Geoscientific
Model Development

A Markov chain method for weighting climate model ensembles

Max Kulinich[1], Yanan Fan[1], Spiridon Penev[1], Jason P. Evans[2], and Roman Olson[3]

[1]School of Mathematics and Statistics, UNSW Sydney, Australia
[2]Climate Change Research Centre and ARC Centre of Excellence for Climate Extremes, UNSW Sydney, Australia
[3]Irreversible Climate Change Research Center, Yonsei University, South Korea

Correspondence: Max Kulinich (m.kulinich@student.unsw.edu.au)

Abstract. Climate change is typically modeled using sophisticated mathematical models (climate models) of physical processes that range in temporal and spatial scales. Multimodel ensemble means of climate models show better correlation with the observations than any of the models separately. Currently, an open research question is how climate models can be combined to create an ensemble mean in an optimal way. We present a novel stochastic approach based on Markov chains to estimate model weights in order to obtain ensemble means. The method was compared to existing alternatives by measuring its performance on training and validation data, as well as model-as-truth experiments. The Markov chain method showed improved performance over those methods when measured by the root mean squared error in validation and comparable performance in model-as-truth experiments. The results of this comparative analysis should serve to motivate further studies in applications of Markov chain and other nonlinear methods that address the issues of finding optimal model weight for constructing ensemble means.

1 Introduction

Climate change is often modeled using sophisticated mathematical models of physical processes taking place over a range of temporal and spatial scales. These models are inherently limited in their ability to represent all aspects of the modeled physical processes. Simple averages of multimodel ensembles of GCMs (global climate models) often show better correlations with the observations than any of the individual models separately (Kharin and Zweirs, 2002; Feng et al., 2011). Knutti et al. (2010) point out that often the equal-weighted averages ("one model, one vote") approach is used as a best-guess result, assuming that individual model biases will at least partially cancel each other out. This approach assumes that all models are (a) reasonably independent, (b) equally plausible, (c) distributed around reality and (d) that the range of their projections is representative of what we believe is the uncertainty in the projected quantity. However, these assumptions are rarely fulfilled (Knutti et al., 2017), and thus a better way of finding a weighted ensemble mean is required (Herger et al., 2018; Sanderson et al., 2017).

Most studies attempting to define an optimal ensemble weighting either employ linear optimization techniques (Krishnamurti et al., 2000; Majumder et al., 2018; Abramowitz et al., 2019) or are based on a specification of likelihoods for the model and observation data (Murphy et al., 2004; Fan et al., 2017). Such methods are inevitably limited by the strong assumptions used for their design. We seek to weaken those assumptions and to complement the existing methods with a more flexible nonlinear optimization approach. An unresolved issue in using weights for models is that models have interdependence, due to the sharing of computer codes, parameterizations, etc. (Olson et al., 2019). Abramowitz et al. (2019) points out that model dependence can play a crucial role when assembling the models into an ensemble. Mathematically, interdependence often results in closeness of model outputs in model output space. If a large cluster of highly dependent models is included in an ensemble with equal weights, the overall ensemble mean will become close to the dependent models' cluster. Ignoring model dependence can lead to bias and overconfidence in future climate model projections (Leduc et al., 2015; Steinschneider et al., 2015).

Hence, it is desirable that an ensemble weighting method is robust against the dependency issue and has normalized non-negative weights for interpretability. Finally, the methods should work well across a range of different climate variables, such as temperature, precipitation, etc.

In this paper, we propose a novel way to construct a weighted ensemble mean using Markov chains, which we call the Markov Chain Ensemble (MCE) method. Our purpose is to demonstrate that going beyond linear optimization on a vector space of climate models' outputs allows better-performing weighted ensembles to be built. We selected Markov chains as a basis for such nonlinear optimization as one of the most straightforward nonlinear structures. It naturally produces non-negative weights that sum to one and captures some of the nonlinear patterns in the ensemble (here we refer to nonlinear patterns as time-dependent selections of model components rather than considering a complete model output vector). It performs well on a range of datasets when compared to the standard simple mean and linear optimization weighting methods as we demonstrate below. We also examine how the method responds to the introduction of interdependent models.

Although Markov chains have been used frequently in the literature for the prediction of future time series (e.g., Bai and Wang, 2011; Pesch et al., 2015), to the best of our knowledge, this is the first time the method has been applied to building weighted climate model ensemble means. In this paper, we use the "memoryless" property of Markov chains at each time step to capture the dynamic change in models' fit through the time series. This dynamic change, through time, is represented by the transition matrix, which describes the probability of each model being the best fit for the next observation at time $t + 1$, given the best fit for the current time t. The transition matrix is built based on the input data and describes probable future states given the current state. The stationary distribution of this transition matrix is used for weighted ensemble creation and reflects the relative contribution of each model to the total weighted ensemble mean forecast.

We describe the datasets used in this study and the proposed MCE method in Sect. 2. We compare the proposed method (MCE) to the commonly used multi-model ensemble average (AVE) method (Lambert and Boer, 2001) and the convex optimization (COE) method proposed by Bishop and Abramowitz (2013) and present the results in Sect. 3, followed by a discussion in Sect. 4 and conclusion in Sect. 5.

2 Methods

2.1 Data

Here we first describe the datasets used in this study. We have chosen three publicly available datasets with differing numbers of models, historical period lengths and model interde-

pendence levels to evaluate and compare the performance of the MCE method with alternative approaches.

2.1.1 CMIP5 data

The first dataset we use is the temperature anomaly (°C) data from Coupled Model Intercomparison Project (CMIP5) with 39 different global climate model (GCM) outputs (one ensemble member per model) and Hadley Centre/Climatic Research Unit Temperature observations (HadCRUT4). The data are obtained from https://climexp.knmi.nl (last access: 21 July 2020), and the period of 1900–2099 is selected for the analysis. It contains temperature anomalies (monthly averages) compared to the reference period of 1961–1990 (Taylor et al., 2011). This dataset contains several clusters of dependent models and has both positive and negative data values, a relatively low variability, and long time series.

2.1.2 NARCliM data

The second dataset contains temperature output from the New South Wales (NSW) and Australian Capital Territory Regional Climate Modelling project (Evans et al., 2014). It contains regional climate model (RCM) simulations over southeastern Australia. Specifically, three RCM versions were forced with 4 global climate models each, for a total of 12 ensemble members. The data contain annual time series of mean summer temperature (°C) for the Far West NSW state planning region as modeled by the NARCliM domain RCMs for the periods 1990–2019 and 2030–2039 (Olson et al., 2016). Corresponding temperature observations are obtained from the Australian Water Availability Project (AWAP) (Jones et al., 2009). The dataset has a high ratio of the number of models to the number of observations. While NARCliM model choice explicitly considered model dependence for both the RCMs as well as the driving GCMs, the resulting ensemble demonstrates an apparent similarity between the simulations (i.e., model interdependence) in small clusters.

2.1.3 KMA data

The third dataset contains yearly heatwave amplitudes (HWAs) for the Korean Peninsula from 29 CMIP5 climate models and observations between 1973 and 2005 (Shin et al., 2017). In particular, HWA contains the difference between the highest temperature during the heatwave events for the corresponding year and the 95th percentile of daily maximum summer temperatures from 1973 to 2005. This framework was discussed in detail in Fischer and Schär (2010). Here a heatwave event occurs when the daily maximum temperature is above the 95th percentile of daily maximum summer temperatures (32.82 °C) for two consecutive days. Daily maximum temperature data used for the calculation of observed HWAs are the mean of 59 weather stations operated by the Korea Meteorological Administration (KMA). Shin et

al. (2017) provides the list of CMIP5 models included in the study. HWA data are non-negative and can be highly skewed with long upper tails as they measure extreme events; therefore, the dataset is highly non-Gaussian. These properties allow us to test methods in more challenging scenarios, where likelihood-based approaches are more difficult to apply.

