rainfall variable: in this case there is no obvious impact on the SNR values. Fig. 2c sub-divides by RCP. RCP8.5 shows higher SNR values, as we might expect, since in the later years RCP8.5 is based on larger changes in external forcing. Fig. 2d sub-divides by time-period: there is a strong gradient in SNR from the first of the three time-periods to the last. This is also as expected since both RCP scenarios are based on increasing external forcing with time. We would expect these varying SNRs to influence the results from the MMA methods. This will be explored in the results we present below.

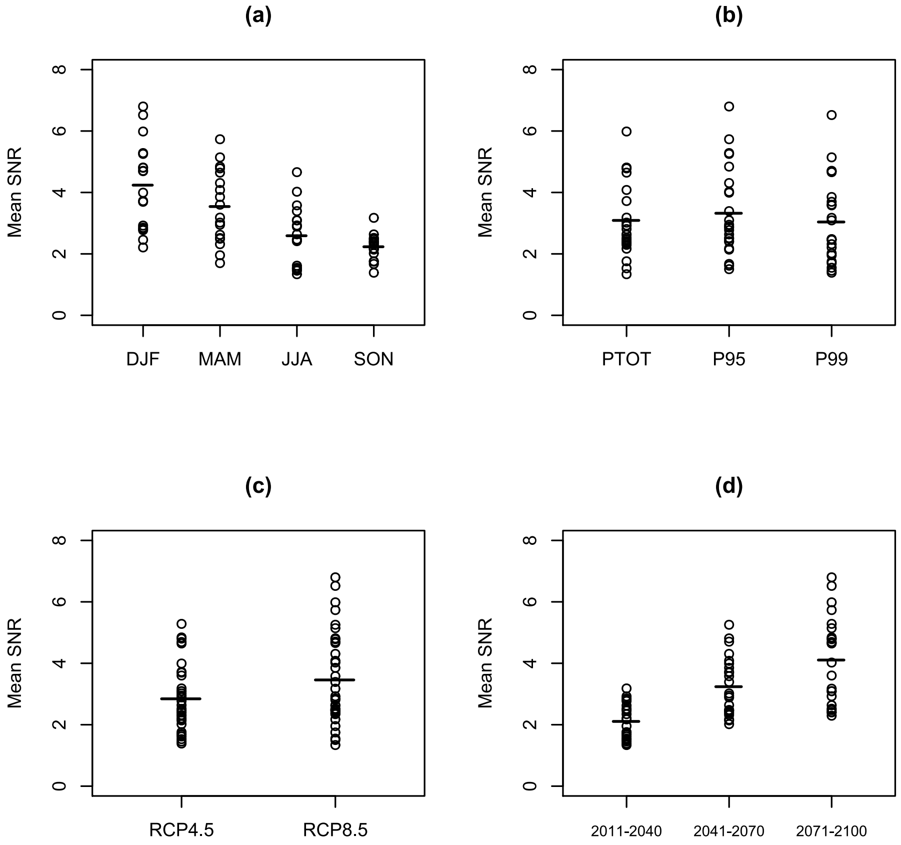


Figure 2: Each panel shows 72 values of the spatial average SNR value (black circles) derived from each of the 72 EURO-CORDEX climate change projections described in the text, along with means within each subset (horizontal lines). Panel (a) shows the 72 values as a function of season, panel (b) shows them as a function of rainfall variable, panel (c) shows them as a function of RCP and panel (d) shows them as a function of time period.

https://doi.org/10.5194/npg-2021-12
Preprint. Discussion started: 19 March 2021

3 Model Averaging Methodologies

The model averaging methodologies we apply are based on standard bias-variance trade-off arguments and are used to average together uncertain projections of change with projections of no change, in such a way as to try and improve predictive skill. The derivations of the methods follow standard mathematical arguments and proceed as follows.

3.1 Assumptions

For each location within each of the 72 cases, we first make some assumptions about the variability of the climate model results, the variability of future reality, and the relationship between the climate model ensemble and future reality. All quantities are considered as changes from the 1981-2010 baseline. We assume that the actual future value is a sample from a distribution with unknown mean μ and variance σ^2. We assume that the climate model values are independent samples from a distribution with unknown mean μ_c and variance ${\sigma_c}^2$. For the BMMA method we will additionally assume that these distributions are normal distributions. With regards to the assumption of independence of samples, this is an approximation, since the models are not entirely independent. Issues related to model dependence and independence have been discussed in, for example, Knutti, et al. (2010) and DelSole, et al. (2013). We assume that methods to address model dependence have been applied before applying MMA. In terms of how the climate models and reality relate to each other, we assume that the climate model ensemble is realistic in the sense that the mean and variance parameters agree, and so $\mu_c = \mu$ and ${\sigma_c}^2 = \sigma^2$. This is a "perfect ensemble" assumption: the models themselves are not perfect, but the distribution they are sampled from perfectly accounts for their biases. This is not likely to be strictly correct, and real climate model ensembles do contain errors and biases, but is a useful working assumption. We will write the future climate state as y, and the ensemble mean, estimated in the usual way from the ensemble, as $\hat{\mu}_c$. Since the usual estimator for the mean is unbiased, we can then say:

$$E(\hat{\mu}_c) = \mu_c = \mu = E(y) \tag{1}$$

If we write the ensemble variance, estimated using the usual unbiased estimator, as ${\hat{\sigma}_c}^2$, then we can say:

$$E\left({\hat{\sigma}_c}^2\right) = {\sigma_c}^2 = \sigma^2 = V(y) \tag{2}$$

Uncertainty around the estimate of the ensemble mean is given by:

$$V(\hat{\mu}_c) = \frac{{\sigma_c}^2}{n} = \frac{\sigma^2}{n} \approx \frac{{\hat{\sigma}_c}^2}{n} \tag{3}$$

3.2 The Simple Mean-squared-error Model Averaging (SMMA) Methodology

In the SMMA method we make a new prediction of future climate in which we adjust the ensemble mean change using a multiplicative factor k. k is an averaging weight such that the weight on the ensemble mean is k and the weight on a change of zero is $1 - k$ .We write the new prediction $\hat{y}$ as:

$$\hat{y} = k\,\hat{\mu}_c \tag{4}$$

where the factor k, for which we derive an expression below, varies from 0 to 1 as a function of all the parameters of the prediction: season, variable, RCP, time-period and spatial location. The intuitive idea behind this prediction is that if for one particular set of prediction parameters the SNR in the ensemble is large, and the ensemble mean prediction $\hat{\mu}_c$ is statistically significant, then it makes sense to use the ensemble mean more or less as is, and k should be close to 1. On the other hand, if the SNR in the change in the ensemble mean is small, and the change in the ensemble mean is far from statistically significant, then perhaps it is better to use a k value closer to zero. Statistical testing sets k to either 1 or 0 depending on whether the change is significant or not: the SMMA method (and the BMMA method described later) allow it to vary continuously from 1 to 0.

The ensemble mean is the unique value that minimises MSE within the ensemble. However, when considering applications of ensembles, it is generally more appropriate to consider out of sample, or predictive, MSE (PMSE). We can calculate the statistical properties of the prediction errors for the prediction $\hat{y}$, and the PMSE, as follows:

$$\text{prediction error} = e = y - \hat{y} = y - k\,\hat{\mu}_c \tag{5}$$

$$\text{bias} = E(e) = E(y - k\,\hat{\mu}_c) = E(y) - E(k\,\hat{\mu}_c) = \mu - k\mu = \mu(1-k) \tag{6}$$

$$\text{error variance} = V(e) = V(y - k\,\hat{\mu}_c) = V(y) + V(k\,\hat{\mu}_c) = \sigma^2 + k^2\frac{\sigma^2}{n} = \sigma^2\left(1 + \frac{k^2}{n}\right) \tag{7}$$

$$\text{PMSE} = E(y - k\,\hat{\mu}_c\,)^2 = E\left(y^2 - 2\,k\,y\,\hat{\mu}_c + k^2\,\hat{\mu}_c{}^2\right) = \mu^2 + \sigma^2 - 2k\,\mu^2 + k^2\left(\mu^2 + \frac{\sigma^2}{n}\right)$$

$$= \mu^2(1-k)^2 + \sigma^2\left(1 + \frac{k^2}{n}\right) = \text{bias}^2 + \text{error variance} \tag{8}$$

