Received: 3 December 2017 | Revised: 16 March 2018 | Accepted: 2 May 2018 | Published on: 27 July 2018

DOI: 10.1002/joc.5705

International Journal of Climatology RMetS

RESEARCH ARTICLE

Using multi-model ensembles of CMIP5 global climate models to reproduce observed monthly rainfall and temperature with machine learning methods in Australia

Bin Wang[1] | Lihong Zheng[2] | De Li Liu[1,3] | Fei Ji[4] | Anthony Clark[5] | Qiang Yu[6,7,8]

[1]NSW Department of Primary Industries, Wagga Wagga Agricultural Institute, Wagga Wagga, New South Wales, Australia

[2]School of Computing and Mathematics, Charles Sturt University, Wagga Wagga, New South Wales, Australia

[3]Climate Change Research Centre and ARC Centre of Excellence for Climate Extremes, University of New South Wales, Sydney, New South Wales, Australia

[4]NSW Office of Environment and Heritage, Department of Planning and Environment, Sydney, New South Wales, Australia

[5]NSW Department of Primary Industries, Orange Agricultural Institute, Orange, New South Wales, Australia

[6]Faculty of Science, School of Life Sciences, University of Technology Sydney, Sydney, New South Wales, Australia

[7]Key Laboratory of Water Cycle and Related Land Surface Processes, Institute of Geographic Sciences and Natural Resources Research, University of Chinese Academy of Sciences, Beijing, China

[8]State Key Laboratory of Soil Erosion and Dryland Farming on the Loess Plateau, Northwest A&F University, Yangling, China

Correspondence
Bin Wang, NSW Department of Primary Industries, Wagga Wagga Agricultural Institute, Wagga Wagga, NSW 2650, Australia.
Email: bin.a.wang@dpi.nsw.gov.au

Global climate models (GCMs) are useful tools for assessing climate change impacts on temperature and rainfall. Although climate data from various GCMs have been increasingly used in climate change impact studies, GCMs configurations and module characteristics vary from one to another. Therefore, it is crucial to assess different GCMs to confirm the extent to which they can reproduce the observed temperature and rainfall. Rather than assessing the interdependence of each GCM, the purpose of this study is to compare the capacity of four different multi-model ensemble (MME) methods (random forest [RF], support vector machine [SVM], Bayesian model averaging [BMA] and the arithmetic ensemble mean [EM]) in reproducing observed monthly rainfall and temperature. Of these four methods, the RF and SVM demonstrated a significant improvement over EM and BMA in terms of performance criteria. The relative importance of each GCM based on the RF ensemble in reproducing rainfall and temperature could also be ranked. We compared the GCMs importance and Taylor skill score and found that their correlation was 0.95 for temperature and 0.54 for rainfall. Our results also demonstrated that the number of GCMs ensemble simulations could be reduced from 33 to 25 in RF model while maintaining predictive error less than 2%. Having such a representative subset of simulations could reduce computational costs for climate impact modelling and maintain the quality of ensemble at the same time. We conclude that machine learning MME could be efficient and useful with improved accuracy in reproducing historical climate variables.

KEYWORDS

GCMs, machine learning, multi-model ensemble, random forest, support vector machine

1 | INTRODUCTION

The Coupled Model Intercomparison Project phase 5 (CMIP5) multi-model data set (Taylor *et al*., 2012) contains output from more than 40 different global climate models (GCMs). Not only does this data set facilitate comprehensive GCM diagnoses, validation and inter-comparison for historical periods, it provides opportunities to explore projected future changes in climate conditions. Indeed, the CMIP5 data set is the basis of global and regional climate projections presented in the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC) (IPCC, 2013). However, because the structural differences in the initialization and the mathematical parameterizations of various physical processes vary largely from one GCM to another, climate change projections (e.g., rainfall and temperature) produced by GCMs are often uncertain, with different GCMs simulating different climate changes for the same forcing of the climate system, and this uncertainty cascades to downstream impact studies (Zhuang *et al*., 2016).

For the last few years, researchers have been proposing numerous techniques to tackle the uncertainty in climate

projections. A promising way to assess the uncertainty is through working with multi-model ensembles (MME), which have the potential to reduce uncertainties in present-day simulations and to improve confidence in some aspects of future climate projections. Recently, various types of ensemble methods have been developed and utilized to enhance the quality of prediction. Generally, ensemble methods can be divided into two groups: a simple composite method where each model is weighted equally regardless of the simulation skill, and a weighted ensemble in which a different weight is applied based on its simulation skill (Oh and Suh, 2016). The multi-model ensemble mean (EM) is the most common and widely used approach. There is some evidence that the "mean model" result, obtained by averaging over the ensemble of models, provides an overall best comparison to observations for climatological mean fields if the simulation data used in the ensemble are produced independently (Lambert and Boer, 2001). However, different model development groups share ideas for parameterizations and even sections of model code (model components), so this makes us suspect that each of these models provides a dependent estimate of future climate change. In fact, there is extensive replication of model code in the CMIP5 GCMs, especially within institutions but also in some cases between institutions (Sanderson *et al*., 2015b). Therefore, ignoring the dependence of GCMs might result in a false model consensus, poor accuracy and poor estimation of uncertainty (Herger *et al*., 2017). In addition, the EM approach is clearly not appropriate for all types of evaluation, as multi-model mean data sets have significantly less spatial and temporal variance than data sets from individual models or from observations.