These three datasets cover different scenarios, data structures, parameter distributions and scales (see Table 1). Such coverage allows us to analyze the performance and the inherent limitations of the proposed method. In this pilot study we use spatially averaged data, which limits physical interpretability of the model weights, but the method can be extended to spatially distributed data.

2.2 Markov chain ensemble (MCE) method

Generally, a homogeneous Markov chain is a sequence of random system states evolving through time, where each next state is defined sequentially based on its predecessor and predefined transition probabilities (Del Moral and Penev, 2016, p. 121). Suppose that there is a finite number of probable system states $S = \{s_1, \ldots, s_N\}$; then this dependency can be described through a transition matrix $\mathbf{P}$ (with $\mathbf{P}(x, y) \in [0, 1]$ and $\sum_y \mathbf{P}(x, y) = 1$, for any $x, y \in S$):

$$\forall x, y \in S, \ \mathbf{Pr}(X_{n+1} = y | X_n = x) = \mathbf{P}(x, y). \quad (1)$$

In this study, we want to utilize the "fundamental limit theorem for regular chain," which states that if $\mathbf{P}$ is a transition matrix for a regular Markov chain (where $\forall x, y \in S$, $\mathbf{P}(x, y) > 0$), then $\lim_{n \to \infty} \mathbf{P}^n = \mathbf{P}^\infty$, where $\mathbf{P}^\infty$ is a matrix with all rows being equal and having strictly positive entries.

This property allows us to construct a non-negative transition matrix $\mathbf{P}$ by distillation of input information (i.e., model outputs and historical observations) and allows $\mathbf{P}$ to converge to a unique vector of model weights $\boldsymbol{w} = (w_1, w_2, \ldots, w_N)$, where N is the total number of models in a given ensemble. The vector $\boldsymbol{w}$ can be obtained by solving the equation $\boldsymbol{w}\mathbf{P} = \boldsymbol{w}$. The converged transition matrix represents a probability of selecting one of the models for any of the time steps in the future when observations are not available. Hence, we propose to use it as a weighting vector for constructing a weighted ensemble mean forecast and test this proposition using cross-validation in the following sections.

More precisely, we start by constructing a transition vector $\boldsymbol{v}$ (based on the input data), which specifies a choice of the optimal model at any given time step t. Using the vector $\boldsymbol{v}$ we construct a transition matrix $\mathbf{P}$ and find its stationary distribution $\boldsymbol{w}$. The resulting weighted ensemble mean is constructed by applying $\boldsymbol{w}$ on the given climate model outputs. We call this process the Markov Chain Ensemble (MCE) algorithm, and it uses historical observations and equivalent climate model simulations as the input data to calculate a set of weights for the future ensemble mean as an output. Table 2 gives a step-by-step description of the MCE algorithm.

We provide some details of the algorithm as described in Table 2 in the following paragraphs.

Initialization of transition matrix $\mathbf{P}^0$. In order for the Markov chain to be regular we set $\mathbf{P}^0(x, y) = \lambda$, $\forall x, y \in S$, where λ equals the lowest computationally possible positive number $\lambda = 2.225074 \times 10^{-308}$ in the R software (R Core Team, 2013).

Initialization of σ *interval.* To avoid division by 0 in Eq. (2) and to prevent Eq. (2) from converging to $1/N$ the initial σ interval is set to [0.1, 1].

Step 1. The MCE method proceeds by utilizing each model output in an optimal way based on its ability to resemble observational data at each given time point. This resemblance is measured by a distance-based probability matrix $\mathbf{D}$ of size $N \times T_1$, using a normalized exponential function.

$$d_{i,k} = \frac{e^{-\left(\frac{M_{i,k} - O_k}{\sigma}\right)^2}}{\sum_{j=1}^{N} e^{-\left(\frac{M_{j,k} - O_k}{\sigma}\right)^2}}, \quad (2)$$

where $1 \leq k \leq T_1 \leq T$, T_1 indicates the length of the training period, and T is the length of the entire historical period included in the study. Additionally, $1 \leq j \leq N$, where N is the number of models included, and σ is chosen randomly as described above.

Step 2. Based on the matrix $\mathbf{D}$ a simulation is performed at each time step $1 \leq k \leq T_1$ by randomly selecting one of the models i with probability proportional to its value $d_{i,k}$. In this way we construct a vector $V = (v_1, v_2, v_3, \ldots, v_{T_1})$, which represents the choice of models closest to observations at each time step.

Step 3. Then the initial matrix $\mathbf{P}^0$ is updated step-wise $(\mathbf{P}^1, \mathbf{P}^2, \ldots, \mathbf{P}^{T_1})$ to capture the transitions between models present in vector V. For each t $(1 \leq t \leq T_1 - 1)$, $\mathbf{P}^i_{V_i, V_{i+1}} = \mathbf{P}^{i-1}_{V_i, V_{i+1}} + 1$.

Step 4. The resulting matrix is normalized by row $\mathbf{P}^*_i = \mathbf{P}^{T_1}_i / \sum_{j=1}^{N} \mathbf{P}^{T_1}_{i,j}$, for each $1 \leq i \leq N$.

Step 5. The stationary distribution $\boldsymbol{w}$ is obtained by solving $\boldsymbol{w}\mathbf{P}^* = \boldsymbol{w}$. A standard R software package is used to find the solution in this study.

Step 6. The ensemble mean is constructed based on weights w and calculate its RMSE_{Tr}.

Step 7. Steps 2–6 are repeated L times, where L is selected based on the external requirements of precision of the results and of computational power available.

Step 8. Select the set of weights $\boldsymbol{w}^*$ with the best performance on the training set (with the lowest $\text{RMSE}^*_{\text{Tr}}$).

Step 9. Construct the E_{MCE} ensemble using the selected $\boldsymbol{w}^*$.

2.2.1 Parameter sensitivity

From Eq. (2) it is clear that having a small σ will result in distances d close to $1/N$. Having a large σ will result in all the distances becoming marginal, with the exception of the

Table 1. Summary of CMIP5, NARCLiM and KMA data properties.

Dataset	Climate variable	Minimum	Maximum	Variance	Number of observations	Number of models
CMIP5	Temperature (°C)	−0.80	1.16	0.12	1440	39
NARCLiM	Temperature (°C)	9.38	31.64	36.61	240	12
KMA	HWA (°C)	0	1.49	0.18	33	29

Table 2. The Markov chain ensemble (MCE) algorithm.

Input:		– length of training period T_1,
		– historical observations O_t, at times $t = 1, \ldots, T_1$,
		– climate model output $M_{i,t}$, at times $t = 1, \ldots, T_1$, for $i = 1, \ldots, N$ models,
		– an initialized number of simulations, L,
		– an initialized σ interval $[\sigma_{\min}, \sigma_{\max}]$, and
		– an initialized transition matrix $\mathbf{P}^0$ of $N \times N$ size.