From the above equations we see that for $k = 0$ the bias of the prediction $\hat{y}$ is μ and the variance is σ^2, giving a PMSE of $\mu^2 + \sigma^2$. For $k = 1$ the bias is 0 and the variance is $\sigma^2\left(1 + \frac{1}{n}\right)$, giving a PMSE equal to the variance. We now seek to find the value of k that minimizes the PMSE. The derivative of the PMSE with respect to k is given by

$$\frac{d\text{PMSE}}{dk} = 2k\left(\mu^2 + \frac{\sigma^2}{n}\right) - 2\mu^2 \tag{9}$$

From this we find that the PMSE has a minimum at

$$k = \frac{\mu^2}{\mu^2 + \frac{\sigma^2}{n}} = \frac{1}{1 + \frac{\sigma^2}{n\mu^2}} = \frac{1}{1 + \frac{1}{s^2}} \tag{10}$$

where s is the SNR $s = n^{1/2}|\mu|/\sigma$. Equation (10) shows that the value of k at the minimum always lies in the interval [0,1]. We see from the above derivation that there is a value of k between 0 and 1 which gives a lower PMSE than either the prediction for no change ($k = 0$) or the unadjusted ensemble mean ($k = 1$) . Relative to the ensemble mean, the prediction based on this optimal value of k has a higher bias, but a lower variance, which is why we refer to it as a bias-variance trade-off: in the expression for PMSE we have increased the bias squared term, in return for a bigger reduction in the variance term. The PMSE of this prediction is lower than the PMSE of the prediction based on the ensemble mean because of the reduction in the term $\frac{\sigma^2 k^2}{n}$, which represents the contribution to PMSE of the estimation error of the ensemble mean. For an infinite size ensemble, this term would be zero, and the optimal value of k would be 1. We can therefore see the prediction $\hat{y}$

as a small-sample correction to the ensemble mean, which compensates for the fact that the ensemble mean is partly affected by the variability across a finite ensemble.

If we could determine the optimal value of k then we could, without fail, produce predictions that would have a lower PMSE than the ensemble mean. However, the expression for k given above depends on two unknown quantities, μ^2 and σ^2, and the best we can do is to attempt to estimate k based on the information we have. The most obvious estimator is that formed by simply plugging-in the observed equivalents of μ^2 and σ^2, calculated from the ensemble, which are $\hat{\mu}_c{}^2$ and $\hat{\sigma}_c{}^2$, giving the plug-in estimate for k:

$$\hat{k}_S = \frac{\hat{\mu}_c{}^2}{\hat{\mu}_c{}^2 + \frac{\hat{\sigma}_c{}^2}{n}} = \frac{1}{1+\frac{1}{\hat{s}^2}} \qquad (11)$$

This is the estimate of k that we will use in the SMMA method.

3.2.1 Relation to Statistical Significance

We can relate the value of $\hat{k}_S$ to the threshold for statistical significance, since statistical testing for changes in the mean of a normal distribution also uses the observed SNR, in which context it is known as the t-statistic. For a sample of size 10, two-tail significance at the 95% confidence level is achieved by signals with a SNR value of 2.262 or greater. This corresponds to a $\hat{k}_S$ value of 0.837. All situations with a $\hat{k}_S$ value greater than this are therefore statistically significant at the 95% level.

3.2.2 Generation of Probabilistic Predictions

Applying SMMA to a climate projection adjusts the mean. By making an assumption about the shape of the distribution of uncertainty, we can also derive a corresponding probabilistic forecast, as follows. We will assume that the distribution of uncertainty, for given values of the estimated mean and variance $\hat{\mu}_c$ and $\hat{\sigma}_c{}^2$, is a normal distribution. For the unadjusted ensemble mean, an appropriate predictive distribution can be derived using standard Bayesian methods, which widen and change the predictive distribution so as to take account of parameter uncertainty on the estimates of $\hat{\mu}_c$ and $\hat{\sigma}_c{}^2$. Bayesian methods require priors, and sometimes the choice of prior is difficult, but the normal distribution is one of the few statistical models that have a unique objective prior that is relevant in the context of making predictions (see, for example, Lee (1997)). This prior has a number of attractive properties, including that the resulting predictions match with confidence limits. The predictions based on this prior are t distributions, in which the location parameter is given by the usual estimate for the mean, $\hat{\mu}_c$, the square of the scale parameter (which is not the variance for the t distribution) is given by the usual unbiased estimate for the variance, $\hat{\sigma}_c{}^2$, and the number of degrees of freedom is the ensemble size minus 1. This formulation gives us probabilistic predictions based on the unadjusted ensemble mean. We then modify it to create probabilistic predictions based on the MMA-adjusted ensemble means: the distribution remains a t distribution, the location parameter is given by the

MMA-adjusted mean, the scale parameter is given by the PMSE from Eq. (8), and the numbers of degrees of freedom are again given by the ensemble size minus 1.

3.3 Bayesian Mean-squared-error Model Averaging (BMMA) Methodology

The BMMA method is derived as an extension of the SMMA method as follows. Since the prediction in the SMMA method $\hat{y}$ depends on $\hat{k}_S$ and $\hat{\mu}_c$, and $\hat{k}_S$ depends on $\hat{\mu}_c$ and $\hat{\sigma}_c$, we see that the prediction is affected by parameter estimation uncertainty on $\hat{\mu}_c$ and $\hat{\sigma}_c$. Since we only have 10 ensemble members with which to estimate these parameters, this uncertainty is large. Different values of $\hat{\mu}_c$ and $\hat{\sigma}_c$ from within the range of parameter estimation uncertainty would lead to different values of $\hat{y}$. We take a Bayesian approach to managing this parameter uncertainty and calculate the average value of $\hat{y}$ across all possible parameter values weighted by their probability densities $p(\hat{\mu}_c, \hat{\sigma}_c)$. To calculate the probability densities we use the same objective prior as was used in Sect. 3.2.2 above. To calculate the average we simulate 250 independent pairs of values $\hat{\mu}_c$ and $\hat{\sigma}_c$ for each location for each case, calculate 250 values of $\hat{y}$, and average the 250 values of $\hat{y}$ together to create our final prediction, which we write as $\hat{y}_B$. For purposes of comparison with the SMMA method we can then reverse-engineer an effective value of k, given by $\hat{k}_B = \hat{y}_B/\hat{\mu}_c$.

3.4 Simulation results

Given that $\hat{k}_S$ and $\hat{k}_B$ are only estimated, there is no guarantee that the predictions from the SMMA and BMMA methods will actually have a lower PMSE than the ensemble mean, in spite of the derivation which implies that they should. Before we test SMMA and BMMA on actual climate model output, we can explore whether they may or may not give better predictions using simulations, as follows. We vary a SNR parameter from 0 to 7, in 100 steps. For each value, we simulate 1 million synthetic ensembles, each of 10 points from a normal distribution. For each ensemble we apply the two MMA methodologies, statistical testing and conventional AICc model averaging and compare the resulting predictions with the underlying known mean, which we know in this case because these are ensembles we have generated ourselves. We then calculate the PRMSE of each method relative to the PRMSE of the ensemble mean. Results are shown in Fig. 3a. The horizontal line shows the performance of the unadjusted ensemble mean, which is constant with SNR, and which is determined simply by the variance of the variable being predicted and the parameter uncertainty on the ensemble mean. The red dashed line shows the performance of the SMMA method. We see that it does better than the ensemble mean for small values of SNR, up to around 1.45, and worse thereafter. For large values of the SNR its performance asymptotes to that of the ensemble mean. The worst performance is for values of SNR of around 2.5. The blue dotted line shows the performance of the BMMA method. It shows a similar pattern of behaviour to the SMMA method: it does better than the ensemble mean for small values of SNR, now up to around 1.9, and worse thereafter. For both small and large SNR values it performs worse than the SMMA method, while for a range of intermediate values it performs better. The purple dot-dashed line shows the performance of statistical testing, which gives the best predictions for the smallest values of SNR, but the poorest predictions over a large range of intermediate SNR values. This poor predictive performance is related to the use of a high threshold that has to be crossed before the ensemble mean is used. The green long-dashed line shows the performance of AICc model averaging, which shows results in between statistical testing and the MMA methods. Comparing the four methods, we see there is a trade-off whereby those methods that perform best for small and large SNR values perform the least well for intermediate values. The spatial average performance on a real data-set will then depend on the range of SNR values in that dataset. Although this graph gives us insight into the performance of the various methods, and suggests that, depending on the range of actual SNR values, they may all perform better than the ensemble mean in some cases, it cannot be used as a look-up table to determine which of the methods to use. This is because the results are shown as a function of the actual SNR value (as opposed to the estimated SNR value), and in real cases this actual value is unknown.