Numerous studies have demonstrated that a weighted ensemble method, which is based on the simulation skills of the models and sometimes also accounts for dependence between different models (Sanderson *et al*., 2015a; Leduc *et al*., 2016; Annan and Hargreaves, 2017), can have better projection skills than the ensembles with equal weighting (Bishop and Abramowitz, 2013; Oh and Suh, 2016; Wang *et al*., 2016a). Recently, several different means of weighting different climate models in an ensemble have been developed. For example, Bishop and Abramowitz (2013) derived weights that explicitly account for model dependence defined using covariance of model errors. They concluded that such a weighing scheme based on the correlation of model errors outperformed the simple model mean. Sanderson *et al*. (2015a) developed a novel approach to weight models by taking the model observation distance matrix as a measure of model performance and interdependence. Bayesian model averaging (BMA) (Raftery *et al*., 2005; Yang *et al*., 2012a) has been widely used to combine climate forecasts from individual models and characterize the uncertainty caused by model structure (Madadgar and Moradkhani, 2014). In this approach a priori weights for each model (often equal weights) is given, and then the weights are updated based on model agreement with observations, which are able to improve both uncertainty estimation and prediction (Wallach *et al*., 2016). However, model weighting is not a standard procedure in climate modelling. It is acknowledged by the latest Intergovernmental Panel on Climate Change (IPCC) report (IPCC, 2013) that the climate community does not know how to weight models to determine the best estimate of future climate change (Wallach *et al*., 2016). The performance of climate models varies from variable to variable and region to region among different aspects of climate system (Kerkhoff *et al*., 2015; Qi *et al*., 2016) and how to quantify model skill and derive models weights as well as use the best way to combine model results is difficult to determine and still controversial (Wallach *et al*., 2016).

Recently, state-of-the-art machine learning (ML) techniques have become appealing in a wide variety of climate change research or prediction problems. ML turns out to be especially suitable because of its key advantage of investigating nonlinear and hierarchical relationships between the predictors and the response using an ensemble leaning approach. Acharya *et al*. (2014) employed extreme learning machine on seven GCMs to make an MME-based estimation of the northeast monsoon rainfall over south peninsular India. They found that extreme learning machine can capture these seasonal rainfall extremes reasonably well compared to the other MME methods (e.g., simple arithmetic mean). Similarly, Kumar *et al*. (2012) used artificial neural network (ANN) to develop MME system to forecast summer monsoon rainfall and concluded such method has a higher skill than individual GCM projection and the simple EM. Tao *et al*. (2018) developed residual-based bagging tree (RBT) model to correct biases between GCMs simulations and observations. They found that RBT approach performed better in reducing biases when compared with the raw EM, the EM with simple additive bias correction and the single best model. ML has also been applied in climate projections to statistically downscale monthly temperature and rainfall with different input (predictive) variables (Salcedo-Sanz *et al*., 2016; Sarhadi *et al*., 2016; Vu *et al*., 2016).

GCM uncertainties and biases are the two major obstacles for realistic assessments of climate change impacts. Bias correction techniques including mean and/or variance-based method, quantile mapping and transfer function are very popular and have been widely used to reduce model bias in climate modelling (Ines and Hansen, 2006; Wang *et al*., 2016b). However, bias correction is mainly able to correct some GCM systematic biases, but insufficient in correction of non-stationary GCM biases and inter-GCM uncertainties. Using ML for ensemble of multiple GCMs is evidently able to reduce the GCM uncertainties (Kumar *et al*., 2012; Acharya *et al*., 2014).

There are numerous promising approaches such as ANN (Acharya *et al*., 2014), extreme ML (Deo and Şahin, 2015), supervised principal component analysis (Sarhadi *et al*., 2017), relevance vector machine (Okkan and Inan, 2015) which are used to develop MME approach and statistical downscaling. In this study, we consider two different ML approaches, random forest (RF) and support vector machine (SVM) as statistically based models, in a problem of MMEs of CMIP5-based GCMs to reproduce observed monthly rainfall and temperature. RF and SVM have been widely used in handling nonlinear and hierarchical relationships between the predictors and the responses. These two methods work well with few parameters and are easy to implement. For instance, SVM has been successfully applied in some recent cases such as managing nonlinear meteorological events (Salcedo-Sanz *et al*., 2016; Sarhadi *et al*., 2017). On the other hand, RF and SVM have been proven to perform better compared to other ML techniques in some agriculture-related areas. For example, Were *et al*. (2015) found that SVM had lower RMSE and high R^2 values in predicting soil organic carbon stock than ANN in eastern Africa. Wang *et al*. (2018) reported that the RF model outperformed the boosted regression tree model regardless of input features, expressed by larger R^2 and smaller prediction errors in quantifying soil properties in eastern Australia. To our knowledge, no previous study has applied RF and SVM in the multiple GCMs ensemble evaluation. In addition, although a plethora of works has discussed the application of ML in rainfall and temperature prediction, as far as we know, there has been no published work to date on calibrating and validating ensemble models to reproduce the observed historical climate data using these two ML approaches.

The main objective of this study is to test whether employing ML techniques (RF and SVM) to develop MME could perform better in reproducing historical monthly temperature and rainfall than traditional approaches (EM and BMA). We are endeavouring to provide a robust and accurate multi-GCMs ensemble approach to reproduce rainfall and temperature in Australia. We also compared the importance of each GCM based on the RF with its Taylor skill score. The ranked GCMs would be used to test how the number of ensemble size affects model prediction. This study will provide a possibility for using ML method to conduct MME and enhance the existing forecast skills of GCMs ensemble, and it is likely to project more accurate future climate change using these calibrated ML models. This paper is organized in four sections. The introduction is followed by methodology in section 2, which describes the data and study area and includes a description of the four techniques that are used in the study, as well as the forecast verification metrics that are employed to evaluate the generated predictions. Results and discussion are elaborated in section 3, and finally, the summary and conclusion are provided in section 4.