Step 1.	Randomly select $\sigma \in [\sigma_{\min}, \sigma_{\max}]$ and compute the distance matrix $\mathbf{D}$ according to Eq. (2).
Step 2.	Construct a sequence vector $\boldsymbol{v}$ based on $\mathbf{D}$ using stochastic simulations.
Step 3.	Update $\mathbf{P}^0$ step-wise by increasing probability of transitions contained in $\boldsymbol{v}$: $\mathbf{P}^0 \to \mathbf{P}^1 \to \ldots \to \mathbf{P}^{T_1}$.
Step 4.	Obtain normalized transition matrix $\mathbf{P}^*$, by normalizing $\mathbf{P}^{T_1}$ row-wise so that each row sums to 1.
Step 5.	Find $\boldsymbol{w}$ by solving $\boldsymbol{w}\mathbf{P}^* = \boldsymbol{w}$ and store its value.
Step 6.	Construct the ensemble mean based on weights w and calculate its RMSE_{Tr}.
Step 7.	Repeat Step 2–6 until L sets of weights $\boldsymbol{w}^1, \boldsymbol{w}^2, \ldots, \boldsymbol{w}^L$ and respective $\text{RMSE}_{\text{Tr}}{}^1, \text{RMSE}_{\text{Tr}}{}^2, \ldots, \text{RMSE}_{\text{Tr}}{}^L$ have been obtained.
Step 8.	Select a set of weights $\boldsymbol{w}^*$ corresponding to the minimal $\text{RMSE}_{\text{Tr}}{}^*$.
Step 9.	Construct the final E_{MCE} using the selected $\boldsymbol{w}^*$.

largest one. To optimize the properties of the simulations we control σ by randomly choosing it from the [0.1, 1] interval.

As we select only one of the simulations, the MCE method is not sensitive to the number of simulations L after a certain threshold. This threshold is set based on the requirements for precision of the results and on the calculation time. In Fig. 1 we illustrate the simulation performance dynamics (simulation index and performance on training and validation NAR-CLiM data) depending on the value of $L \in [1, 1\,000\,000]$. The simulation index $i^* \leq L$ represents the index of the best-performing simulation at each value of L (with $\boldsymbol{w}^*$ vector of weights and $\text{RMSE}_{\text{Tr}}^*$ as described in Step 8 of Table 2). The cross-validation procedure and RMSE metrics are described below in Sect. 2.5–2.6.

Though better RMSE results can be achieved with larger L, the marginal improvement in RMSE has a high computational time cost. For demonstration purposes in this study we select $L = 3000$ to accommodate for possible differences in RMSE changes between different datasets. As we will show below, even with a sub-optimal value of L, MCE method has high performance and stable results.

2.2.2 Model interdependence

While we do not claim that the proposed method explicitly addresses the issue of model dependence, it is implicitly addressed to some degree at Step 3 in Table 2 of the MCE method. If there are two or more highly correlated models only one of them can be chosen at each step, and thus the resulting sum of such models' weights will be close to the scenario when only one of those models is kept in the ensemble.

We demonstrate this property of the MCE method on modified NARCLiM data by adding a copy of one of the models with a small random error added and comparing the resulting weights as shown in Fig. 2. To mitigate difference in weight values between random simulations we repeat the calculation 100 times and compare the mean values of the weights.

As we can see from Fig. 2, adding a highly correlated ensemble member does not significantly change the weight distribution significantly, and more pleasingly when a high performing model is duplicated, the weights are shared between the two copies (see Model 3 and Model 9). Consequently the performance of E_{MCE} remains approximately the same. Though we can not guarantee this behavior in all types of
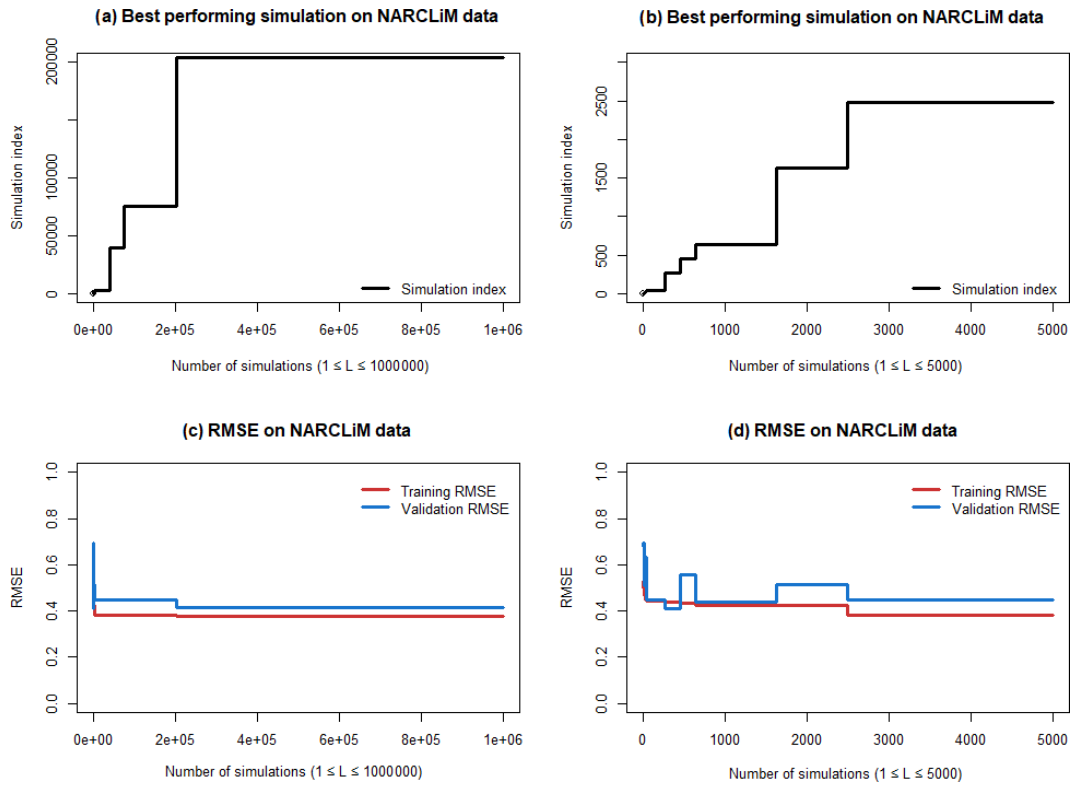
Figure 1. Sensitivity of the ensemble properties to the value of L. Left panels (**a**) and (**c**) contain results from all the simulations. Right panels (**b**) and (**d**) contain the results from the first 5000 simulations.
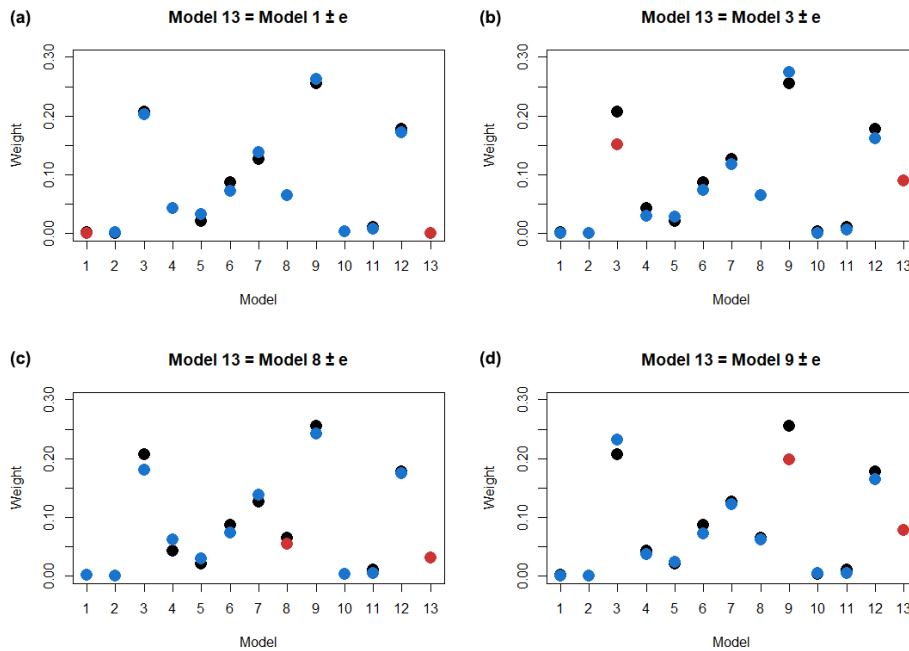
Figure 2. Change of MCE weights after adding a copy of Model 1, 3, 8 and 9 (clockwise from panel **a**) to the NARCliM ensemble. The original MCE weights are in black. The weights of the modified ensemble are in blue, and the weights of the highly correlated models are in red.