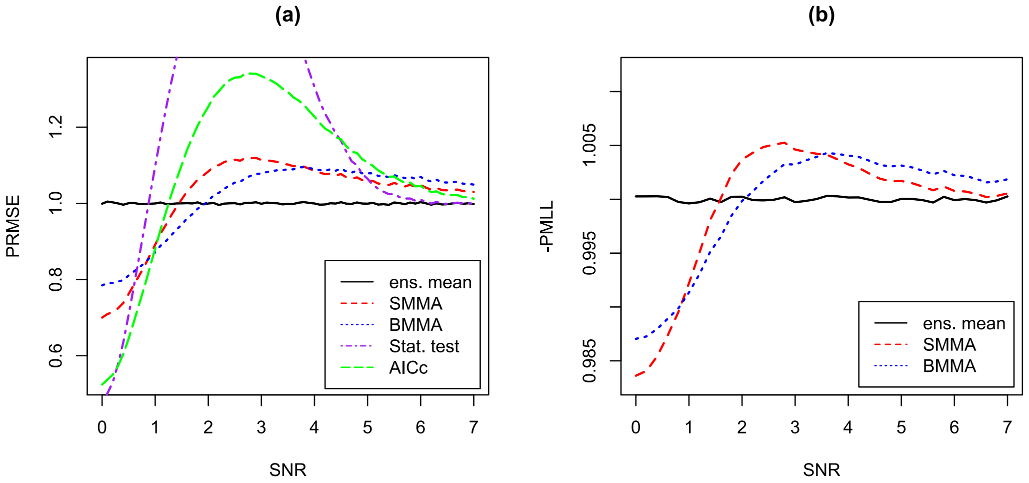


Figure 3: panel (a) shows the results of a simulation experiment for quantifying the performance of the two Mean-squared-error Model Averaging (MMA) methods and comparing with statistical testing and AICc model averaging. Panel (a) shows performance of point forecasts in terms of predictive root mean squared error (PRMSE). Panel (b) shows performance of probabilistic forecasts in terms of predictive mean log-likelihood (PMLL). The horizontal black solid line in both panels is the performance of the unadjusted ensemble mean, versus the real SNR, which would usually be unknown. The red dashed line in both panels shows the performance of the Simple MMA (SMMA) scheme and the blue dotted line in both panels shows the performance of the Bayesian MMA (BMMA) scheme. In panel (a) the purple dot-dashed line shows the performance of statistical testing and the green long-dashed line shows the performance of AICc model averaging.

We can also use simulations to test whether MMA gives better probabilistic predictions. Fig. 3b follows Fig. 3a, but now shows validation of probabilistic predictions using Predictive Mean Log-Likelihood (PMLL), which evaluates the ability of a prediction method to give reasonable probabilities across the whole probability distribution. We show the PMLL values as minus one times PMLL, to highlight the similarities between the results in panels (a) and (b). We only show probabilistic results for the ensemble mean, SMMA and BMMA. We see that the pattern of change in PMLL from using the two MMA methods is almost identical to the pattern of change in PRMSE: for small values of SNR, the MMA methods give better probabilistic predictions than the ensemble mean, while for large values of SNR, the MMA methods give less good probabilistic predictions than the ensemble mean. The relativity between SMMA and BMMA is also the same as for PRMSE.

One of the implications of these simulation results is that for variables for which the impact of climate change is large and unambiguous, corresponding to large SNR, such as temperature or sea-level rise in most cases, there is little justification for applying the MMA methods, or statistical testing, or AICc, since they would likely make predictions slightly worse. However, for variables such as rainfall, where the impact of climate change is often highly uncertain, corresponding to low SNR, these simulation results suggest it is worth exploring these methods since, depending on the size of the SNR values, they imply that they should improve the predictions relative to using the ensemble mean.

3.5 Previous Literature

Similar methods to SMMA have been studied in the statistics literature (see e.g., Copas (1983)) and are sometimes known as parameter shrinkage or damping. The SMMA method is adapted from a method used in commercial applied meteorology, where the same principles of bias-variance trade-off were used to derive better methods for fitting trends to observed temperature data for the pricing of weather derivatives (Jewson & Penzer, 2006). The adaptation of the method to ensemble climate predictions is described in a non-peer-reviewed technical report (Jewson & Hawkins, 2009a). The BMMA method was described and tested using simulations in a second non-peer-reviewed technical report (Jewson & Hawkins, 2009b). The present study is, we believe, the first attempt at large-scale testing of these methods using real climate predictions to evaluate whether they really are likely to improve predictions in practice.

4 Results for RCP4.5, 2011-2040, RTOT, Winter

We now show results for the SMMA method for the single case that was previously illustrated in Fig. 1. For this case, Fig. 4 shows values of the reduction factor $\hat{k}_S$, the adjusted ensemble mean $\hat{k}_S\hat{\mu}_c$ and the difference between the ensemble mean and the adjusted ensemble mean $\hat{\mu}_c(1-\hat{k}_S)$ as both absolute and relative values.

In Fig. 4a we see that in the regions where the ensemble mean is statistically significant (as shown in Fig. 1d), $\hat{k}_S$ is close to 1 and the SMMA method will have little effect. In the other regions it takes a range of values, and in some regions, e.g., parts of Spain, it is close to zero. These values of $\hat{k}_S$ lead to the prediction shown in Fig. 4b. The prediction does not, overall, look much different from the unadjusted prediction shown in Fig. 1a. The changes in the prediction are more clearly illustrated by the differences shown in Fig. 4c. We see the largest changes in the regions of low standard deviation and low SNR such as Portugal and the Alps. Fig. 4d shows that some of the changes are large relative to the ensemble mean (e.g., in parts of

https://doi.org/10.5194/npg-2021-12
Preprint. Discussion started: 19 March 2021

Spain, Italy and Greece). Overall, SMMA does not radically alter the patterns of climate change in the ensemble mean: it selectively identifies locations where the changes have high uncertainty and makes adjustments in those locations. The 330 impact is therefore local rather than large-scale.