2 | MATERIAL AND METHODS

2.1 | The study area and climate data

We used 108 weather stations based on the recently released Australian Climate Observations Reference Network-Surface Air Temperature data set (ACORN-SAT; Trewin, 2013), which are available in the SILO (Scientific Information for Land Owners) patched point data set (PPD). As the ACORN-SAT data set extends from 1910 to the present with 60 locations having data for the full post-1910 period, to maintain the consistency of existing databases with long-term climate data, daily temperature and rainfall data during 1900–2016 for these 108 meteorological sites across Australia were extracted from the SILO PPD (Jeffrey *et al*., 2001; http://www.longpaddock.qld.gov.au/silo/ppd/index.php; see Figure S1, Supporting Information).

This study used an ensemble of 33 CMIP5 GCMs. These models and their respective modelling centres are listed in Table S1. We examined gridded monthly surface air temperature and total monthly rainfall data. These gridded monthly data were spatially downscaled to each of 108 weather stations. The spatial interpolation was achieved by using an inverse distance-weighted (IDW) interpolation method in this study. The IDW interpolation method was used to compute rainfall and temperature values for each weather station based on its distance to the geographical centres of the four nearest GCM grid cells (Liu and Zuo, 2012),

$$S_i = \sum_{k=1}^{4} \left[\left(\frac{1}{d_{i,k}}\right)^3 \left(\sum_{k=1}^{4} \left(\frac{1}{d_{i,k}}\right)^3 \right)^{-1} P_k \right], \qquad (1)$$

where S_i is the downscaled site-specific GCM projection at site i, P_k is the GCM projection at the cell k, $d_{i,k}$ is the distance between site i and the centre of cell k.

2.2 | Multi-model ensembles mean

The multi-model EM method is defined as

$$S(t) = \frac{1}{N} \sum_{i=n}^{N} P_n(t), \qquad (2)$$

where $S(t)$ is an EM for time t, N is the total number of GCMs and $P_n(t)$ is the projection of the nth GCM for time t.

2.3 | Bayesian model averaging

BMA has been widely employed as an effective way of correcting under-dispersion in ensemble forecasts (Yang *et al*., 2012a; Wang *et al*., 2014). BMA is a standard statistical procedure for combining predictive distribution from different sources and provides a way of combining statistical models at the same time calibrating them using a training data set (Raftery *et al*., 2005; Yang *et al*., 2012a). The output of BMA is a weighted average of probability density functions

(PDFs) which are centred on the bias-corrected forecast. The BMA weights reflect the relative contributions of the component models to the prediction over the training samples and can be used to assess the usefulness of each ensemble member. In the case of a variable y to be forecast on the basis of training data y^T using K models M_1, …, M_k, the forecast PDF $p(y|\ y^T)$ can be given by

$$p\left(y|y^T\right)=\sum_{k=1}^{K}p\left(y|M_k,y^T\right)p\left(M_k|y^T\right), \tag{3}$$

where $p(y|M_k,y^T)$ is the forecast PDF based on model M_k alone; K is the number of models; $p(M_k|y^T)$ is the posterior probability of model M_k being correct given the training data and indicates how well the model M_k represents the training data. The total sum of the posterior model probabilities is one, that is $\sum_{k=1}^{K}p(M_k|y^T)=1$, and they can thus be treated as weights. The detailed information on BMA can be referred to Raftery *et al*. (2005) and Hoeting *et al*. (1999), which have shown how to estimate weight for each model M_k in BMA. In the present study, BMA analysis was conducted via the "BMS"-package of R software (Zeugner and Feldkircher, 2015; https://cran.r-project.org/web/packages/BMS/BMS.pdf).

2.4 | Random forest

RF was first developed by Breiman (2001) and is a very flexible and powerful tree-based ensemble method in order to improve the regression accuracy. Initially, RF modelling draws bootstrap samples (63%) with replacement from the entire sample population in the training set to grow each tree. The bootstrap sampling leads to RF less sensitive to over-fitting in comparison to decision trees (Heung *et al*., 2014). Therefore, RF contains not a single standard regression tree but many regression trees, like a forest. Unlike most common methods based on ML, RF only needs two parameters to be tuned for generating a prediction model: (a) the number of regression trees to grow in the forest (n_{tree}), (b) the number of randomly selected evidential features at each node (m_{try}). By default, the random subset size is the square root of the number of the entire predictors for model. The RF analysis is a nonparametric algorithm that can handle nonlinear and additive relationships and is used to rank the relative importance of each predictor variable in controlling the response variable. Variable importance is based on the regression prediction error of the out-of-bag, also called the OOB, which is left out of the bootstrap samples (37%). It is computed as a function of change prediction error by permuting with input variable and expressed using mean decrease in accuracy (Heung *et al*., 2014). In error estimation, the OOB sample is predicted by the respective trees and by aggregating the predictions, the mean square error (MSE_{OOB}) is calculated using Equation (3),

$$\text{MSE}_{\text{OOB}}=\frac{1}{n}\sum_{i=1}^{n}\left(O_i-\hat{P}_{i\text{OOB}}\right)^2, \tag{4}$$

where $\hat{P}_{i\text{OOB}}$ is the average of all OOB predictions across all trees.