data, we believe that the MCE method's design helps to mitigate the model interdependence problem.

2.2.3 MCE method limitations

Though the MCE method can be used on any climate dataset which contains the required inputs, its relative performance differs depending on the properties of the dataset. We will demonstrate that in the case of normally distributed data, its performance is competitive with the simple averaging and other more sophisticated methods. In more challenging scenarios, when data are not normally distributed, MCE performs better than the common alternatives.

As the MCE method is based on a stochastic process, the results between runs can vary. To mitigate this effect and to have reproducible results we set the seed of R software's random number generator to a constant for all simulations. The MCE method in its current implementation does not provide an uncertainty quantification, and this limitation is a subject for future nonlinear ensemble weighting method development.

Finally, as the MCE method does not consider spatial information, the resulting weights have limited physical interpretability. Extending the MCE method to utilize such information is a subject for future research.

2.3 Multi-model ensemble average (AVE) method

In order to evaluate the relative performance of the MCE method we select two other popular approaches to constructing ensemble weighted average. The first approach is the widely used average of individual climate model outputs (Lambert and Boer, 2001; Gleckler et al., 2008):

$$E_{\text{AVE}_t} = 1/N \sum_{j=1}^{N} M_{j,t}, \tag{3}$$

for each $1 \le t \le T$. If model differences from observations are random and independent, they will cancel on averaging, and the resulting ensemble average will perform better than individual climate models (Lambert and Boer, 2001).

2.4 Convex optimization (COE) method

The second approach that has been selected for relative performance evaluation in this study is a convex optimization as proposed by Bishop and Abramowitz (2013). It represents a family of other methods based on a linear optimization over the vector space of individual climate model outputs.

The purpose of this method is to find a linear combination of climate model outputs with $w_1, w_2, \ldots, w_N$ weights which would minimize mean squared differences with respect to observations:

$$E_{\text{COE}_t} = \sum_{j=1}^{N} w_j M_{j,t}, \tag{4}$$

for each $1 \le t \le T$, so that $\sum_{t=1}^{T} (E_{\text{COE}_t} - O_t)^2$ is minimized under restrictions $\sum_{j=1}^{N} w_j = 1$ and $w_j \ge 0$ for each $1 \le j \le N$.

This method and its implementation are discussed in detail in Bishop and Abramowitz (2013), and we show that it has relatively high performance on the chosen datasets. However, like any other linear optimization technique, it naturally has some limitations that nonlinear optimizations like the MCE method do not. In particular, the COE method assumes there is a large enough sample size to rule out spurious fluctuations in the weights associated with a sample size that is too small. Such an assumption is not required for the MCE method. In addition, convex optimization tends to set a large portion of weights equal to 0, as is shown in the examples below, which results in a lower effective number of models used for prediction.

2.5 Performance metrics

2.5.1 RMSE

The root mean squared error (RMSE, Eq. 5) is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed. RMSE is positive, and a value of 0 indicates a perfect fit to the data. In general, a lower RMSE is better than a higher one. However, comparisons across different types of data would be invalid because the measure is dependent on the scale of the numbers used. Minimizing RMSE is commonly used for finding optimal ensemble weight vectors (e.g., Herger et al., 2018; Krishnamurti et al., 2000).

$$\text{RMSE} = \sqrt{\frac{1}{T} \sum_{t=1}^{T} \left(\sum_{j=1}^{N} w_j M_{j,t} - O_t \right)^2}, \tag{5}$$

with $\sum_{j=1}^{N} w_j = 1$ and $w_j > 0$ for $j = 1, \ldots, N$. T is the total number of time steps, $M_{j,t}$ denotes the value of model j at time step t and O_t is the observed value at point t.

2.5.2 Trend bias

The monthly trend bias is calculated as the difference between the inclination parameter a in weighted ensembles and observations estimated using a linear function $y = ax + b$ on validation data for each month. The total weighted ensemble trend bias metric is calculated as a mean of the monthly trend biases.

2.5.3 Climatology monthly bias

The monthly bias is calculated as the difference between the mean of the weighted ensemble and the observation for each month on validation data. The total climatology monthly bias metric is calculated as a mean of the monthly biases.

2.5.4 Interannual variability

Interannual variability for each month is calculated as the difference between the standard deviation of a detrended weighted ensemble and the standard deviations of detrended observations on validation data. The total interannual variability metric is calculated as the mean of interannual variability for each month.

2.5.5 Climatological monthly RMSE

Climatological RMSE is calculated according to Eq. (5) on climatological monthly means of weighted ensemble values and observations on validation data.

2.6 Cross-validation procedures

2.6.1 Holdout method

In this method the dataset, which contains the observations, is split into a training (or calibration) set and a validation (or testing) set. The goal of cross-validation is to examine the model's ability to predict new data that were not used in estimating the required parameters.

We partition our data into two sets, with 70 % of data used for training and 30 % for validation. This is a specific case of the K-fold validation procedure (Refaeilzadeh et al., 2009, p. 532–538), which is relatively simple to apply and discuss, facilitating the sharing of our findings with other members of the research and non-research communities.

2.6.2 Model-as-truth performance assessment

To evaluate each method's performance on the future model projections, we use the model-as-truth approach and analyze the metrics described in Sect. 2.5. At each step of model-as-truth performance assessment one model is selected as a true model (pseudo-observations) and the remaining models are used to build a weighted ensemble mean that best estimates the true model over the historical period. This weighted ensemble mean is then tested against the future projections of the true model. For a given ensemble this is repeated as many times as the number of the ensemble members with a different member being chosen as the true model each time. The median and spread of these results is reported.

3 Results

3.1 CMIP5 data

Though the selected monthly CMIP5 data contain annual variation, it is not predominant due to the length and trend of the dataset as shown in Fig. 3a. The CMIP5 model output distribution is close to normal as shown in Fig. 3b.

Applying the MCE method on the selected CMIP5 data with $T = 120$ (1900–2019) and a training period $T_1 = 80$ (1900–1979), we obtain a weighted ensemble mean E_{MCE} and compare it with outputs from other methods. We summarize CMIP5 data properties together with the resulting ensemble's weights in Fig. 3 and holdout cross-validation results in Table 3.