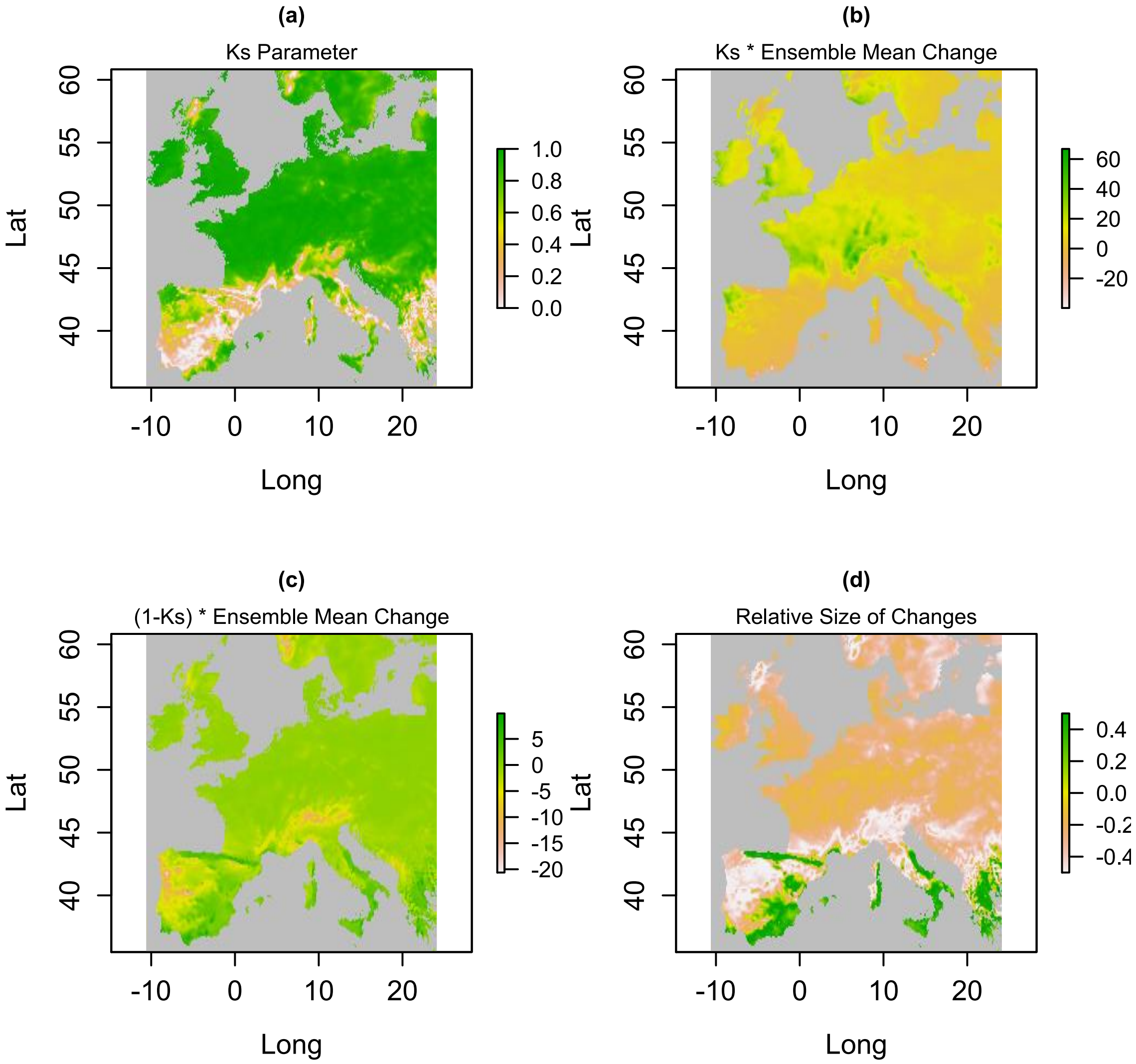


Figure 4: Various metrics derived from the EURO-CORDEX data shown in Fig. 1. Panel (a) shows the reduction parameter $\hat{k}_s$ for the SMMA method, panel (b) shows the ensemble mean reduced by the parameter $\hat{k}_s$, panel (c) shows the difference between the unadjusted ensemble mean and panel (b), and panel (d) shows the same difference as a fraction of the ensemble 335 mean.

Figure 5a shows a histogram of the values of SNR shown on the map in Fig. 1c. There are a large number of values below 2, which correspond to non-significant changes in the ensemble mean. Figure 5b shows a histogram of the values of $\hat{k}_S$ shown on the map in Fig. 4a. Many of the $\hat{k}_S$ values are close to one, corresponding to regions where the change in the ensemble is significant, and where the MMA methods will have little impact. However, there are also values all the way down to zero, corresponding to regions where the ensemble mean change is not significant, and where the MMA methods will have a larger impact.

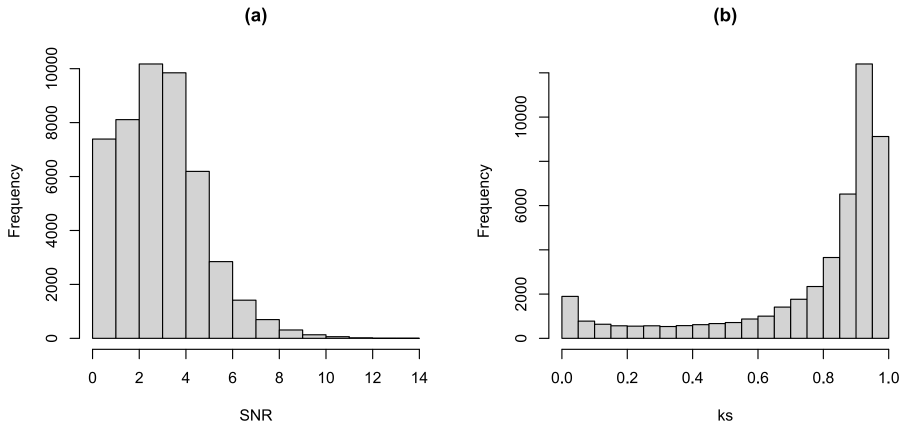


Figure 5: The left-hand panel shows the frequency distribution of the SNR values shown in Fig. 1c and the right-hand panel shows the frequency distribution of the k values shown in Fig. 3a.

4.1 Cross-validation

We can test whether the adjusted ensemble means created by the MMA methods are really likely to give more accurate predictions than the unadjusted ensemble mean, as the theory suggests they might, by using leave-one-out cross-validation within the ensemble. Cross-validation is commonly used for evaluating methods for processing climate model output in this way (see e.g., Raisanen and Ylhaisi (2010)). This only evaluates *potential* predictive skill, however, since as we are considering projections of future climate, it cannot involve observations. We apply the following steps:

- At each location, for each of the 72 cases, we cycle through the ten climate models, missing out each model in turn
- We use the nine remaining climate models to estimate the reduction factors $\hat{k}_S$ and $\hat{k}_B$.
- We make five predictions using the ensemble mean, the SMMA method, the BMMA method, statistical significance testing and AICc model averaging.
- We compare each of the five predictions with the value from the model that was missed out
- We calculate the PMSE over all ten models and all locations, for each of the predictions.

- We calculate the PMLL for the ensemble mean and the MMA methods
- We calculate the ratio of the PRMSE for the adjusted ensemble mean and statistical significance predictions to the PRMSE of the unadjusted ensemble mean prediction, so that values less than one indicate a better prediction than the unadjusted ensemble mean prediction.
- We also calculate the corresponding ratio for the PMLL results.

For the case illustrated in Fig. 1 and Fig. 4, we find a value of the PRMSE ratio of 0.960 for the SMMA method, 0.930 for the BMMA method, 1.100 for significance testing and 0.964 for AICc. Since the SMMA, BMMA and AICc methods give values that are less than 1, we see that the adjusted ensemble means are, on average over the whole spatial field, giving predictions with a lower PMSE than the ensemble mean prediction. The predictions are 4%, 8% and 4% more accurate, respectively, as estimates of the unknown mean. Since statistical testing gives a value greater than one, we see that it is giving predictions with higher PMSE than the ensemble mean prediction. All these values are a combination of results from all locations across Europe. The PMSE values from the SMMA, BMMA and AICc methods are lower than those from the ensemble mean in the spatial average but are unlikely to be lower at every location. From the simulation results shown in Sect. 3.4 above we know that the MMA and AICc methods are likely giving better results than the unadjusted ensemble mean in regions where the SNR is low, but less good results where the SNR is high. The final average values given above are therefore in part a reflection of the relative sizes of the regions with low and high SNR.

The values of the PMLL ratio for SMMA and BMMA are 0.9983 and 0.9982, and we see that the probabilistic predictions based on the MMA-adjusted ensemble means are also improved relative to probabilistic predictions based on the unadjusted ensemble mean. The changes in PMLL are small, but our experience is that small changes are typical when using PMLL as a metric.

5 Results for 72 Cases

We now expand our cross-validation testing from one case to all 72 cases, across four seasons, three variables, two RCPs and three time-horizons. Fig. 6 shows the spatial means of the estimates of k for both MMA methods for all these cases, stratified by season, RCP, variable and time horizon. The format of Fig. 6 follows the format of Fig. 2: each panel contains 72 black circles and 72 red crosses. Each black circle is the spatial mean over all the estimates of k from the SMMA method for one of the 72 cases. Each red cross is the corresponding spatial mean estimate of k from the BMMA method. The horizontal lines show the means of the estimates within each sub-set. Figure 6a shows that the estimates of k from both methods decrease from DJF to SON. This is because of the decreasing SNR values shown in Fig. 2a. The BMMA method gives higher k estimates than the SMMA method on average, and a lower spread of values. There is no clear impact of rainfall variable on the k values (Fig. 6b). Figure 6c shows higher k values for RCP8.5 than RCP4.5, reflecting the SNR values shown in Fig. 2c. Figure 6d shows k values increasing with time into the future, reflecting the increasing SNR values shown in Fig. 2d.

