https://doi.org/10.5194/gmd-2020-253
Preprint. Discussion started: 1 October 2020

Geoscientific Model Development Discussions
Open Access
EGU

A Markov chain method for weighting climate model ensembles

Max Kulinich[1], Yanan Fan[2], Spiridon Penev[3], and Jason P. Evans[4]

[1,2,3]School of Mathematics and Statistics, UNSW Sydney, Australia.
[4]Climate Change Research Centre and ARC Centre of Excellence for Climate Extremes, UNSW Sydney, Australia

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

Abstract. Climate change is typically modelled using sophisticated mathematical models (Climate Models) of physical processes taking place over long periods of time. Multi-model ensembles 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 in an optimal way. We present a novel 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. The Markov chain method showed improved performance over those methods when measured by the root mean squared error and the R-squared metrics. The results of this comparative analysis should serve to motivate further studies in Markov chain and other nonlinear methods application, that address the issues of finding optimal model weight for constructing ensemble means.

1 Introduction

Climate change is often modelled using sophisticated mathematical models of physical processes taking place over long periods of time. These models are inherently limited in their ability to represent all aspects of the modelled physical processes. Simple averages of multi-model 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) points 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 be at least partly cancelled. 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 employ linear optimisation techniques (Krishnamurti et al. (2000); Majumder et al. (2018); Abramowitz et al. (2018)) 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 optimisation approach. An unresolved issue in using weights for models is that models have interdependence, due to the sharing of codes, sharing of initialisation parameters, etc. Abramowitz et al. (2018) points out that model dependence can play a crucial role when assembling the models into an ensemble. If highly dependent models are included in an ensemble mean with equal

https://doi.org/10.5194/gmd-2020-253
Preprint. Discussion started: 1 October 2020


Geoscientific Model Development Discussions

weights as independent models, the overall ensemble mean will become close to the dependent models' cluster of values if the cluster is large.

Hence, it is desirable that an ensemble weighting method is robust against the dependency issue, and has normalised non-negative weights for interpretability. Finally, the methods should work well across a range of different data types, such as temperature, precipitation and other kinds of climate variables. 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 optimisation on a vector space of climate models' outputs allows building better performing weighted ensembles. We selected Markov chains as a basis for such nonlinear optimisation as one of the most straightforward nonlinear structures. It naturally produces non-negative weights that sum to one. It performs well on a range of data sets when compared to the standard simple mean and linear optimisation weighting methods as we demonstrate below. While it does not directly address the issue of dependence, it can mitigate some of the problems faced having dependent models in an ensemble.

Although Markov chains have been used frequently in the literature for 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 data sets used in this study and the proposed MCE method in Section 2. We compare the proposed method (MCE) to the commonly used multi-model ensemble average (AVE) method (Lambert and Boer (2001)) and the convex optimisation (COE) method proposed by Bishop and Abramowitz (2013) to explicitly address the issue of dependent models and present the results in Section 3, followed by a discussion in Section 4 and conclusion in Section 5.

2 Methods

2.1 Data

Here we first describe the data sets used in this study. We have chosen three publicly available data sets with differing numbers of models, historical period lengths and model interdependence levels to evaluate and compare the performance of the MCE method with alternative approaches.

CMIP5 Data: The first data set we use is the temperature anomalies (°C) data from Coupled Model Intercomparison Project (CMIP5) with 80 different Global Climate Model (GCM) outputs and Hadley Centre/Climatic Research Unit Temperature observations (HadCRUT4). The data is obtained from https://climexp.knmi.nl and the period of 1900 - 2019 is selected for the analysis. It contains temperature anomalies (yearly averages) compared to the reference period of 1961-1990 and is commonly

https://doi.org/10.5194/gmd-2020-253
Preprint. Discussion started: 1 October 2020


Geoscientific Model Development Discussions
Open Access
EGU

used for different research projects (Taylor et al. (2011)). This data set contains several clusters of dependent models, has both positive and non-positive data values, has a low variability and relatively long time series, with a large ensemble.

NARCliM Data: The second data set contains temperature output from the New South Wales and Australian Capital Territory Regional Climate Modelling project (Evans et al. (2014)). It contains regional climate model (RCM) simulations over southeastern Australia. The data contains seasonal–mean temperature (°C) as modelled by the NARCliM domain regional climate models (RCMs) between the years 1990–2008 (Olson et al. (2016)). The temperatures are averaged over southeastern Australian regions which include New South Wales (NSW) planning regions, Australian Capital Territory (ACT), and Victoria. Corresponding temperature observations are obtained from the Australian Water Availability Project (AWAP) (Jones et al. (2009)). The data set 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 inter-dependence) in small clusters.

KMA Data: The third data set contains data from 59 weather stations operated by the Korea Meteorological Administration (KMA) and includes summertime (June, July, and August) observed daily maximum temperatures between 1973 and 2005, together with historical model output from 29 CMIP5 models, Shin et al. (2017) provides the list of models included in the study. The projected heatwave characteristics were then calculated from the temperature data. In particular, heatwave amplitude (HWA) contains the difference between the highest temperature during the heatwave events and the 95th percentile of reference 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 reference summer temperatures (32.82°C) for two consecutive days. HWA is non-negative and can be highly skewed with long upper tails as it measures extremal events; therefore, the data set is highly non-Gaussian (see Figure 1). These properties allow us to test methods in a more challenging scenario, where likelihood-based approaches are more difficult to apply.