We can see that E_{AVE} and E_{MCE} perform at a similar RMSE level, with E_{COE} performance decreasing comparatively more in validation, a possible indication of overfitting to the training data. We can see from Fig. 3 that the COE method tends to set zero weights to some models but builds a weighted ensemble mean that performs best on the training period (1900–1979). Due to some models having zero weights, some of the models' diversity is lost, and this results in worse performance on the validation period (RMSE$_{\text{V}}$ and RMSE$_{\text{CM}}$ in Table 3). The MCE method, on the other hand, produces model weights that vary around $1/N$, where N is the number of the models. The MCE method does not give any model zero weighting and hence preserves the ensembles' diversity. The climatological biases B_{T}, B_{CM} and B_{IV} are nearly equal for all three methods.

The model-as-truth performance assessment is done on $T = 200$ (1900–2099) and a training period $T_1 = 120$ (1900–2019) as described in Sect. 2.6.2. The results are summarized in Fig. 4 and Table 4 in the form of median, 25 % and 75 % percentiles of the $N = 39$ (number of models) values.

All the methods perform similarly in model-as-truth assessment, with E_{COE} having better RMSE$_{\text{CM}}$.

3.2 NARCliM data

The seasonal variation in NARCLiM data is larger than in CMIP5 data as shown in Fig. 5a. The NARCLiM models output distribution is not normal as shown in Fig. 5b due to summer time and winter time temperature peaks.

We apply the MCE method on the selected NARCliM data with $T = 20$ (1990–2009) and a training period $T_1 = 14$ (1990–2003), obtain a weighted ensemble mean E_{MCE}, and compare it with outputs from other methods. We summarize NARCLiM data properties together with the resulting ensemble's weights in Fig. 5 and holdout cross-validation results in Table 5.

As in CMIP5 data analysis (Fig. 3), we see that the MCE method maintains (i.e., assigns non-zero weights to) more models in the final weighted ensemble than the COE method. As the number of models is significantly smaller than in the CMIP5 case, the difference between the MCE output weights and the equal weights is also considerably larger. The MCE method shows itself to be capable of maintaining much of the ensembles' diversity during the optimization process. This allows MCE to substantially improve performance over the AVE method on both training and validation periods and perform at the same level as COE on the validation period even with lower RMSE$_{\text{T}}$. Again, COE has a larger decline in performance from training to validation periods, indicating possible overfitting.
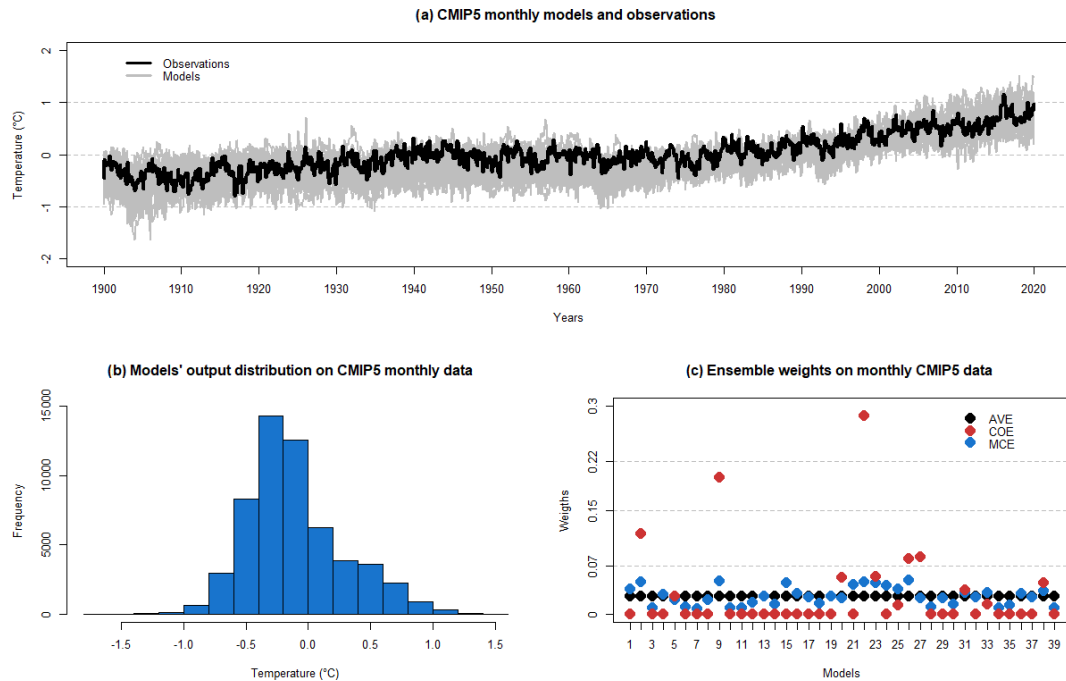
Figure 3. CMIP5 data properties. **(a)** Model outputs and observations. **(b)** Model output distribution. **(c)** AVE, COE and MCE weights.

Table 3. Performance comparison of different methods on CMIP5 data; RMSE on training ($RMSE_T$) and validation ($RMSE_V$) data; and trend bias (B_T), climatological monthly bias (B_{CM}), interannual variability bias (B_{IV}) and climatological monthly RMSE ($RMSE_{CM}$) on validation data.

Ensemble	$RMSE_T$	$RMSE_V$	B_T	B_{CM}	B_{IV}	$RMSE_{CM}$
E_{AVE}	0.22	0.17	0.01	-0.09	-0.05	0.10
E_{COE}	0.15	0.19	0.00	-0.12	-0.04	0.13
E_{MCE}	0.18	0.17	0.01	-0.10	-0.05	0.10

Table 4. Model-as-truth performance comparison of different methods on CMIP5 data, median of trend bias (B_T), climatological monthly bias (B_{CM}), interannual variability bias (B_{IV}) and climatological monthly RMSE ($RMSE_{CM}$) on validation data.

Ensemble	B_T	B_{CM}	B_{IV}	$RMSE_{CM}$
E_{AVE}	0.00	0.09	-0.09	0.25
E_{COE}	0.00	0.06	-0.07	0.17
E_{MCE}	0.00	0.10	-0.09	0.22

The model-as-truth performance assessment is done on $T = 30$ (1990–2019 and 2030–2039) and a training period $T_1 = 20$ (1990–2019) as described in Sect. 2.6.2. The results are summarized in Fig. 6 and Table 6 in form of median, 25 % and 75 % percentiles of the $N = 12$ (number of models) values.

As in the CMIP5 results (Fig. 4 and Table 4) all the methods perform at the same level in B_T and B_{CM} metrics of the model-as-truth assessment. In B_{IV} and $RMSE_{CM}$ met-

rics E_{COE} performs better, while E_{AVE} performs worse than E_{MCE}.

3.3 KMA data

The KMA data are non-negative with a non-normal distribution of model outputs and observations as shown in Fig. 7a and b.

Applying the MCE method on the selected data with $T = 33$ (1973–2005) and a training period $T_1 = 22$ (1973–1994), we obtain a weighted ensemble mean E_{MCE} and compare it with outputs from other methods. As KMA data contain only summertime months, we analyze only $RMSE_T$ and $RMSE_V$. We summarize KMA data properties together with the resulting ensemble's weights in Fig. 7 and holdout cross-validation results in Table 7.