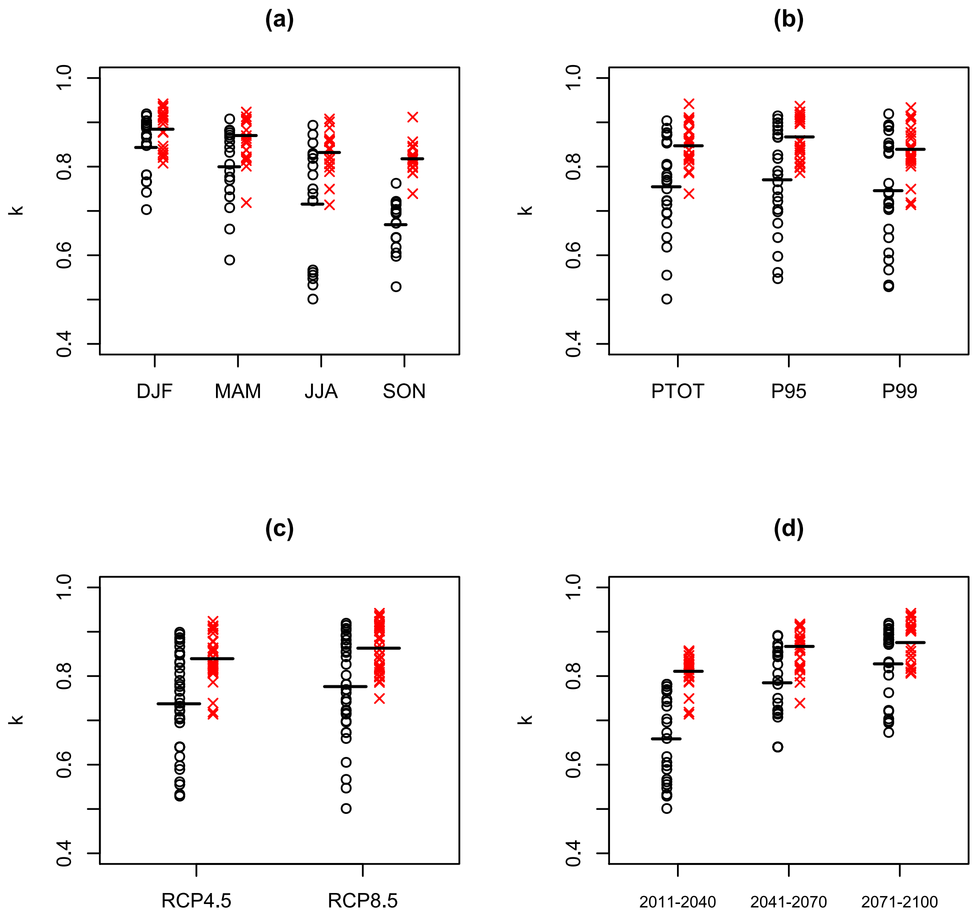


Figure 6: Each panel shows the same 72 values of the Europe-wide spatial means of the parameters $\hat{k}_S$ (black circles) and $\hat{k}_B$ (red X's) derived from the 72 EURO-CORDEX climate change projections described in the text, along with means within each subset (horizontal lines). Panel (a) shows the 72 values as a function of season, panel (b) shows them as a function of rainfall variable, panel (c) shows them as a function of RCP and panel (d) shows them as a function of time period. 395

Figure 7 shows corresponding spatial mean PRMSE results and includes results for significance testing (blue plus signs) and AICc (purple triangles). For the SMMA method the PRMSE reduces (relative to the PRMSE of the unadjusted ensemble mean) for 45 out of 72 cases, while for the BMMA method the PRMSE reduces for 51 out of 72 cases. Significance testing performs much worse than the other methods, and only reduces the PRMSE for 5 out of 72 cases. AICc reduces PRMSE for 27 out of 72 cases and so performs better than statistical testing but less well than the unadjusted ensemble mean.

Considering the relativities of the results between SMMA, BMMA, significance testing and AICc by subset: BMMA gives the best results overall and beats SMMA for 10 out of 12 of the subsets tested. Significance testing gives the worst results and is beaten by SMMA, BMMA and AICc model averaging in every subset. Considering the results of SMMA, BMMA significance testing and AICc relative to the unadjusted ensemble mean by subset: SMMA beats the ensemble mean for 11 out of 12 of the subsets tested, BMMA beats the ensemble mean for 12 out of 12 of the subsets tested, significance testing never beats the ensemble mean and AICc beats the ensemble mean for 2 out of 12 of the subsets tested. Considering the variation of PRMSE values by season (Fig. 7a) we see that the SMMA, BMMA, significance testing and AICc all perform gradually better through the year, and best in SON, as the SNR ratio reduces (see Fig. 2a). In SON the results for SMMA and BMMA for each of the 18 cases in that season are individually better than the ensemble mean. Considering the variation of PRMSE values by rainfall variable and RCP (Fig. 7b and Fig. 7c), we see little obvious pattern. Considering the variation of PRMSE values by time period, we see that SMMA and BMMA show the largest advantage over the unadjusted ensemble mean for the earliest time period, again because of the low SNR values (Fig. 2d).

Considering results over all 72 cases we find average PRMSE ratios of 0.956 and 0.946 for the SMMA and BMMA methods respectively, corresponding to estimates of the future mean climate that are roughly 4% and 5% more accurate than the predictions made using the unadjusted ensemble mean. For significance testing we find average PRMSE ratios of 1.226, corresponding to estimates of the future mean climate that are roughly 23% less accurate than the predictions made using the unadjusted ensemble mean. For AICc we find average PRMSE ratios of 1.02, corresponding to estimates of the future mean climate that are roughly 2% less accurate those from the unadjusted ensemble mean.