In the current study, in order to optimize two parameters of n_{tree} and m_{try}, a number of experiments were conducted using different combinations of n_{tree} and m_{try}. The range of number of n_{tree} was set between 500 and 1,300 at intervals of 200 for temperature (n_{tree} ranging from 100 to 1,000 at a step length of 200 for rainfall), and the number of selected evidential features m_{try} was between 5 and 20 at 1 intervals for temperature (m_{try} ranging from 1 to 15 at a step length of 1 for rainfall). To save computing time, the training data were partitioned into threefolds for cross validations and the error rates for each of the three cross-validation partitions were aggregated into a mean error rate. Three replicates of the threefold cross validation were conducted. The final model (optimal model) was selected when the prediction error was lowest. Figures S2 and S3 showed the process of tuning parameters for RF using grid research method to determine the optimal parameters that produce the minimum forecasting error (Kuhn, 2008). The optimal value for m_{try} and n_{tree} for both rainfall and temperature can be found in Table S2. In short, RF had optimal m_{try} between [12, 18], n_{tree} between [900, 1,300] for monthly temperature. For monthly rainfall, optimal m_{try} was between [7, 15], and n_{tree} was between [500900].

2.5 | Support vector machine

SVM analysis is another popular supervised ML tool for classification and regression, proposed by Cortes and Vapnik (1995). There are many successful applications of SVM in image segmentation, object detection, image classification, handwriting recognition, text and hypertext categorization, and applications in the biological and other sciences. SVM uses hyperplanes to divide all of the data into different classes optimally. It has a better learning capability and smaller prediction errors than many other methods (Chen *et al*., 2010; Sarhadi *et al*., 2016; Hou *et al*., 2017).

For a given observation sample set of N input and output data,

$$D=\{(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)\}\in R^K\times R. \tag{5}$$

It is assumed that a regression function is expressed as

$$F=\left\{f|f(x_i)=w^T*x_i+b,w\in R^K\right\}, \tag{6}$$

where w is the unit normal vector to the hyperplane, b is the distance from the origin to the hyperplane and x_i is the ith input vector. Essentially, SVM follows an idea of maximal margin that allows treating SVM regression as a convex optimization problem. SVM is a useful technique which provides the user with high flexibility in terms of distribution of

underlying variables, relationship between independent and dependent variables.

The function used to predict new values depends only on the support vectors,

$$f(x)=\sum_{n=1}^{N}\left(\alpha_n-\alpha_n^*\right)k(X_n,X)+b, \tag{7}$$

where $k(X_n, X)$ is the kernel function. In general, there are four types of kernel functions commonly seen in SVMs.

- Linear: $<x, x'>$.
- Polynomial: $(\gamma < x, x' > + C)^d$, The most common degree is $d = 2$ (quadratic), since larger degrees tend to overfit. As one problem with the polynomial kernel is that it may suffer from numerical instability: when $<x, x'> + c<1$, $K < x, x' > = (<x, x' > + c)^d$ tends to zero with increasing d, whereas when $<x, x' > + c > 1$, $K < x, x' > = (<x, x' > + c)^d$ tends to infinity.
- Radial basis function kernel: $\exp^{-\gamma|x-x'|^2}$, γ must be greater than 0.
- Sigmoid: $\tanh(\gamma < x, x^{'} > + r)$, where $r < 0$.

As the relationship in our case is nonlinear, we select radial basis function (RBF) as the kernel model for SVM in the following experimental test. SVM RBF kernel needs two parameters to be tuned including penalty (*cost*) that controls the trade-off between margin and training errors, and the kernel width (*sigma*) that controls the degree of nonlinearity of the model (Naghibi *et al*., 2017). The regression problem can be solved by a standard quadratic programming form to obtain the optimal solution.

Similarly, in order to optimize two parameters of *cost* and *sigma*, a number of experiments were conducted using different combinations of *cost* and *sigma.* The number of *cost* was set between 4, 8, 16, 32 and 64 for temperature and rainfall. The range of number of *sigma* was set between 0.005 and 0.05 at intervals of 0.005. In the current study, three replicates of the threefold cross validation were used to select the optimal parameters of SVM. Figures S4 and S5 showed the process of tuning parameters for SVM using grid research method to determine the optimal parameters that produce the minimum forecasting error. The optimal value for *cost* and *sigma* for both rainfall and temperature can be found in Table S2. In short, SVM had an optimal *cost* value between [8, 32], and the value of *sigma* between [0.065, 0.070] for monthly temperature. The value of cost was between [4, 8] and sigma was between [0.035, 0.050] for monthly rainfall.

2.6 | Model evaluation

We used a random 90% of climate data including 108 sites over 1900–2016 for training ("calibration data set") and the remaining 10% was used as the "validation data set" to validate the prediction of historical monthly rainfall and temperature. This process ensured that the calibration and validation data are independent but sampled from the same overall population of data. To assess the accuracy of the prediction versus observations, two statistical indices were considered: regression coefficients of the coefficient of determination (R^2) measuring the percentage of variation explained by each model; the root-mean-square error (RMSE) measuring the overall accuracy of the prediction,

$$R^2=\left(\frac{\sum_{i=1}^{n}\left(O_i-\overline{O}\right)\left(P_i-\overline{P}\right)}{\sqrt{\sum_{i=1}^{n}\left(O_i-\overline{O}\right)^2}\sqrt{\sum_{i=1}^{n}\left(P_i-\overline{P}\right)^2}}\right)^2, \tag{8}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i-O_i)^2}, \tag{9}$$

where P_i and O_i are the predicted and observed monthly rainfall and temperature; n is the number of samples; $\overline{P}$ and $\overline{O}$ are the means for the predicted and observed rainfall and temperature; a good model will have and R^2 close to 1 and RMSE of almost 0.