We can see that MCE has the lowest RMSE and maintains the ensembles' diversity with a few models receiving zero weights. The COE method gives non-zero weights to only a small subset of models, which results in its performance on the validation period being lower compared to MCE.
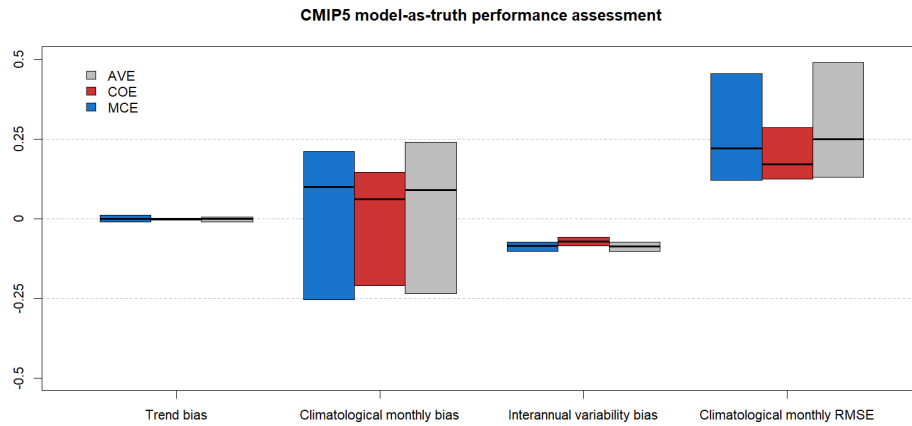
Figure 4. CMIP5 model-as-truth performance assessment results. Median, 25 % and 75 % percentiles of $N = 39$ models.
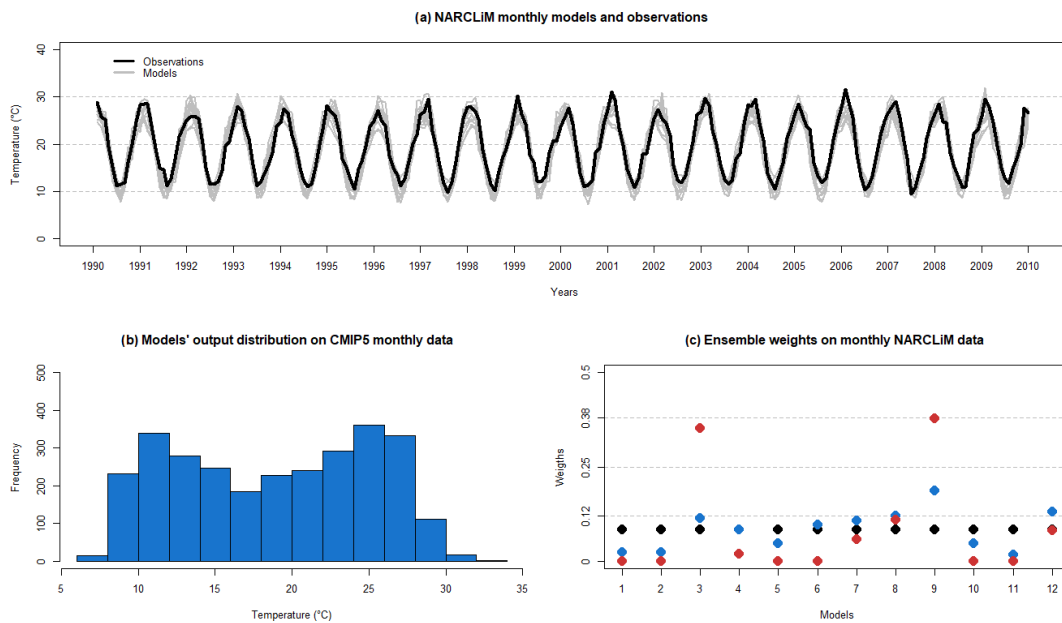

Figure 5. NARCLiM data properties. **(a)** Model outputs and observations. **(b)** Model output distribution. **(c)** AVE, COE and MCE weights.

4 Discussion

The obtained results indicate that Markov chains can be used to construct a better performing weighted ensemble mean with lower RMSE on validation data than commonly used methods like multi-model ensemble averaging and convex optimization (Tables 3, 5 and 7). As the method's performance did not degrade from training to validation as much as COE, we are confident that it is less prone to over-fitting than linear optimization methods. We attribute this advantage of the MCE method to its ability to maintain the ensemble's diversity while optimizing its weights on the training period (Figs. 3, 5 and 7), to mitigate model interdependence and to capture some of the nonlinear patterns in the data.

The MCE method also performs at the same level as other methods in terms of climatological metrics and model-as-
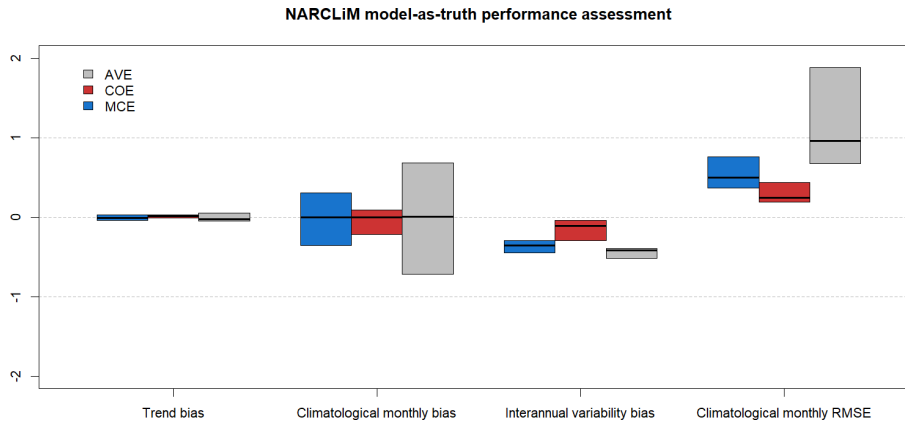
truth performance assessment, which gives us confidence in its ability to be used for future estimation of climate variables.

However, as previous studies show (e.g., Mason and Knutti, 2011; Sanderson et al., 2017) and as discussed in Sect. 2.2.3, extending the MCE method to include spatial information would improve our ability to interpret the physical meaning of the resulting weights.

As the number of models increases, MCE tends to become closer to AVE weights (Fig. 3), while being closer to COE with a smaller number of models (Fig. 7). This phenomenon can be explained by a higher effect of diversity on performance in larger ensembles with normally distributed data (observations and model outputs) than in smaller ensembles like NARCLiM. The KMA data have an intermediate number of models and MCE produces a hybrid response

Table 5. Performance comparison of different methods on NARCLiM data; RMSE on training ($RMSE_T$) and validation ($RMSE_V$) data; and trend bias (B_T), climatological monthly bias (B_{CM}), interannual variability bias (B_{IV}) and climatological monthly RMSE ($RMSE_{CM}$) on validation data.

Ensemble	$RMSE_T$	$RMSE_V$	B_T	B_{CM}	B_{IV}	$RMSE_{CM}$
E_{AVE}	1.6	1.85	0.04	−1.16	−0.73	1.19
E_{COE}	1.32	1.58	0.00	−0.49	−0.59	0.64
E_{MCE}	1.4	1.58	0.04	−0.64	−0.69	0.70

Figure 6. NARCLiM model-as-truth performance assessment results. Median, 25 % and 75 % percentiles of the $N = 12$ models.

Table 6. Model-as-truth performance comparison of different methods on NARCLiM data, median of trend bias (B_T), climatological monthly bias (B_{CM}), interannual variability bias (B_{IV}) and climatological monthly RMSE ($RMSE_{CM}$) on validation data.