These three data sets cover different scenarios, data structures, parameter distributions and scales, and we summarise these in Figure 1. Such coverage allows us to analyse the performance and the inherent limitations of the proposed method.

https://doi.org/10.5194/gmd-2020-253
Preprint. Discussion started: 1 October 2020

Geoscientific Model Development Discussions

Data	Measurement	Number of observations	Range	Variance	Number of models
CMIP5	Temperature anomaly (°C)	120	From -0.47 to 0.83	0.09	80
NARCliM	Temperature (°C)	19	From 25.58 to 29.44	1.34	12
KMA	HWA (°C)	33	From 0 to 1.49	0.17	29

Figure 1. Histogram densities of the observational data (top) in the chosen data sets, CMIP5 (left), NARCliM (middle), KMA (right). The corresponding correlation matrices (middle) and a summary of the properties of the three data set used in this study.

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 P (with $P(x,y) \in [0,1]$ and $\sum_y P(x,y) = 1$, for any $x, y \in S$):

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

In this study, we want to utilise the "Fundamental Limit Theorem for Regular Chain" which states that if P is a transition matrix for a regular Markov chain (where $\forall x, y \in S$, $P(x,y) > 0$), then

$$\lim_{n \to \infty} P^n = P^\infty$$

https://doi.org/10.5194/gmd-2020-253
Preprint. Discussion started: 1 October 2020

Geoscientific Model Development Discussions

where P^{∞} is a matrix with all rows being equal and having strictly positive entries.

This property allows us to construct a non-negative transition matrix P by distillation of input information (i.e., model outputs and historical observations) and allows P to converge to a unique vector of model weights $w = (w_1, w_2, ..., w_N)$, where N is a total number of models in a given ensemble. Vector w can be obtained by solving the equation $wP = w$. 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 chapters.

More precisely, we start by constructing a transition vector v (based on the input data) which specifies a choice of the optimal model at any given time step t. Using vector v we construct a transition matrix P and find its stationary distribution w. The resulting weighted ensemble mean is constructed by applying w on the given (future) 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 1 gives a step by step description of the MCE algorithm.

Input:

- length of training period T_1, and
- historical observations at O_t, at times $t = 1, \ldots, T_1$, and
- climate model output $M_{i,t}$, at times $t = 1, \ldots, T_1$, for $i = 1, \ldots, N$ models, and
- an initialised transition matrix P^0 of N x N size, and
- control parameters σ and L

Step 1. Compute the distance matrix D according to Equation 2.

Step 2. Construct a sequence vector v based on D using stochastic simulations.

Step 3. Update P^0 step-wise by increasing probability of transitions contained in v : $P^0 \rightarrow P^1 \rightarrow ... \rightarrow P^{T_1}$.

Step 4. Obtain normalised transition matrix P^*, by normalising P^{T_1} row-wise so that each row sums to 1.

Step 5. Find w by solving $wP^* = w$ and store its value.

Step 6. Repeat Step 2 - 5 until L sets of weights $w^1, w^2, ..., w^L$ have been obtained.

Step 7. Calculate average $\overline{w} = \sum_{l=1}^{L} w_l / L$ and use it to construct the ensemble mean.

Table 1. The Markov Chain Ensemble (MCE) algorithm

We provide some details of the algorithm as described in Table 1 in the following paragraph.

https://doi.org/10.5194/gmd-2020-253
Preprint. Discussion started: 1 October 2020

Geoscientific Model Development Discussions

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

Step 1: The MCE method proceeds by utilising 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 D of size $N \times T_1$, using a normalised exponential function.

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

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, $1 \leq i \leq N$ where N is the number of models included, and σ is a control factor, which determines the degree to which models with larger distances are included. The control factor σ is selected by optimising MCE performance on training data as described below in section 2.2.1.

Step 2: Based on the matrix 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 $d_{i,k}$. This way we construct a vector $V = (v_1, v_2, v_3, ..., v_{T_1})$, which represents choice of models closest to observations at each time step.

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

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

Step 5: The stationary distribution w is obtained by solving $wP^* = w$. A standard R software package is used to find the solution in this study.

Step 6: Steps 2 - 5 are repeated L times, where L is selected by optimising MCE performance on training data as described below in section 2.2.1.

Step 7: The average $\overline{w} = \sum_{l=1}^{L} w_l / L$ is used to construct the weighted ensemble mean.

2.2.1 Choosing control parameters

We select control parameters σ and L by running MCE on training data and analysing its performance with different values of these parameters. The best performing set of parameters is chosen to construct the final MCE output.

In order to be able to apply MCE, we set a reasonably high initial $L = 300$, run the MCE algorithm on different $\sigma \in [0.005, 0.5]$, and choose the best performing σ from this range. The optimal σ is achieved when a clear trend of lower and higher σ values with lower R-squared is observed. Then we use this optimal σ to find $L \in [50, 1000]$, which would give relatively high performance and low computational time.