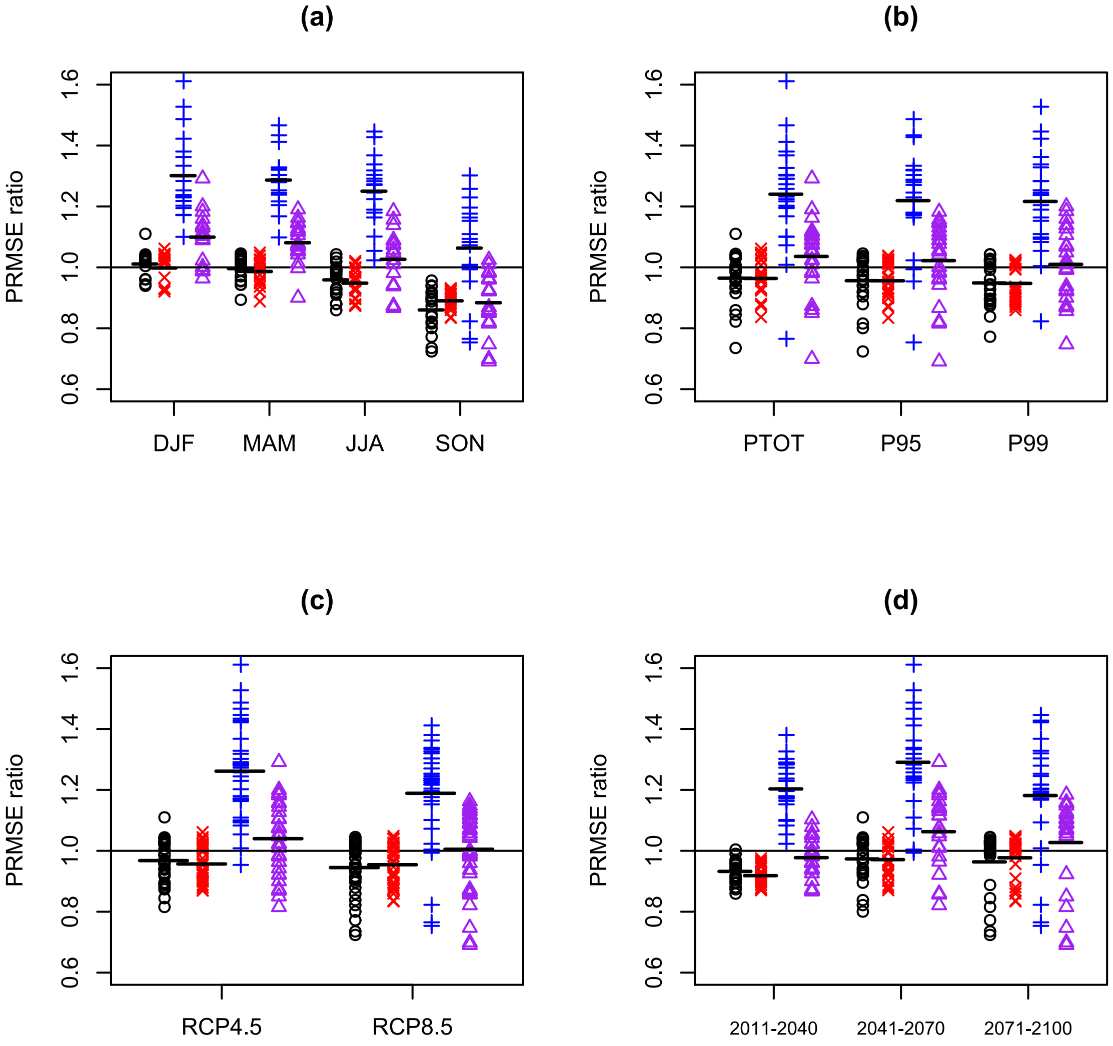


Figure 7: Each panel shows 72 values of the PRMSE ratio from the SMMA scheme (black circles), 72 values of the PRMSE ratio from the BMMA scheme (red X's), 72 values of the PRMSE ratio from significance testing (blue triangles), and 72 values of the PRMSE ratio from AICc model averaging (purple triangles), all derived from the 72 EURO-CORDEX climate change projections described in the text, along with means within each subset (horizontal lines). Panel (a) shows the 72 values as a function of season, panel (b) shows them as a function of rainfall variable, panel (c) shows them as a function of RCP and panel (d) shows them as a function of time period.

https://doi.org/10.5194/npg-2021-12
Preprint. Discussion started: 19 March 2021

Figure 8 is equivalent to Fig. 7, but shows results for PMLL i.e., evaluates the performance of probabilistic predictions. Given the poor performance of statistical testing and AICc in terms of PRMSE we do not show their results for PMLL. We see that the PMLL results are very similar to the PMSE results in Fig. 7, with BMMA showing the best results, followed by SSMA, followed by the unadjusted ensemble mean. For our EURO-CORDEX data, we conclude that making the mean of the prediction more accurate also makes the probabilistic prediction more accurate, which implies that the distribution shape being used in the probabilistic predictions is appropriate.

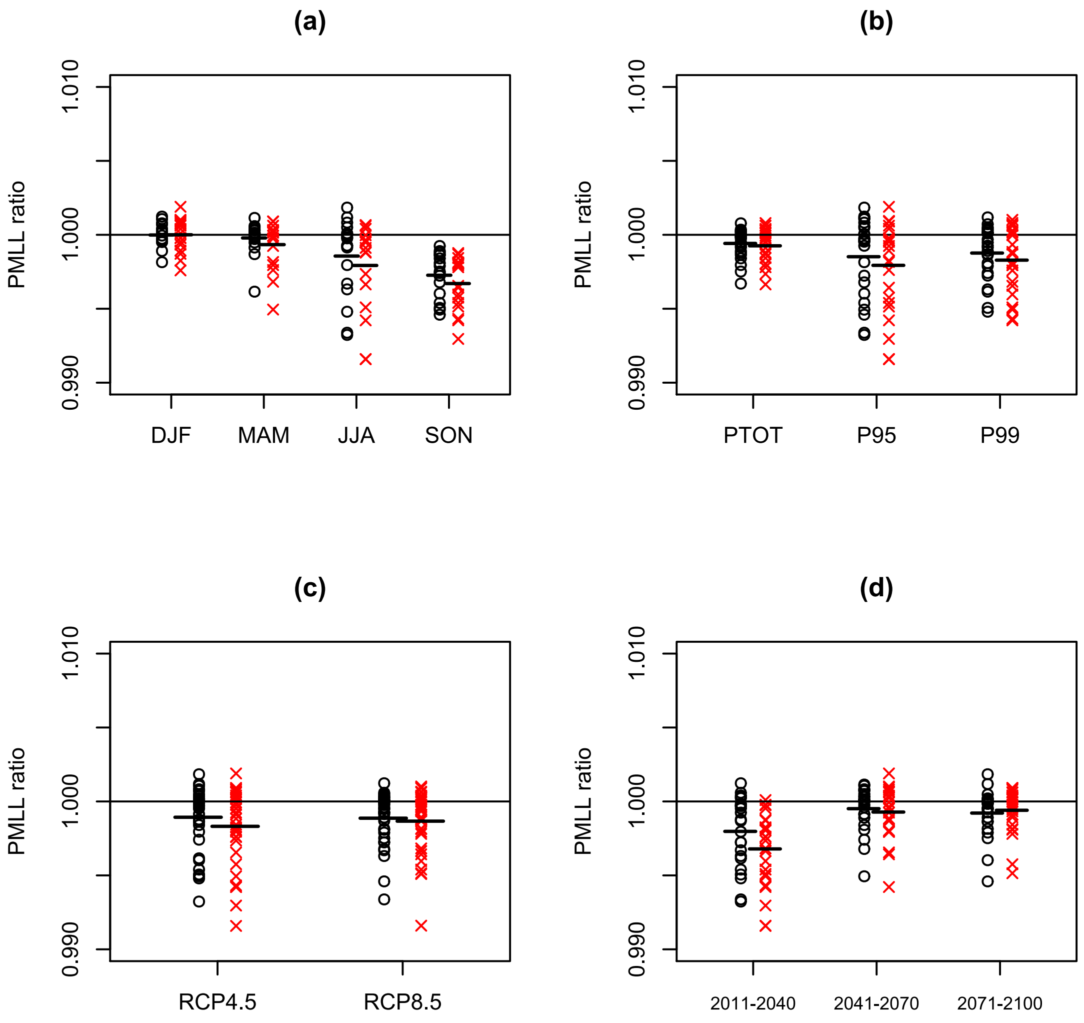
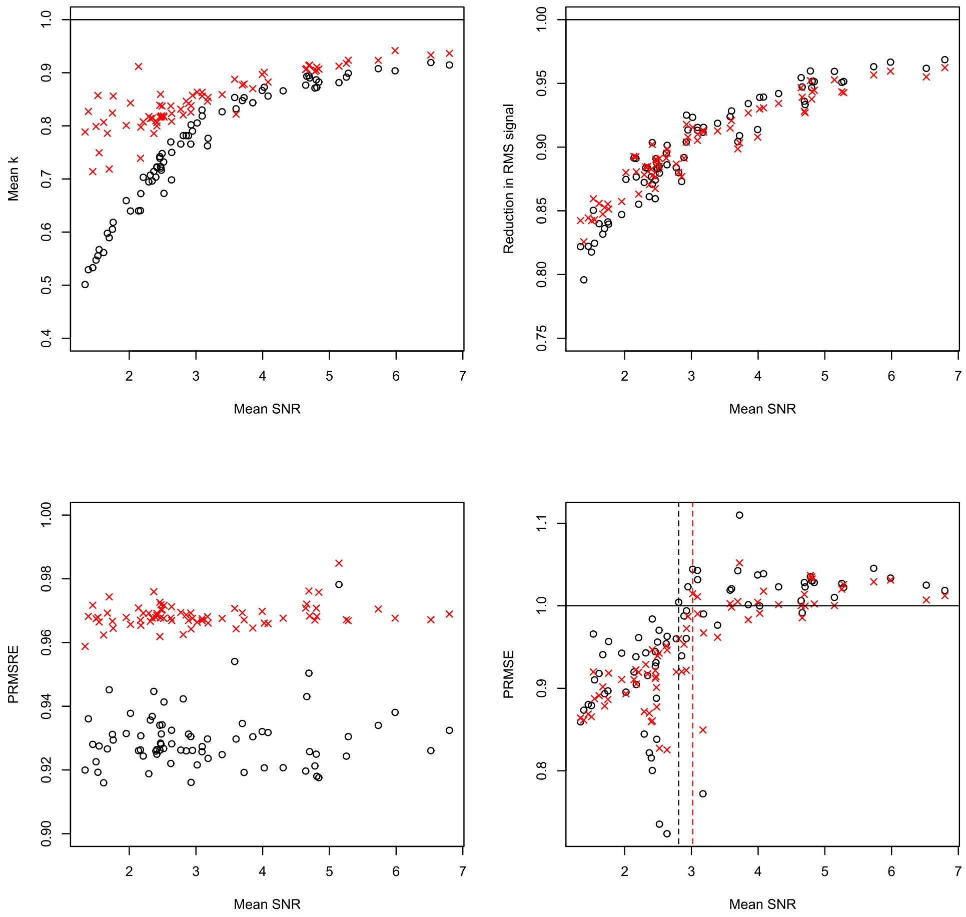