Ensemble	B_T	B_{CM}	B_{IV}	$RMSE_{CM}$
E_{AVE}	−0.03	0.01	−0.42	0.96
E_{COE}	0.01	−0.01	−0.11	0.24
E_{MCE}	−0.02	−0.01	−0.36	0.50

Table 7. Performance comparison of different methods on KMA data, RMSE on training ($RMSE_T$) and validation ($RMSE_V$) data.

Ensemble	$RMSE_T$	$RMSE_V$
E_{AVE}	0.36	0.5
E_{COE}	0.23	0.52
E_{MCE}	0.29	0.44

which maintains ensemble diversity (a few models with zero weights) but does weight a small number of models more highly.

The MCE method is computationally cheap and is limited only by a software's ability to handle extreme numerical values. One limitation of the MCE method is its current inability to quantify the uncertainty of the resulting weighted ensemble mean. However, we believe that given the stochastic na-

ture of the method, this limitation can be overcome in future implementations. MCE performance can be further improved by combining it with other types of optimization, e.g., linear. In addition, other nonlinear optimization techniques, which would include more complex structures than simple Markov chains, can be developed based on our demonstrated results.

Finally, the MCE method does not require some of the assumptions necessary for the multi-model ensemble average method (e.g., models being reasonably independent and equally plausible as discussed by Knutti et al., 2017), and it does not produce as many zero weights as the convex optimization method, hence maintaining more of the models' diversity. We attribute the tendency of the COE method to set zero weights to some models to its property below.

Geometrically, the restrictions $w_j \geq 0, \sum_{j=1}^{N} w_j = 1$ describe a simplex in R^N that is a subset of the hyperplane with the equation $\sum_{j=1}^{N} w_j = 1$. Denote $w = (w_1, w_2, \ldots, w_N)$. The potential choice of weights that only satisfy the constraint $\sum_{j=1}^{N} w_j = 1$ without the non-negativity restriction represents any point in the hyperplane $P = \{w : \sum_{j=1}^{N} w_j = 1\}$. This hyperplane contains the simplex $S = \{w \in P : w_j \geq 0\}$. In general, the optimal point $\boldsymbol{w}^*$ for the unrestricted solution of the optimization problem

$$\min_{w} \sum_{i=1}^{T} \left(\sum_{j=1}^{N} w_j M_{j,i} - O_i \right)^2, w \in P$$
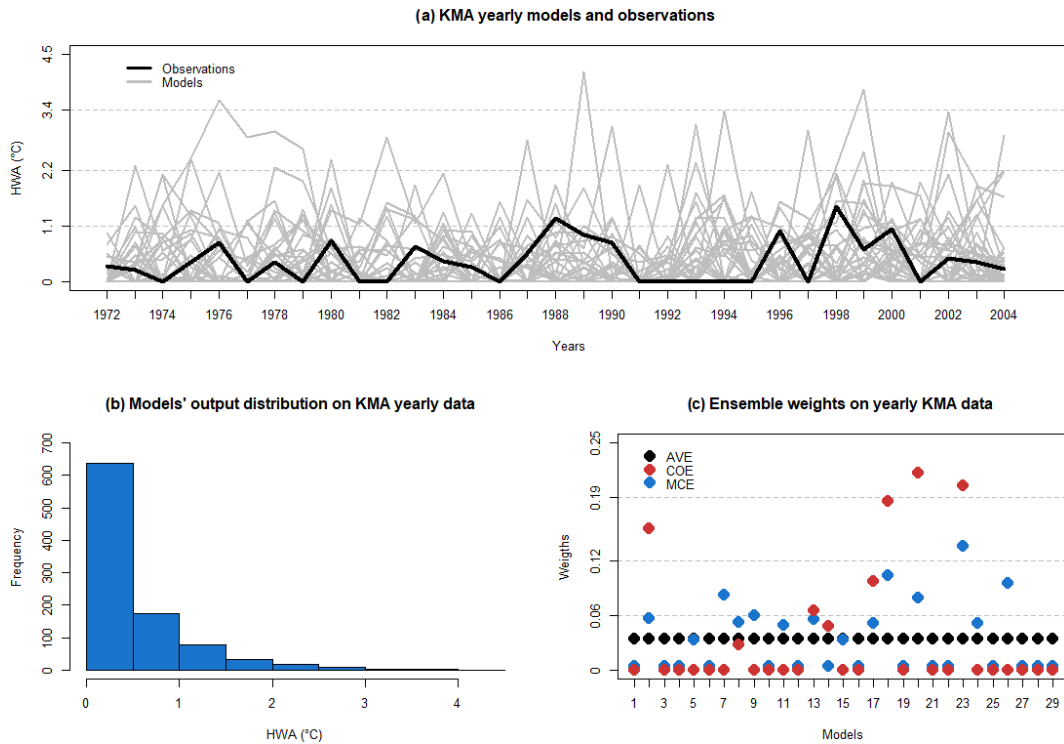
Figure 7. KMA data properties. **(a)** Model outputs and observations. **(b)** Model output distribution. **(c)** AVE, COE and MCE weights.

will be outside the simplex. It is clear that the optimal point for the constrained solution on the simplex:

$$\min_{w} \sum_{i=1}^{T} \left(\sum_{j=1}^{N} w_j M_{j,i} - O_i \right)^2 , w \in S$$

would be on the boundary of the simplex rather than in its interior. Indeed, if we assume that the optimal point for the constrained problem is a certain $\tilde{w}$ in the interior of the simplex, we immediately arrive at a contradiction. Take then the point $\hat{w} = \boldsymbol{w}^* + \lambda(\tilde{w} - \boldsymbol{w}^*)$ with $\lambda \in (0, 1)$ chosen such that $\hat{w}$ is on the intersection of the line connecting $\boldsymbol{w}^*$ and $\tilde{w}$ with the boundary of the simplex. Because of the strict convexity of the function

$$f(w) = \sum_{i=1}^{T} \left(\sum_{j=1}^{N} w_j M_{j,i} - O_i \right)^2 ,$$

we have

$$f(\hat{w}) = f(\boldsymbol{w}^* + \lambda(\tilde{w} - \boldsymbol{w}^*))$$
$$= f(\lambda \tilde{w} + (1 - \lambda)\boldsymbol{w}^*) < \lambda f(\tilde{w})$$
$$+ (1 - \lambda)f(\boldsymbol{w}^*) < f(\tilde{w}),$$

in contradiction to the assumption that $\tilde{w}$ delivers the minimum over the simplex. Hence the optimization on the simplex tends to deliver optimal points with some components equal to zero because they tend to be on the boundary of the simplex.

5 Conclusions

In this study, we presented a novel approach based on Markov chains to estimate model weights in constructing weighted climate model ensemble means. The complete MCE method was applied to selected climate datasets, and its performance was compared to two other common approaches (AVE and COE) using a cross-validation hold-out method and model-as-truth performance assessment with RMSE, trend bias, climatology monthly bias, interannual variability and climatological monthly RMSE metrics. The MCE method was discussed in detail, and its step-wise implementation, including mathematical background, was presented (Table 2).