3 | RESULTS AND DISCUSSION

3.1 | Model performance

The 90% of data sets were selected as training data in our study because we would like to include more data to establish the relationship between GCMs and observed climate data. In previous literatures, the proportion of training data can range from 60 to 95% (Deo and Şahin, 2015; Vu *et al*., 2016; Hou *et al*., 2017). Therefore, 90% was within the range when we designed the study. However, we used different training and testing data set to test whether they have impacts on BMA and EM results in our study. The results show that different data set for validation had a small change in R^2 and RMSE (Table S3).

To assess the improvement of RF and SVM based MME, we compared their performance against EM and BMA. As shown in Figure 1, specifically, two evaluation parameters R^2 and RMSE were used to evaluate model accuracy in reproducing monthly temperature. We also added values of two statistical indicators for the individual GCM models. It was clear to see that MME techniques improved model performance in reproducing monthly temperature compared to each single GCM, despite the fact that the simulations of the individual model varied considerably. This result was consistent with SVM providing the best agreement with observations in all cases with R^2 being no lower than 0.96 (Figure 1). RF's performance was ranked as the second with R^2 varying between 0.92 and 0.96. By contrast, EM and BMA achieved R^2 in the average of 0.89 and 0.91

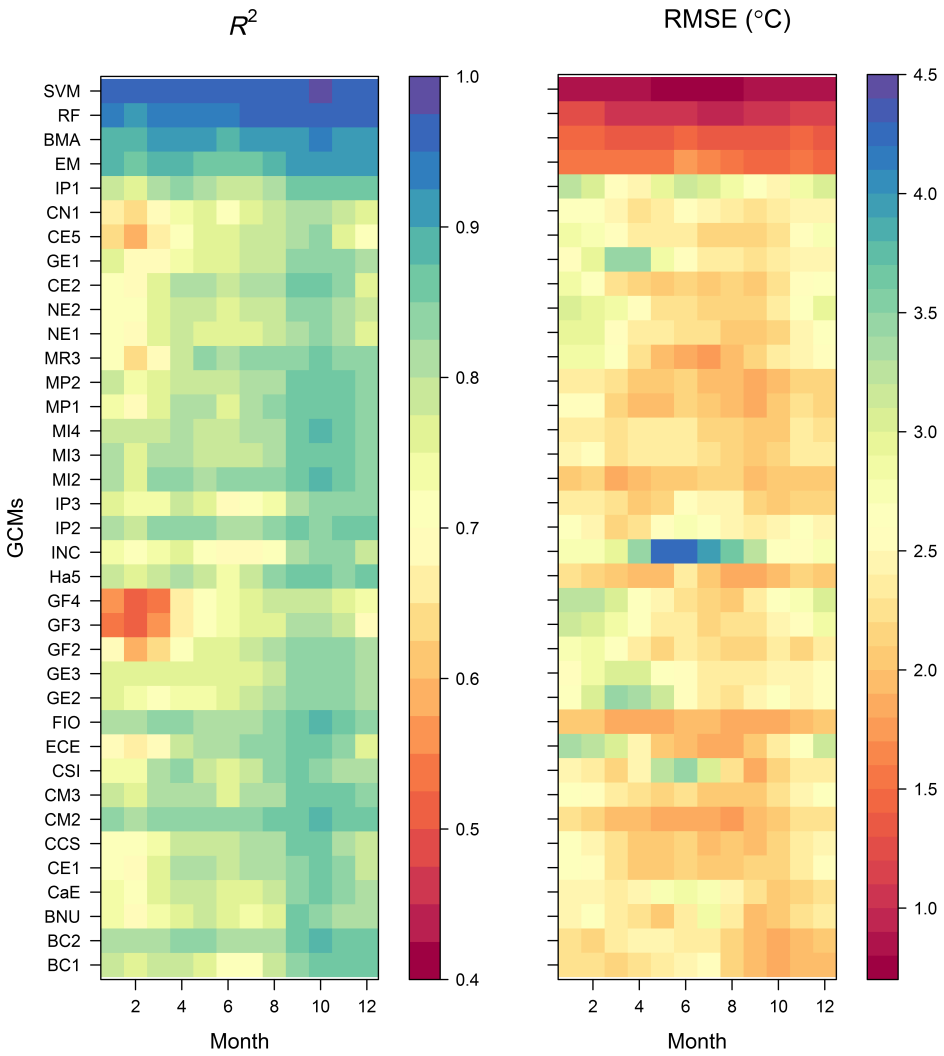

FIGURE 1 Summary statistics of the predictive quality of simple arithmetic mean (EM), BMA, RF and SVM together with 33 GCMs for monthly temperature for testing data sets; the coefficient of determination (R^2) and RMSE were used to evaluate model accuracy [Colour figure can be viewed at wileyonlinelibrary.com]

across the months, respectively. Our results also demonstrated that SVM had the smallest predictive error with the RMSE of 0.82 °C on average. Meanwhile, the RMSE (1.08 °C) produced by RF were only little bit higher than the SVM. Although EM was simple, quick and its predictive error was less than individual GCMs, the overall accuracy of EM did not outperform two ML techniques. SVM did show a superior performance in every aspect, which showed its advantage in handling the nonlinear relationship between observations and GCMs simulations. Table 1 shows the difference between two statistical indices estimated from two ML methods and their estimated values from EM and BMA. On average across the months, RF had an improvement of 6.5 and 4.2% for R^2 while RMSE decreased by 30.6 and 23.2% compared to EM and BMA, respectively. The predictive power of SVM increased most by 9.1 and 6.7% for R^2 compared to EM and BMA, respectively. RMSE produced by SVM was lower than EM and BMA by 47.1 and 41.3%, respectively.