Figure 8: As Fig. 7, but now for 72 values of the PMLL ratio derived from probabilistic forecasts from SMMA (black circles) and BMMA (red X's).

https://doi.org/10.5194/npg-2021-12
Preprint. Discussion started: 19 March 2021

5.1 Further analysis

Figure 9 shows further analysis of these results. Figure 9a shows the mean values of the estimates of k for the SMMA and BMMA methods, versus the SNR for all 72 cases. The connection between the mean SNR and the mean k is now very clear, with mean k increasing with mean SNR. This panel also shows that the BMMA method gives higher k values on average for all values of SNR, but particularly for low values of SNR. Figure 9b shows the reduction in the root mean squared size (as opposed to error) of the prediction, which is a measure of the impact of the model averaging. The two methods give very similar results, in which the impact is greatest for the cases with low SNR. These are average reductions over the whole of Europe: locally, the reduction takes values in the whole range from 0 to 1. Figure 9c shows the PRMSE, but now calculated from relative errors, relative to the spatially varying ensemble mean. Values less than one for all 72 cases indicate that the MMA methods perform better than the unadjusted ensemble mean more comprehensively by this measure. The difference between these results and the straight PRMSE results arises because the locations where the MMA methods improve predictions the most on a relative basis tend to be the ones with small signals, which tend to have small prediction errors. These locations do not contribute very much to the straight PRMSE but contribute more when the errors are expressed in a relative sense. Figure 9d shows a scatter plot of the PRMSE versus the SNR for the two methods for all 72 cases. There is a clear relation in which the MMA methods perform best for small SNR values. The relation is similar to that shown in the simulation experiment results shown in Fig. 2, but with the cross-over points (shown by vertical lines) shifted to the right, because these are now relations between averages over many cases with different underlying values for the unknown real SNR. We see that for every case in which the mean SNR is less than 2.81 the SMMA method performs better than the unadjusted ensemble mean on average, and for every case in which the mean SNR is less than 3.02 the BMMA method performs better than the unadjusted ensemble mean on average.



460 Figure 9: various diagnostics for each of the 72 EURO-CORDEX climate change projections, plotted versus mean SNR. Results from applying the SMMA scheme are shown with black circles, and results from applying the BMMA scheme are shown with red X's. Panel (a) shows mean values of the parameters $\hat{k}_S$ and $\hat{k}_B$; panel (b) shows the reduction in the root mean square of the ensemble mean; panel (c) shows the reduction in the relative PRMSE and panel (d) shows the PRMSE. Panel (d) has additional vertical lines showing the cross-over points, below which the MMA results are all better than the ensemble mean results. 465

The results in Sect. 5 can be summarized as follows: for the EURO-CORDEX rainfall data, SMMA and BMMA give more accurate predictions on average, in both a point and probabilistic sense, than the unadjusted ensemble mean. BMMA gives more accurate results than SMMA. The MMA methods do well because the ensemble mean is uncertain and has low SNR values at many locations. The benefits of SMMA and BMMA are greatest in the cases with the lowest SNR values.

6 Discussion and Conclusions

Ensemble climate projections can be used to derive probability distributions for future climate, and the ensemble mean can be used as an estimate of the mean of the probability distribution. Because climate model ensembles are always finite in size, changes in the ensemble mean are always uncertain, relative to the changes in the ensemble mean that would be given by an infinite ensemble. The uncertainty varies in space. In regions where the signal-to-noise ratio (SNR) of the change in the ensemble mean is high, the change in the ensemble mean gives a precise estimate of the change in the mean climate that would be estimated from the infinite ensemble. However, in regions where the SNR is low, the interpretation of the change in the ensemble mean is a little more difficult. For instance, when the SNR is very low, the change in the ensemble mean is little more than random noise generated by variability in the members of the ensemble, and cannot be taken as a precise estimate of the change in mean climate of the infinite ensemble. In these cases, it might be unfortunate if the ensemble mean were interpreted too literally, or were used to drive adaptation decisions.

We have presented two bias-variance trade-off model averaging algorithms that adjust the change in the ensemble mean as a function of the SNR in an attempt to improve predictive accuracy. We call the methods Mean-squared error Model Averaging (MMA). These methods are designed to be applied after bias correction and corrections for inter-model correlation have been applied. One method is very simple (simple MMA, SMMA), and the other is a more complex Bayesian extension (Bayesian MMA, BMMA). The methods can both be thought as continuous generalisations of statistical testing, where instead of accepting or rejecting the change in the ensemble mean they apply continuous adjustment. They can also be thought of as small-sample corrections to the estimate of the ensemble mean. When the SNR is large the ensemble mean is hardly changed by these methods, while when the SNR is small the change in the ensemble mean is reduced towards zero in an attempt to maximise the predictive skill of the resulting predictions.

We have applied the MMA methods to a large set of rainfall projections from the EURO-CORDEX ensemble, for 72 different cases across four seasons, three different rainfall variables, two different RCPs and three future time periods during the 21st century. This data shows large variations in the SNR, which results in large variations of the extent to which the ensemble mean is adjusted by the methods.

We have used cross-validation within the ensemble to test whether the adjusted ensemble means achieve greater potential predictive skill for point predictions and probabilistic predictions. To assess point predictions we used predictive mean squared error (PMSE) and to assess probabilistic predictions we used predictive mean log-likelihood (PMLL). For both measures, we compared against results based on the unadjusted ensemble mean. For PMSE we have additionally compared against results based on statistical testing and small-sample Akaike Information Criterion model averaging (a standard method for model averaging). We emphasize that these calculations can only tell us about the potential accuracy of the method, not the actual accuracy, since we cannot compare projections of future climate with observations. On average over all 72 cases and all locations, the MMA methods reduce the PMSE, corresponding to what is roughly a 5% increase in potential accuracy in the estimate of the future mean climate. For the SMMA method, the PMSE reduces for 45 of the 72 cases, while for the BMMA method the PMSE reduces for 51 out of 72 cases. Which cases show a reduction in PMSE and which not depends strongly on the mean SNR within each case, in the sense that the MMA methods perform better when the SNR is low. For instance, the winter SNRs are high, and the average PMSE benefits of the MMA methods are marginal. The autumn SNRs are much lower, and the MMA methods beat the unadjusted ensemble mean in every case. Significance testing, by comparison, gives much worse PMSE values than the unadjusted ensemble mean, and AICc model averaging gives slightly worse PMSE values than the unadjusted ensemble mean. For PMLL we also found that the MMA methods beat the unadjusted ensemble mean.

The ensemble mean can be used as a standalone indication of the possible change in climate, or as the mean of a distribution of possible changes in a probabilistic analysis. We conclude that in both cases, when the ensemble mean is highly uncertain, the MMA-adjusted ensemble means described above can be used in its place. Applying MMA has various advantages: (a) it reduces the possibility of over-interpreting changes in the ensemble mean that are very uncertain, while not affecting more certain changes, (b) relative to significance testing, it avoids jumps in the ensemble mean change in space and between scenarios and (c) when the SNR is low, it will likely produce more accurate predictions than predictions based on either the unadjusted ensemble mean or statistical testing. In addition to the above advantages, relative to statistical testing the MMA-adjusted ensemble mean reduces the likelihood of false negatives (i.e., missing a change due to climate change) and increases the likelihood of false positives (i.e., falsely identifying a change as being due to climate change). Whether this is an advantage or not depends on the application, but is typically beneficial for risk modelling.