The interval for choosing σ can vary depending on the data set to accommodate computational limitations. In addition, we want to have at least $L = 300$ simulations to have robust results. However, the MCE method is robust against variations of

https://doi.org/10.5194/gmd-2020-253
Preprint. Discussion started: 1 October 2020


Geoscientific Model Development Discussions
Open Access
EGU

both σ and L as we show below, hence finding the exact optimal set of parameters is not necessary for the purpose of our demonstration.

2.2.2 MCE method limitations

Though the MCE method can be used on any climate data set, which contains the required inputs, its relative performance differs depending on the properties of the data set. We will demonstrate that in the case of a standard normally distributed data, its performance is slightly better than simple averaging and other more sophisticated methods. In more challenging scenarios, when data is not normally distributed or has a limited number of data points, MCE is performing significantly better than the common alternatives.

There are only two arbitrary tuning parameters: σ and L, both of them can be chosen by analysing MCE performance on training data sets. Though this approach does not guarantee that the chosen parameters are optimal on a validation set as well, it does provide an easy and objective way to make these choices, while keeping the MCE method performance high.

The MCE method in its current implementation does not provide an uncertainty quantification, and this limitation is a subject for future nonlinear ensemble weighting methods development.

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 a widely used average of individual climate model outputs (Lambert and Boer (2001); Gleckler et al. (2008)):

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

for each $1 \leq t \leq T$. 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 optimisation (COE) method

The second approach that has been selected for relative performance evaluation in this study is a convex optimisation as proposed by Bishop and Abramowitz (2013). It represents a family of other methods based on a linear optimisation 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 $w_1, w_2, ..., w_N$ weights which would minimise mean squared differences with respect to observations:

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

for each $1 \leq t \leq T$, so that $\sum_{t=1}^{T} (E_{COE_t} - O_t)^2$ is minimised under restrictions $\sum_{j=1}^{N} w_j = 1$ and $w_j \geq 0$ for each $1 \leq j \leq N$.

https://doi.org/10.5194/gmd-2020-253
Preprint. Discussion started: 1 October 2020


Geoscientific Model Development Discussions

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

2.5 Performance metrics

We consider the following performance metrics in this paper when comparing competing methods.

RMSE

The root mean squared error (RMSE), Equation 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 always non-negative, and a value of 0 would indicate 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. Minimising RMSE is commonly used for finding optimal ensemble weight vectors (e.g. Herger et al. (2018); Krishnamurti et al. (2000)).

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

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.

R-squared

R-squared is a statistical measure of how close the observational data are to the fitted estimator. It represents the proportion of the variance in the dependent variable (weighted ensemble mean) that is predictable from the independent variables (GCMs).

The definition of R-squared used here is

$$R^2 = 1 - \frac{\sum_{t=1}^{T}(\sum_{j=1}^{N} w_j M_{j,t} - O_t)^2}{\sum_{t=1}^{T}(O_t - \overline{O})^2} \tag{6}$$

It is important to note that this definition of R-squared allows its values to lay within a range of $(-\infty, 1)$, and a high value of R^2 means high performance.

https://doi.org/10.5194/gmd-2020-253
Preprint. Discussion started: 1 October 2020

Geoscientific Model Development Discussions

2.6 Cross-validation procedures

In a future prediction problem, a method is usually given a data set of known data on which training is run (training data set), and a data set of unknown data against which the model is tested (called validation data set or testing data set). The goal of cross-validation is to examine the model's ability to predict new data that was 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 K-fold validation (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.

3 Results

3.1 CMIP5 data

We select σ and L by analysing MCE performance on training data set as described in section 2.2.1 and shown in Figure 2.