Similarly, we tested these four methods on monthly rainfall. Although SVM had better performance than other three methods (Figure 2) and MME outperformed the individual GCMs, the overall predictive accuracy for rainfall was less satisfactory than temperature. The main reason was likely to be the high variability in rainfall, which introduced much challenges to accurately simulate rainfall (Yang *et al.*, 2012b). In summary, SVM had the best performance with R^2 being between 0.45 and 0.67 in each month and the results were relatively satisfactory. RF was the second place with R^2 varying between 0.42 and 0.62. By contrast, EM achieved R^2 only between the range of 0.15 and 0.49 and the R^2 of BMA ranged between 0.22 and 0.53. SVM had an average increase of 106.8 and 69.7% for R^2 compared to EM and BMA across the months, respectively (Table 1). The predictive errors RMSE decreased by 25.4 and 19.2% compared with EM and BMA, respectively. However, the performance of RF did not improve as much as that of SVM.

Overall, for monthly rainfall, the ensemble SVM simulations showed the best performance, followed by RF and EM. This result coincided with what we have achieved when employing a MME for monthly temperature. Our study

TABLE 1 The relative changes (%) between two statistical indicators estimated from two ML methods (RF and SVM) and their estimated values from simple arithmetic mean (EM) and BMA in reproducing observed monthly temperature (rainfall in parentheses)

	Compared to EM				Compared to BMA			
	R^2		RMSE		R^2		RMSE	
	RF	SVM	RF	SVM	RF	SVM	RF	SVM
Jan	5.5 (28.2)	9.4 (30.7)	−24.1 (−19.4)	−45.5 (−20.1)	4.0 (18.4)	7.8 (20.7)	−18.9 (−10.8)	−41.8 (−11.6)
Feb	6.5 (35.9)	10.4 (45.7)	−24.5 (−16.6)	−43.8 (−20.9)	4.3 (29.5)	8.1 (38.8)	−18.0 (−12.4)	−39.1 (−16.9)
Mar	6.8 (35.5)	9.9 (46.7)	−28.8 (−16.7)	−46.6 (−21.0)	3.7 (24.5)	6.7 (34.7)	−19.4 (−9.6)	−39.5 (−14.2)
Apr	6.2 (172.4)	9.2 (188.9)	−27.6 (−18.0)	−45.9 (−19.6)	3.5 (94.1)	6.5 (105.8)	−19.1 (−13.5)	−39.6 (−15.2)
May	8.6 (150.6)	11.5 (186.1)	−35.5 (−23.2)	−52.3 (−28.6)	4.9 (80.4)	7.7 (106.0)	−25.3 (−14.2)	−44.7 (−20.3)
Jun	10.0 (123.8)	13.0 (169.6)	−39.3 (−22.7)	−54.5 (−31.5)	5.8 (79.2)	8.6 (115.9)	−27.5 (−14.6)	−45.7 (−24.3)
Jul	9.7 (50.7)	11.2 (65.2)	−43.9 (−25.1)	−52.7 (−32.3)	5.9 (36.9)	7.3 (50.2)	−32.4 (−14.9)	−43.0 (−23.0)
Aug	8.2 (45.9)	10.0 (60.2)	−38.8 (−23.3)	−50.8 (−29.9)	5.2 (32.3)	7.0 (45.3)	−29.2 (−14.4)	−43.0 (−21.8)
Sep	4.9 (71.5)	6.5 (85.7)	−31.1 (−22.7)	−43.9 (−27.7)	4.0 (48.3)	5.6 (60.5)	−26.4 (−17.1)	−40.1 (−22.5)
Oct	3.8 (132.8)	5.6 (160.8)	−26.6 (−23.9)	−43.7 (−30.5)	3.0 (89.7)	4.8 (112.5)	−22.7 (−18.7)	−40.8 (−25.8)
Nov	3.7 (132.8)	6.1 (158.7)	−23.3 (−18.3)	−43.1 (−22.5)	3.0 (74.9)	5.4 (94.4)	−19.9 (−14.5)	−40.6 (−18.9)
Dec	3.8 (78.9)	6.0 (82.7)	−24.1 (−17.8)	−41.8 (−20.1)	2.9 (48.7)	5.1 (51.9)	−19.2 (−13.0)	−38.0 (−15.5)
Average	6.5 (88.3)	9.1 (106.8)	−30.6 (−20.6)	−47.1 (−25.4)	4.2 (54.7)	6.7 (69.7)	−23.2 (−14.0)	−41.3 (−19.2)