Author Contribution

SJ designed the study and the algorithms, wrote and ran the analysis code, produced the graphics and wrote the text. GB, PM and JM extracted the EURO-CORDEX data. MS wrote the code to read the EURO-CORDEX data. All the authors contributed to proof-reading.

Acknowledgements

The authors would like to thank Tim DelSole, Ed Hawkins, Arno Hilberts, Luke Jackson, Shree Khare, Ludovico Nicotina and Jeremy Penzer for interesting discussions on this topic, Francesco Repola from CMCC for assisting with data extraction, and Casper Christophersen, Marie Scholer and Luisa Mazzotta from EIOPA for arranging the collaboration with CMCC. We also thank EURO-CORDEX, and the climate modelling groups (listed in Table 1 of this paper) for producing and making their model output freely available.

Data Availability

The EURO-CORDEX data used in this study is freely available. Details are given at https://euro-cordex.net.

Competing Interests

MS works for RMS Ltd, a company that quantifies the impacts of weather, climate variability and climate change.

References

Barnes, E., Hurrell, J., Ebert-Uphoff, I., Anderson, C., & Anderson, D. (2019). Viewing Forced Climate Patterns Through an AI Lens. *GRL, 46*(22), 13389-13398.

Benestad, R., Haensler, A., Hennemuth, B., Illy, T., Jacob, D., Keup-Thiel, E., . . . Zsebehazi, G. (2017, 08 1). *Guidance for EURO-CORDEX*. Retrieved 1 9, 2021, from https://www.euro-cordex.net/imperia/md/content/csc/cordex/euro-cordex-guidelines-version1.0-2017.08.pdf

Burnham, K., & Anderson, D. (2010). *Model Selection and Multimodel Inference.* New York: Springer-Verlag.

Chen, J., Brissette, F., Zhang, X., Chen, H., Guo, S., & Zhao, Y. (2019). Bias correcting climate model multi-member ensembles to assess climate change impacts on hydrology. *Climatic Change*, *153*(3), 361-377.

Copas, J. (1983). Regression, Prediction and Shrinkage. *Journal of the Royal Statistical Society, Series B, 45*(3), 311-354.

DelSole, T., Yang, X., & Tippett, M. (2013). Is Unequal Weighting Significantly Better than Equal Weighting for Multi-Model Forecasting? *Q. J. R. Meteorol. Soc.*, *139*(670), 176-183.

Deser, C., Phillips, A., Bourdette, V., & Teng, H. (2010). Uncertainty in climate change projections: the role of internal variability. *Climate Dynamics, 38*, 527–546.

European Environment Agency. (2017). *Indicator Assessment: Mean Precipitation*. Retrieved 4 15, 2020, from https://www.eea.europa.eu/data-and-maps/indicators/european-precipitation-2/assessment

Eyring, V., Bony, S., Meehl, G., Senior, C., Stevens, B., Stouffer, R., & Taylor, K. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.*, *9*, 1937-1958.

Frankcombe, L., England, M., Mann, M., & Steinman, B. (2015). Separating Internal Variability from the Externally Forced Climate Response. *J. Clim.* , *28*(20), 8184–8202.

Hawkins, E., & Sutton , R. (2009). The Potential to Narrow Uncertainty in Regional Climate Predictions. *BAMS*, *90*(8), 1095–1108.

Hawkins, E., & Sutton , R. (2012). Time of emergence of climate signals. *GRL*, *39*(1), 1-6.

Hingray, B., & Said, M. (2014). Partitioning Internal Variability and Model Uncertainty Components in a Multimember Multimodel Ensemble of Climate Projections. *J. Clim.*, *27*, 6779–6798.

Jacob, D., Petersen, J., & authors. (2014). EURO-CORDEX: new high-resolution climate change projections for European impact research. *Reg Environ Change*, *14*, 563-578.

Jacob, D., Teichmann, C., & authors. (2020). Regional climate downscaling over Europe: perspectives from the EURO-CORDEX community. *Regional Environmental Change, 20*.

Jewson, S., & Hawkins, E. (2009a). *Improving the expected accuracy of forecasts of future climate using a simple bias-variance tradeoff*. Retrieved from arXiv: https://arxiv.org/abs/0911.1904

Jewson, S., & Hawkins, E. (2009b). *Improving Uncertain Climate Forecasts Using a New Minimum Mean Square Error Estimator for the Mean of the Normal Distribution*. Retrieved from arXiv: https://arxiv.org/abs/0912.4395

Jewson, S., & Penzer, J. (2006). Estimating Trends in Weather Series: Consequences for Pricing Derivatives. *Studies in Nonlinear Dynamics & Econometrics, 10*(3), 1-10.

Jewson, S., Barnes, C., Cusack, S., & Bellone, E. (2019). Adjusting catastrophe model ensembles using importance sampling, with application to damage estimation for varying levels of hurricane activity. *Met Apps, 27*, 1-14.

Kaczmarska, J., Jewson, S., & Bellone, E. (2018). Quantifying the sources of simulation uncertainty in natural catastrophe models. *SERRA, 32*(3), 591-605.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., & Meehl, G. (2010). Challenges in combining projections from multiple climate models. *J. Clim., 23*(10), 2739–2758.

Knutti, R., Sedlacek, J., Sanderson, B., Lorenz, R., Fischer, E., & Eyring, V. (2017). A climate model projection weighting scheme accounting for performance and interdependence. *GRL, 44*, 1909-1918.

Lee, P. (1997). *Bayesian Statistics* (2 ed.). London: Arnold.

Lehner, F., Deser, C., & Terray, L. (2017). Toward a New Estimate of "Time of Emergence" of Anthropogenic Warming: Insights from Dynamical Adjustment and a Large Initial-Condition Model Ensemble. *J. Clim., 30*(19), 7739–7756.

Meinshausen, M., Smith, S., Calvin, K., Daniel, J., Kainuma, M., Lamarque, J., . . . Vuuren, D. (2011). The RCP greenhouse gas concentrations and their extensions from 1765 to 2300. *Climatic Change, 109*.

Moss, R., Edmonds, J., Hibbard, K., Manning, M., Rose, S., van Vuuren, D., . . . Wilbanks, T. (2010). The next generation of scenarios for climate change research and assessment. *Nature, 463*, 747-756.

Pachauri, K., & Meyer, L. (2014). *IPCC, 2014: 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.* Geneva: IPCC.

Raisanen, J., & Ylhaisi, J. (2010). How Much Should Climate Model Output Be Smoothed in Space? *J. Clim, 24*, 867-880.

Sanderson, B., Knutti, R., & Caldwell, P. (2015b). Addressing interdependency in a multimodel ensemble by interpolation of model properties. *J. Clim., 28*, 5150-5170.

Sassi, M., Nicotina, L., Pall, P., Stone, D., Hilberts, A., Wehner, M., & Jewson, S. (2019). Impact of climate change on European winter and summer flood losses. *Advances in Water Resources, 129*, 165-177.

Sippel, S., Meinshausen, N., Merrifield, A., Lehner, F., Pendergrass, A., Fischer, E., & Knutti, R. (2019). Uncovering the Forced Climate Response from a Single Ensemble Member Using Statistical Learning. *J. Clim. , 32*(17), 5677–5699.

Taylor, K., Stouffer, R., & Meehl, G. (2012). An Overview of CMIP5 and the Experiment Design. *BAMS, 93*(4), 485-498.

Thompson, D., Barnes, E., Deser, C., Foust, W., & Phillips, A. (2015). Quantifying the Role of Internal Climate Variability in Future Climate Trends. *J. Clim., 28*(16), 6443–6456.

Wilks, D. (2011). *Statistical Methods in the Atmospheric Sciences* (3 ed.). Oxford: AP.

Wills, R., Battisti, D., Armour, K., Schneider, T., & Deser, C. (2020). Pattern Recognition Methods to Separate Forced Responses from Internal Variability in Climate Model Ensembles and Observations. *J. Clim.* , *33*(20), 8693–8719.

Yip, S., Ferro, C., Stephenson, D., & Hawkins, E. (2011). A Simple, Coherent Framework for Partitioning Uncertainty in Climate Predictions. *J. Clim., 24*(17), 4634–4643.