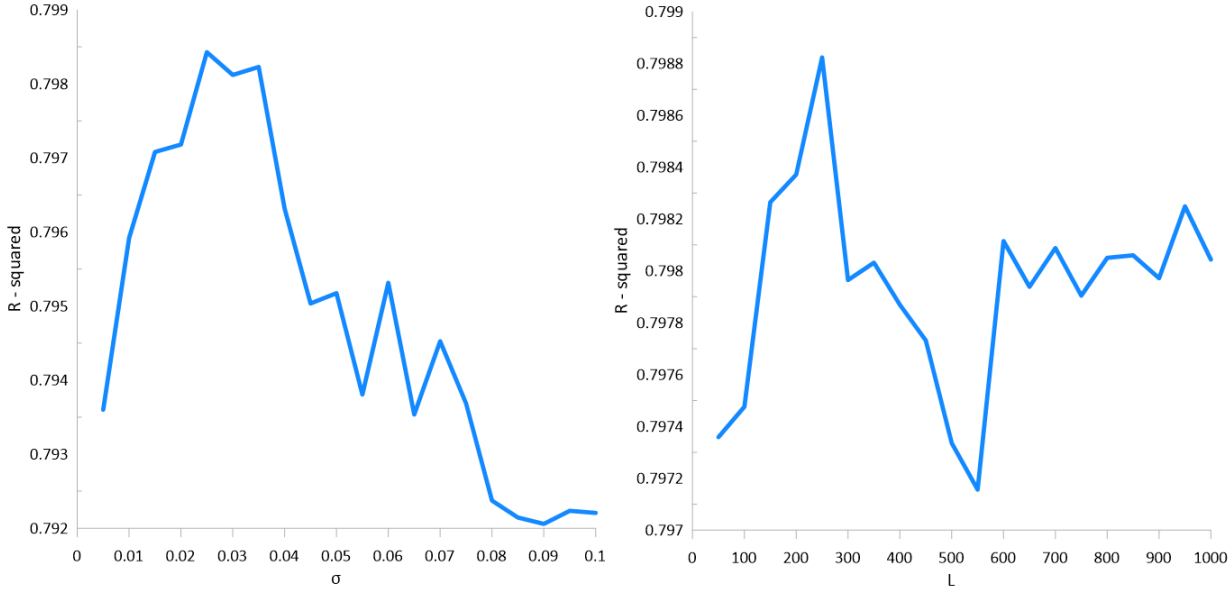

Figure 2. MCE method performance on training data for CMIP5 data depending on L with σ ($L = 300$ on the left hand side and $\sigma = 0.025$ on the right hand side)

Applying the MCE method on the selected CMIP5 data with $T = 120(1900 - 2019)$, $\sigma = 0.025$, $L = 600$ (as described in section 2.2.1) 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:

https://doi.org/10.5194/gmd-2020-253
Preprint. Discussion started: 1 October 2020

Geoscientific Model Development Discussions

$Ensemble$	$RMSE_T$	R^2_T	$RMSE_V$	R^2_V
E_{AVE}	0.11	0.78	0.15	0.87
E_{COE}	**0.08**	**0.87**	0.15	0.87
E_{MCE}	0.10	0.80	**0.14**	**0.89**

Table 2. Performance comparison of different methods on CMIP5 data, RMSE on training ($RMSE_T$) and validation ($RMSE_V$) data; R^2 on training (R^2_T) and validation (R^2_V) data.

We can see that all methods perform at a similar level, with the MCE method having a slight advantage. We analyse this difference in methods' performance by looking at weighted ensembles and model weights for each method:

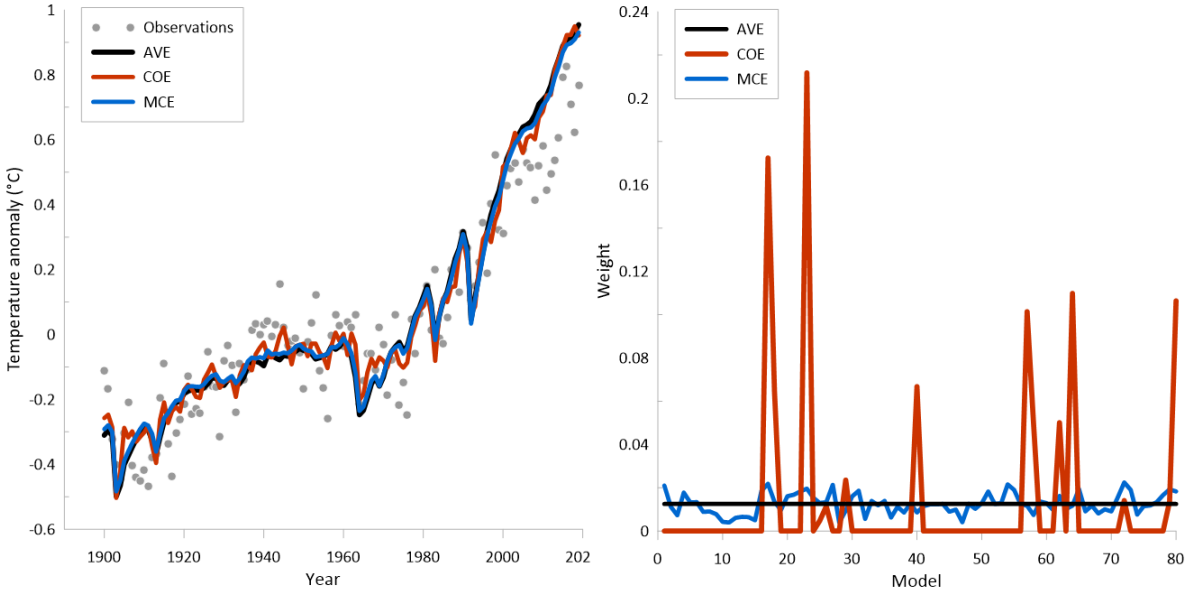

Figure 3. Estimated weighted ensembles (left panel) and weight distributions (right panel), comparison of different methods on CMIP5 data.

We can see from Figure 3 (right panel) 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 (Table 2). 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.

https://doi.org/10.5194/gmd-2020-253
Preprint. Discussion started: 1 October 2020

Geoscientific Model Development Discussions

3.2 NARCliM data

We select σ and L by analysing MCE performance on training data set as described in section 2.2.1 and shown in Figure 4.