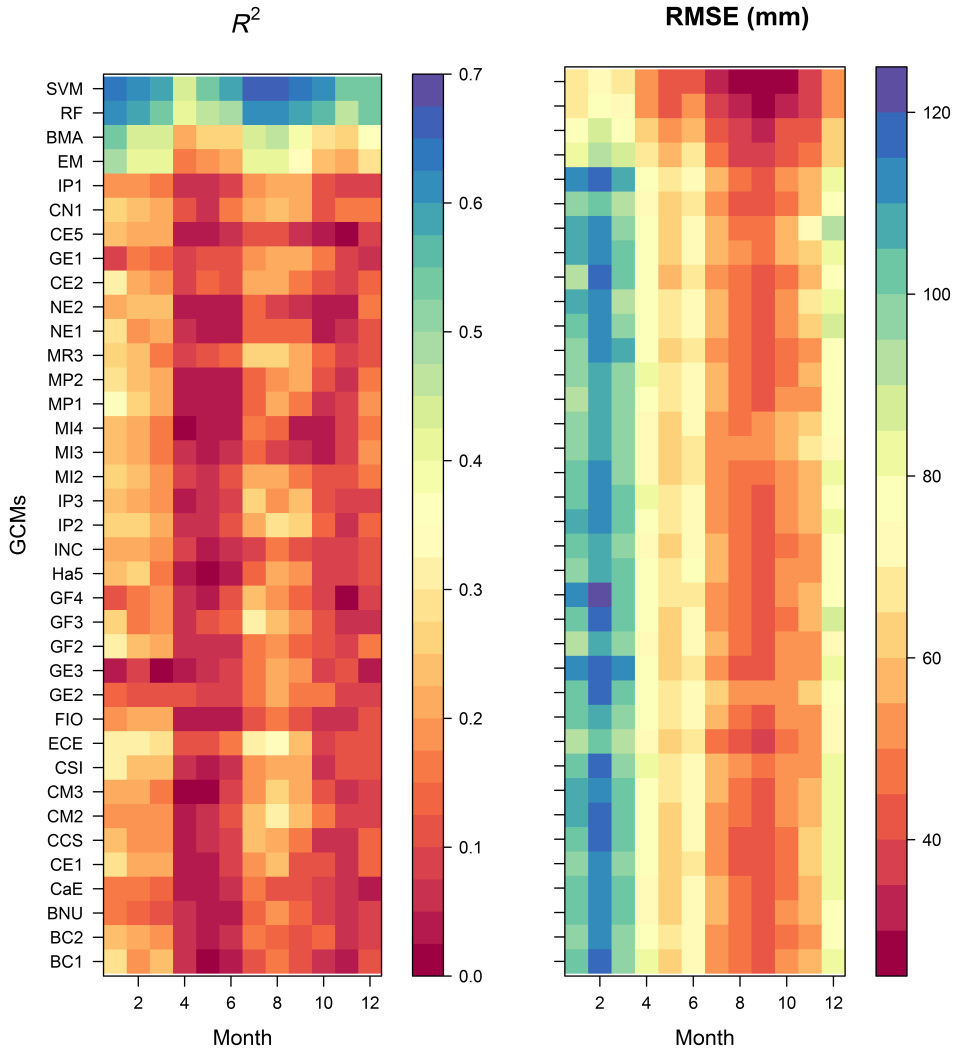

FIGURE 2 Summary statistics of the predictive quality of simple arithmetic mean (EM), BMA, RF and SVM together with 33 GCMs for monthly rainfall for testing data sets; the coefficient of determination (R^2) and RMSE were used to evaluate model accuracy [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 2 The averaged relative importance of the 33 GCMs derived from RF in reproducing historical monthly rainfall and temperature. Taylor skill score was considered to rank the GCM model. The results were shown in decreasing order of RF importance for temperature and rainfall, respectively. The bold number indicated the top 11 GCMs in Taylor skill score, the bold italic number represented GCMs ranked in the middle (12–22) and the remaining were ranked at the bottom. Standard deviation for the importance and skill score of each GCM across the 12 months were shown in bracket