The results of this study indicate that applying nonlinear ensemble weighting methods on climate datasets can improve future climate projection in terms of accuracy. Even a simple nonlinear structure such as Markov chains shows good performance on different commonly used datasets compared to linear optimization approaches. These results are supported by using standard performance metrics, cross-validation procedures and model-as-truth performance assessment. The developed MCE method is objective in terms of parameter selection, has a sound theoretical basis and has a relatively low number of limitations. It maintains ensemble diversity, mitigates model interdependence and captures some of the nonlinear patterns in the data while optimizing ensemble weights. It is also shown to perform well on non-

Gaussian datasets. Based on the above, we are confident in suggesting its application on other datasets and its usage for the future development of new nonlinear optimization methods for weighting climate model ensembles.

References

Abramowitz, G., Herger, N., Gutmann, E., Hammerling, D., Knutti, R., Leduc, M., Lorenz, R., Pincus, R., and Schmidt, G. A.: ESD Reviews: Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing, Earth Syst. Dynam., 10, 91–105, https://doi.org/10.5194/esd-10-91-2019, 2019.

Bai, J. and Wang, P.: Conditional Markov chain and its application in economic time series analysis, J. Appl. Econ., 26, 715–734, https://doi.org/10.1002/jae.1140, 2011.

Bishop, C. H. and Abramowitz, G.: Climate model dependence and the replicate Earth paradigm, Clim. Dynam., 41, 885–900, https://doi.org/10.1007/s00382-012-1610-y, 2013.

Del Moral, P. and Penev, S.: Stochastic Processes. From Applications to Theory, p. 121, Taylor and Francis Group, Boca Raton, 2016.

Evans, J. P., Ji, F., Lee, C., Smith, P., Argüeso, D., and Fita, L.: Design of a regional climate modelling projection ensemble experiment – NARCLiM, Geosci. Model Dev., 7, 621–629, https://doi.org/10.5194/gmd-7-621-2014, 2014.

Fan, Y., Olson, R., and Evans, J. P.: A Bayesian posterior predictive framework for weighting ensemble regional climate models, Geosci. Model Dev., 10, 2321–2332, https://doi.org/10.5194/gmd-10-2321-2017, 2017.

Feng, J., Lee, D., Fu, C., Tang, J., Sato, Y., Kato, H., Mcgregor, J., and Mabuchi, K.: Comparison of four ensemble methods combining regional climate simulations over Asia, Meteorol. Atmos. Phys., 111, 41–53, https://doi.org/10.1007/s00703-010-0115-7, 2011.

Fischer, E. and Schär, C.: Consistent geographical patterns of changes in high-impact European heatwaves, Nat. Geosci., 3, 398–403, https://doi.org/10.1038/ngeo866, 2010.

Gleckler, P. J., Taylor, K. E., and Doutriaux, C.: Performance metrics for climate models, J. Geophys. Res., 113, D06104, https://doi.org/10.1029/2007JD008972, 2008.

Herger, N., Abramowitz, G., Knutti, R., Angélil, O., Lehmann, K., and Sanderson, B. M.: Selecting a climate model subset to optimise key ensemble properties, Earth Syst. Dynam., 9, 135–151, https://doi.org/10.5194/esd-9-135-2018, 2018.

Jones, D., Wang, W., and Fawcett, R.: High-quality spatial climate data-sets for Australia, Aust. Meteorol. Ocean., 58, 233–248, https://doi.org/10.22499/2.5804.003, 2009.

Kharin, V. V. and Zwiers, F. W.: Climate Predictions with Multimodel Ensembles, J. Climate, 15, 793–799, https://doi.org/10.1175/1520-0442(2002)015<0793:CPWME>2.0.CO;2, 2002.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., and Meehl, G. A.: Challenges in Combining Projections from Multiple Climate Models, J. Climate, 23, 2739–2758, https://doi.org/10.1175/2009JCLI3361.1, 2010.

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, https://doi.org/10.1002/2016GL072012, 2017.

Krishnamurti, T. N., Kishtawal, C. M., Zhang, Z., LaRow, T., Bachiochi, D., Williford, E., Gadgil, S., and Surendran, S.: Multimodel Ensemble Forecasts for Weather and Seasonal Climate, J. Climate, 13, 4196–4216, https://doi.org/10.1175/1520-0442(2000)013<4196:MEFFWA>2.0.CO;2, 2000.

Kulinich, M.: Code and data for "A Markov chain method for weighting climate model ensembles" publication (Version 1), Zenodo [code and data set], https://doi.org/10.5281/zenodo.4548417, 2020.

Lambert, S. and Boer, G.: CMIP1 evaluation and intercomparison of coupled climate models, Clim. Dynam., 17, 83–106, https://doi.org/10.1007/PL00013736, 2001.

Leduc, M., Laprise, R., de Elía, R., and Šeparović, L.: Is institutional democracy a good proxy for model independence?, J. Climate, 29, 8301–8316, 2016.

Majumder, S., Balakrishnan Nair, T. M., Sandhya, K. G., Remya, P. G., and Sirisha, P.: Modification of a linear regression-based multi-model super-ensemble technique and its application in forecasting of wave height during extreme weather conditions, J. Oper. Oceanogr., 11, 1–10, https://doi.org/10.1080/1755876X.2018.1438341, 2018.

Masson, D. and Knutti, R.: Climate model genealogy, Geophys. Res. Lett., 38, L08703, https://doi.org/10.1029/2011GL046864, 2011.

Murphy, J., Sexton, D., Barnett, D., Jones, G., Webb, M., Collins, M., and Stainforth, D.: Quantification of modelling uncertainties in a large ensemble of climate change simulations, Nature, 430, 768–772, https://doi.org/10.1038/nature02771, 2004.

Olson, R., Evans, J., Di Luca, A., and Argueso, D.: The NARCliM project: Model agreement and significance of climate projections, Clim. Res., 69, 209–227, https://doi.org/10.3354/cr01403, 2016.

Olson, R., An, S.-I., Fan, Y., Chang, W., Evans, J. P., and Lee, J.-Y.: A novel method to test non-exclusive hypotheses applied to Arctic ice projections from dependent models, Nat. Commun., 10, 3016, https://doi.org/10.1038/s41467-019-10561-x, 2019.

Pesch, T., Schröders, S., Allelein, H. J., and Hake, J. F.: A new Markov-chain-related statistical approach for modelling synthetic wind power time series, New J. Phys., 17, 055001, https://doi.org/10.1088/1367-2630/17/5/055001, 2015.

R Core Team: R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, available at: http://www.R-project.org/ (last access: 11 May 2020), 2013.

Refaeilzadeh, P., Tang, L., and Liu, H.: Cross-Validation, Encyclopedia of Database Systems, 532–538, 532–538, https://doi.org/10.1007/978-0-387-39940-9_565, 2009.

Sanderson, B. M., Wehner, M., and Knutti, R.: Skill and independence weighting for multi-model assessments, Geosci. Model Dev., 10, 2379–2395, https://doi.org/10.5194/gmd-10-2379-2017, 2017.

Shin, J., Olson, R., and An, S.-I.: Projected Heat Wave Characteristics over the Korean Peninsula During the Twenty-First Century, Asia-Pac, J. Atmos. Sci., 54, 1–9, https://doi.org/10.1007/s13143-017-0059-7, 2017.

Steinschneider, S., McCrary, R., Mearns, L. O., and Brown, C.: The effects of climate model similarity on probabilistic climate projections and the implications for local, risk-based adaptation planning, Geophys. Res. Lett., 42, 5014–5022, 2015.

Taylor, K. E., Stouffer, R., and Meehl, G.: An overview of CMIP5 and the Experiment Design, B. Am. Meteorol. Soc., 93, 485–498, https://doi.org/10.1175/BAMS-D-11-00094.1, 2011.