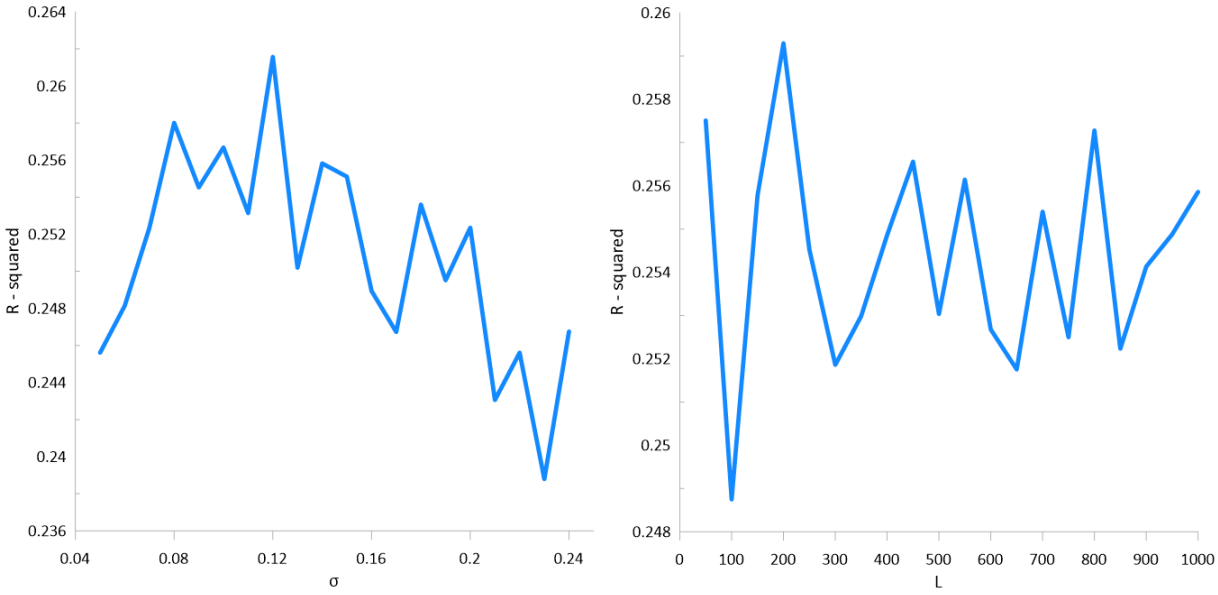

Figure 4. MCE method performance on training data for NARCliM data depending on L with σ ($L = 300$ on the left hand side and $\sigma = 0.12$ on the right hand side).

We apply the MCE method on the selected NARCliM data with $T = 19(1990 - 2008)$, $\sigma = 0.12$, $L = 800$ and a training period $T_1 = 13(1990 - 2002)$, we obtain a weighted ensemble mean E_{MCE} and compare it with outputs from other methods:

$Ensemble$	$RMSE_T$	R^2_T	$RMSE_V$	R^2_V
E_{AVE}	1.36	-0.20	1.86	-1.45
E_{COE}	**0.85**	**0.53**	1.59	-0.78
E_{MCE}	1.07	0.25	**1.49**	**-0.57**

Table 3. Performance comparison of different methods on NARCliM data. RMSE on training ($RMSE_T$) and validation ($RMSE_V$) data; R^2 on training (R^2_T) and validation (R^2_V) data.

Though all three methods have negative R-squared value on the validation data, MCE is still performing slightly better than the other methods. We analyse this difference in the methods' performance by looking at weighted ensembles and model weights for each method:

https://doi.org/10.5194/gmd-2020-253
Preprint. Discussion started: 1 October 2020

Geoscientific Model Development Discussions

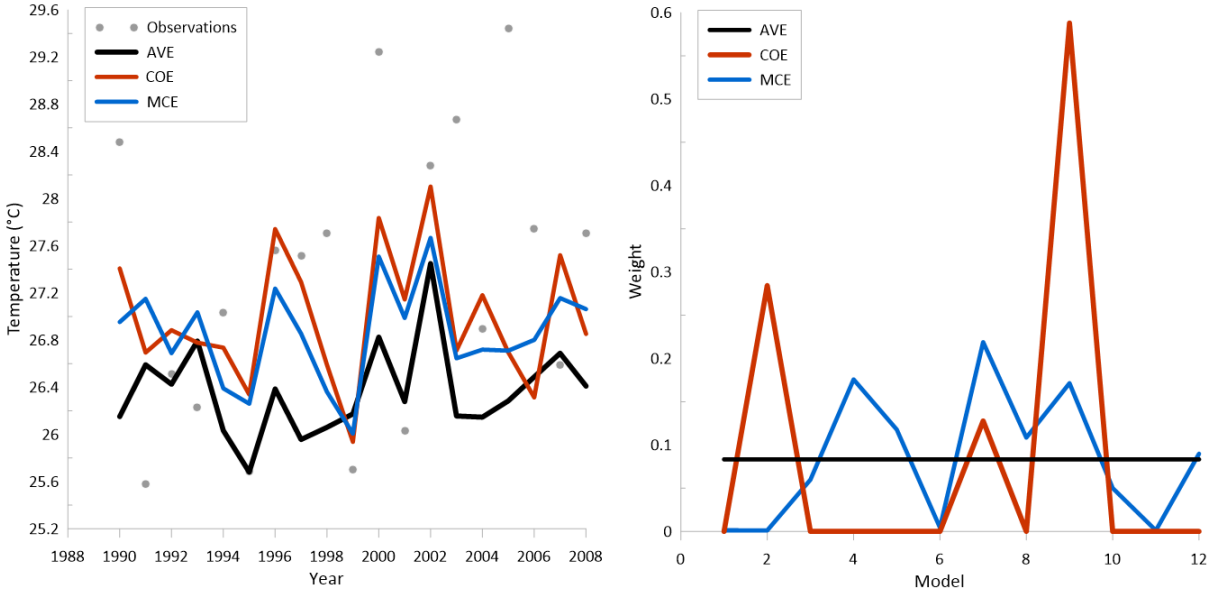

Figure 5. Weighted ensembles (left panel) and weights distributions (right panel), comparison of different methods on NARCliM data.

As in CMIP5 data analysis (Figure 3), we see that the MCE method is maintaining (i.e., assigning 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 CMIP5 case, the difference between the MCE output weights and the equal weights is also considerably larger. The MCE method shows itself being capable of maintaining much of the ensembles' diversity, though a few models receive zero weightings. This allows MCE to substantially improve performance over the AVE method on both training and validation periods.

https://doi.org/10.5194/gmd-2020-253
Preprint. Discussion started: 1 October 2020

Geoscientific Model Development Discussions
Open Access
EGU

3.3 KMA data

We select σ and L by analysing MCE performance on training data set as described in section 2.2.1 and shown in Figure 6.

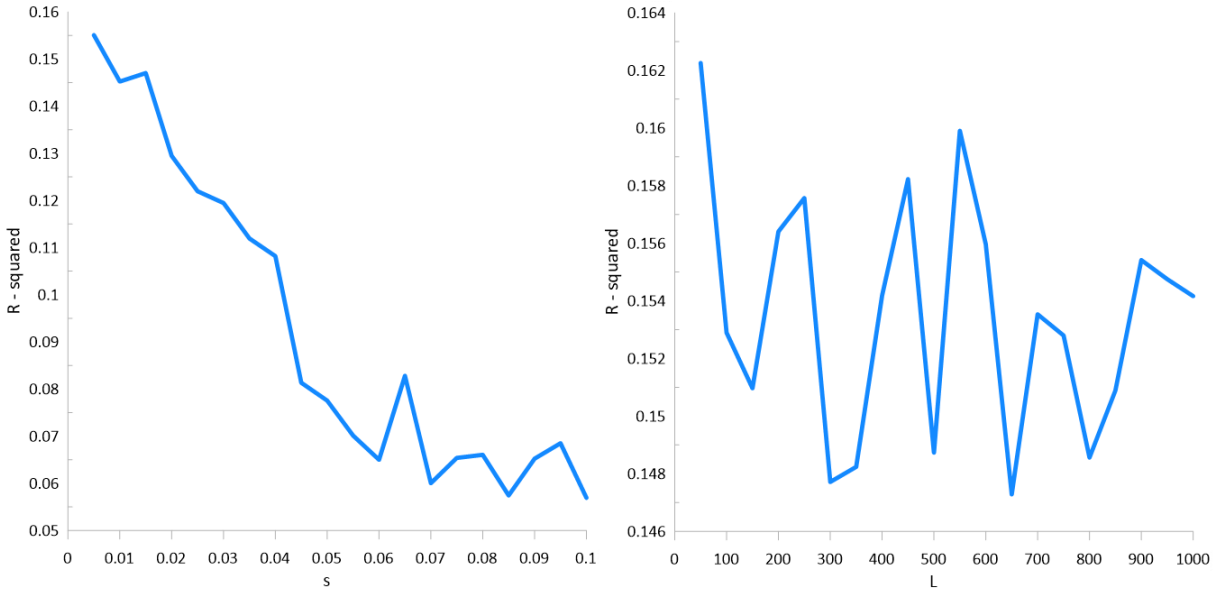

Figure 6. MCE method performance on training data for KMA data depending on L with σ ($L = 300$ on the left hand side and $\sigma = 0.005$ on the right hand side).

Applying the MCE method on the selected data with $T = 33(1973 - 2005)$, $\sigma = 0.005$, $L = 550$ 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:

$Ensemble$	$RMSE_T$	R^2_T	$RMSE_V$	R^2_V
E_{AVE}	0.37	0.07	0.52	-0.01
E_{COE}	**0.24**	**0.61**	0.55	-0.10
E_{MCE}	0.35	0.16	**0.49**	**0.10**

Table 4. Performance comparison of different methods on KMA data. RMSE on training ($RMSE_T$) and validation ($RMSE_V$) data; R^2 on training (R^2_T) and validation (R^2_V) data.

225 We can see that MCE significantly outperforms other methods using R-squared metrics and has the lowest RMSE. We analyse this difference in methods' performance by looking at weighted ensembles and model weights for each method:

https://doi.org/10.5194/gmd-2020-253
Preprint. Discussion started: 1 October 2020

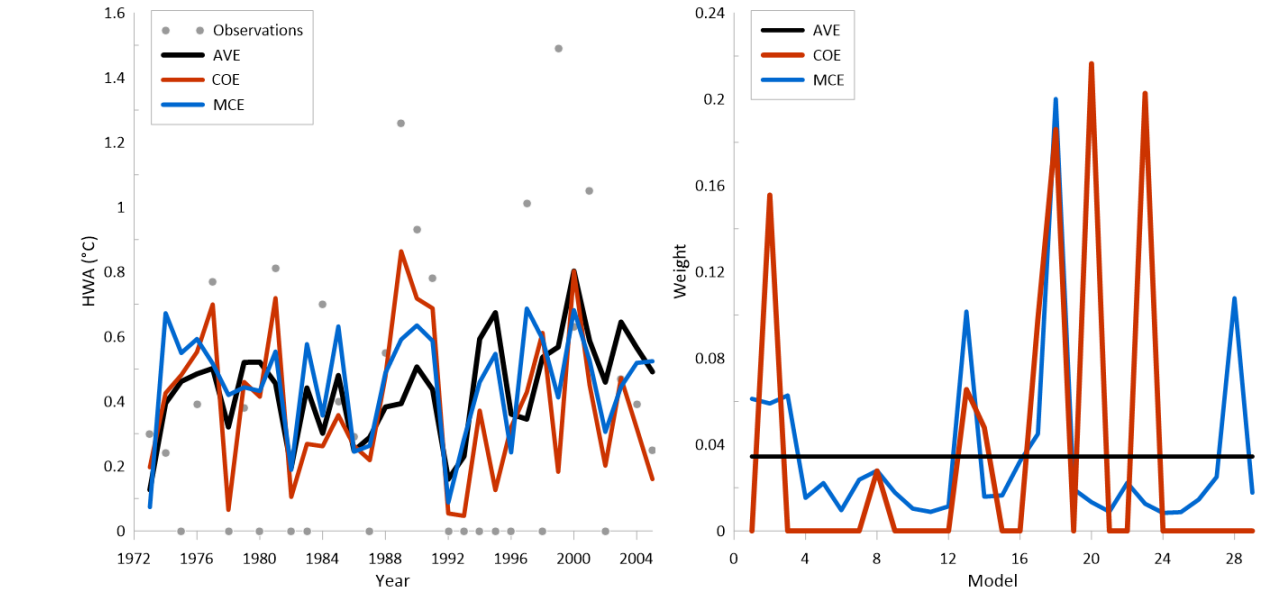

Figure 7. Weighted ensembles (left panel) and weights distributions (right panel), comparison of different methods on KMA data

As before, we see that the MCE method maintains the ensembles' diversity with no models receiving zero weights, but a small number of models having much higher weights than the rest. 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.

4 Discussion

The obtained results indicate that Markov chains can be used to construct a better performing weighted ensemble mean with lower RMSE and higher R-squared values on validation data than commonly used methods like multi-model ensemble average and convex optimisation (Tables 2, 3 and 4). While it performed worse than COE on the training periods, we are confident that it is less prone to over-fitting than linear optimisation methods. We attribute this advantage of the MCE method to its ability to maintain the ensemble's diversity while optimising its weights on the training period (Figures 3, 5 and 7).

As the number of models is increased MCE tends to become closer to AVE weights (Figure 3), while being closer to COE with a smaller number of models (Figure 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) and similar cross-correlations between models. The NARCliM data has a relatively small number of models and has four groups of strongly correlated models (Figure 1). In this case, the MCE method recognises the reduced model diversity and does give some models zero weights. The KMA data has an intermediate number of models with a more random cross-correlation structure. This produces a hybrid response in the MCE which maintains ensemble diversity (no models with zero weights) but does weight a small number of models more

https://doi.org/10.5194/gmd-2020-253
Preprint. Discussion started: 1 October 2020


Geoscientific Model Development Discussions

highly. As this effect is observed on all three data sets, we conclude that the main advantage of the MCE method is its ability to preserve the ensemble's diversity while optimising its weights on the training period.

The MCE method is not computationally expensive and is limited only by a software's ability to handle extreme numerical values. These limitations can be easily overcome by adjusting control parameters as our method is robust to their reasonably small variations (Figures 2, 4 and 6). The main limitation of the MCE method is its current inability to quantify the uncertainty of the resulting weighted ensemble mean. However, we believe that given a stochastic nature of the method, this limitation can be overcome in future implementations. MCE performance can be further improved by combining it with other types of optimisation, e.g. linear. In addition, other nonlinear optimisation techniques, which would include more complex structures than simple Markov Chains, can be developed based on our demonstrated results.

Finally, the MCE method doesn't require some of the assumptions necessary for the multi-model ensemble average method (Knutti et al. (2017); Herger et al. (2018); Sanderson et al. (2017)) and linear optimisation techniques (Krishnamurti et al. (2000); Majumder et al. (2018); Bishop and Abramowitz (2013)). In addition, it doesn't produce as many zero weights as the convex optimisation method, hence maintaining more of the models' diversity. We attribute the tendency of the COE method setting 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 w^* for the unrestricted solution of the optimisation problem

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

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

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

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 certain $\tilde{w}$ in the interior of the simplex, we immediately arrive at a contradiction. Take then the point $\hat{w} = w^* + \lambda(\tilde{w} - w^*)$ with $\lambda \in (0,1)$ chosen such that $\hat{w}$ is on the intersection of the line connecting w^* and $\tilde{w}$ with the boundary of the simplex. Because of the strict convexity of the function

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

we have:

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

in contradiction to the assumption that $\tilde{w}$ delivers the minimum over the simplex. Hence the optimisation 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.

https://doi.org/10.5194/gmd-2020-253
Preprint. Discussion started: 1 October 2020


Geoscientific Model Development Discussions

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 data sets, and its performance was compared to two other common approaches (AVE and COE) using cross-validation with RMSE and R-squared metrics. The MCE method was discussed in detail, and its step-wise implementation, including mathematical background, was presented (Table 1).

The results of this study indicate that applying nonlinear ensemble weighting methods on climate data sets can improve future climate projection in terms of accuracy. Even a simple nonlinear structure such as Markov chains shows better performance on different commonly-used data sets than linear optimisation approaches. These results are supported by using standard performance metrics and cross-validation procedures. The developed MCE method is objective in terms of parameter selection, has a sound theoretical basis and has a relatively low number of limitations. Based on the above, we are confident to suggest its application on other data sets and its usage for the future development of new nonlinear optimisation methods for weighting climate model ensembles.

Code and data availability. Code and data for this study is available at https://doi.org/10.5281/zenodo.3965056.

Author contributions. All co-authors contributed to method development, theoretical framework and designing of experiments. MK, YF and JPE selected climate data for this study. MK developed the model code with contribution from SP and performed the simulations. MK and YF prepared the manuscript with contribution from all co-authors.

Competing interests. The authors declare no competing interests.

Acknowledgements. MK would like to acknowledge the support from UNSW Scientia PhD Scholarship Scheme.

https://doi.org/10.5194/gmd-2020-253
Preprint. Discussion started: 1 October 2020

Geoscientific Model Development Discussions
Open Access
EGU

References

Abramowitz, G. and Bishop, C. H.: Climate Model Dependence and the Ensemble Dependence Transformation of CMIP Projections, J. Climate, 28, 2332–2348, https://doi.org/10.1175/JCLI-D-14-00364.1, 2015

Abramowitz, G., Herger N., Gutmann, E. D., Hammerling, D., Knutti, R., Leduc, M., Lorenz, R., Pincus, R., and Schmidt, G. A.: Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing, Earth Syst. Dynam., 1-20, 2018

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

Bishop, C.H. and Abramowitz, G. Climate model dependence and the replicate Earth paradigm, Clim. Dyn., 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, Taylor and Francis Group, 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.: A Bayesian posterior predictive framework for weighting ensemble regional climate models, Geosci. Model Dev., 1-22, 10.5194/gmd-2016-291, 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, Meteorology and Atmospheric Physics – Meteorol. Atmos. Phys., 111, 41-53, 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, 10.1038/ngeo866, 2010.

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

Herger, N., Abramowitz, G., Knutti, R., Angélil, O., Lehmann, K., and Sanderson, B.M.: Selecting a Climate Model Subset to Optimize Key Ensemble Properties, Earth Syst. Dynam., 9, 135 - 151. https://doi.org/10.5194/esd-9-135-2018, 2018.

Huang, J.-C., Huang, W.-T., Chu, P.-T., Lee, W.-Y., Pai, H.-P., Chuang, C.-C., and Wu, Y.-W.: Applying a Markov chain for the stock pricing of a novel forecasting model, Commun. Stat. Theory, 46:9, 4388-4402, DOI: 10.1080/03610926.2015.1083108, 2017.

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

Kharin, V. V., and F. W. Zwiers: 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, doi:10.1002/2016GL072012, 2017.

Krishnamurti, T., Kishtawal, C., LaRow, T., Bachiochi, D., Zhang, Z., Williford, C., Gadgil, S., and Surendran, S.: Improved Weather and Seasonal Climate Forecasts From Multi-Model Superensemble, Science, 285, 1548-1550, 10.1126/science.285.5433.1548, 1999.

https://doi.org/10.5194/gmd-2020-253
Preprint. Discussion started: 1 October 2020

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.

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.

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, 1-10, DOI: 10.1080/1755876X.2018.1438341, 2018.

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, 768-72., 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, 10.3354/cr01403, 2016.

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(5), 055001, doi: 488 10.1088/1367-2630/17/5/055001, 2015.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Refaeilzadeh, P., Tang, L. and Liu, H.: Cross-Validation, Encyclopedia of Database Systems, 532–538, 532-538, 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, 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. of Atmos. Sci., 54, 1-9, 10.1007/s13143-017-0059-7, 2017.

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