Temperature			Rainfall		
GCM	**Importance**	**Score**	**GCM**	**Importance**	**Score**
CM2	99.30 (±1.78)	**0.92** (±0.01)	ECE	89.31 (±15.16)	**0.45** (±0.10)
BC2	87.57 (±9.21)	**0.91** (±0.02)	CN1	74.31 (±29.25)	**0.43** (±0.06)
IP2	87.45 (±7.65)	**0.90** (±0.02)	MR3	72.94 (±27.32)	**0.39** (±0.05)
MI2	85.24 (±6.09)	**0.91** (±0.02)	IP2	69.31 (±27.94)	0.26 (±0.05)
FIO	81.48 (±13.66)	**0.91** (±0.01)	CE2	68.73 (±13.79)	**0.40** (±0.08)
IP1	76.09 (±16.16)	***0.89*** (±0.03)	GF3	67.48 (±19.43)	**0.39** (±0.08)
MI4	75.38 (±12.66)	**0.90** (±0.02)	CM2	62.09 (±13.55)	0.33 (±0.07)
MI3	74.76 (±14.26)	**0.89** (±0.02)	GF2	61.95 (±23.97)	***0.35*** (±0.09)
Ha5	73.27 (±12.92)	**0.90** (±0.02)	MI2	59.60 (±24.50)	**0.43** (±0.08)
ECE	72.94 (±17.92)	***0.88*** (±0.05)	IP3	59.38 (±23.70)	***0.36*** (±0.08)
CM3	72.47 (±13.40)	**0.90** (±0.02)	CSI	58.45 (±16.92)	***0.36*** (±0.04)
MR3	71.51 (±16.86)	***0.88*** (±0.05)	IP1	56.02 (±18.77)	0.24 (±0.12)
MP1	70.59 (±8.79)	**0.90** (±0.03)	MP2	52.64 (±28.04)	**0.37** (±0.06)
MP2	68.18 (±11.95)	**0.90** (±0.02)	CE1	52.45 (±18.06)	**0.39** (±0.07)
CSI	63.22 (±12.98)	***0.89*** (±0.02)	CCS	52.39 (±23.00)	**0.39** (±0.08)
CCS	60.65 (±6.96)	***0.89*** (±0.03)	CM3	51.38 (±15.20)	0.30 (±0.07)
BC1	60.16 (±26.11)	***0.89*** (±0.03)	MP1	50.63 (±14.79)	***0.35*** (±0.11)
CE2	60.07 (±7.65)	***0.89*** (±0.03)	GE1	49.31 (±23.47)	**0.40** (±0.07)
CE1	58.00 (±10.83)	***0.89*** (±0.03)	GF4	49.11 (±16.08)	***0.35*** (±0.08)
GE3	56.51 (±17.19)	***0.88*** (±0.02)	BC2	48.62 (±14.01)	***0.35*** (±0.06)
CaE	56.31 (±13.34)	***0.88*** (±0.02)	GE2	48.61 (±25.37)	**0.40** (±0.09)
GE2	53.17 (±14.99)	0.88 (±0.03)	Ha5	45.40 (±22.40)	0.32 (±0.06)
NE2	53.08 (±14.19)	***0.88*** (±0.02)	NE1	44.88 (±31.27)	***0.35*** (±0.08)
GE1	50.26 (±14.47)	0.88 (±0.03)	INC	43.75 (±18.05)	0.31 (±0.08)
GF2	49.52 (±12.60)	0.87 (±0.05)	NE2	42.34 (±30.24)	***0.36*** (±0.06)
NE1	48.78 (±12.58)	0.88 (±0.03)	BC1	39.01 (±21.09)	***0.35*** (±0.08)
BNU	48.59 (±16.12)	0.88 (±0.03)	CE5	37.83 (±18.83)	***0.34*** (±0.06)
IP3	44.56 (±27.46)	0.87 (±0.02)	FIO	37.59 (±19.73)	0.30 (±0.07)
CN1	39.00 (±10.88)	0.86 (±0.03)	MI3	36.51 (±14.60)	***0.34*** (±0.10)
INC	37.92 (±28.34)	0.87 (±0.02)	GE3	33.29 (±17.71)	0.31 (±0.09)
CE5	33.04 (±15.08)	0.85 (±0.05)	BNU	30.30 (±26.98)	0.26 (±0.07)
GF3	15.16 (±11.24)	0.83 (±0.08)	MI4	30.22 (±23.19)	0.33 (±0.10)
GF4	10.31 (±12.78)	0.83 (±0.07)	CaE	25.34 (±27.52)	0.26 (±0.10)

aimed at developing an integrated multi-GCM-based ML method to reproduce monthly rainfall and temperature. We have evaluated the capability of two ML methods by investigating the relationship between the observation and ensemble simulations. MME outperformed the individual models in our study, which was consistent with previous studies (Robertson *et al*., 2004; Tebaldi and Knutti, 2007; Chiyuan *et al*., 2014; Zhuang *et al*., 2016). Furthermore, our results showed that MME performed better in reproducing monthly temperature than rainfall. The SVM simulations had lower RMSE than the EM ensemble for both temperature and rainfall, which suggested that the simulation of SVM was more similar to the observation data. However, RF's performance was a little bit weaker than SVM. With the help of the Kernel function, a nonlinear regression can be transformed into a linear regression in a hyper space for SVM. Unlike a general linear regression, SVM tries to identify the hyperplane based on the support vectors. Those support vectors are with a tolerated range beside the hyperplane. In other words, such a way is to minimize the error of the objective function of SVM. RF regression identifies the best tree by building a set of decision trees and aggregates the votes of each of its component trees, giving not only an estimate of how this new example should be classified, but also an estimate of the algorithm's certainty of the guess. RF is an ensemble learning method that produces and combines numerous decision trees. However, EM tries to take consideration of every data point in average, which can be weak

and less useful in predicting the sudden changes (e.g., data with large variability).

3.2 | Relative importance of GCMs

Although both RF and SVM are "black box" approaches, RF is able to provide the relative importance of each predictor in model training. In the RF model, the performance of each GCM was evaluated automatically based on their relative importance using the increased percentage of mean square error when each GCM was held in OOB. More specifically, the significance of each GCM represented as the mean decrease in accuracy. The relative importance of GCMs indicated which GCM was more important for the ensemble simulations. We calculated the mean importance of each GCM in ensemble simulations of monthly temperature (rainfall) for the purpose of discriminating least important GCMs (Figure S6). The models were ranked based on the relative importance (Table 2). The results suggested that some GCMs were more important in the RF model for reproduction of temperature including CM2, BC2, IP2, MI2 and FIO (top five) than the others (Table 2). By contrast, ECE, CN1, MR3, IP2 and CE2 ranked as the top five GCMs in reproducing rainfall, which showed that IP2 was the only model to rank in the top five for both temperature and rainfall simulations.

Traditionally, to evaluate the multiple aspects of the performances of GCMs, the Taylor diagram and skill scores (Taylor, 2001) are commonly employed. The Taylor diagram is a 2D plot that concisely summarizes how well a pattern matches the observation in terms of their correlations and the ratio of their variances. In the present study, we used Taylor skill scores calculated as the Equation (10) to measure the differences between observed temperature (rainfall) and GCMs simulated ones,

$$S=\frac{4(1+R)^2}{\left(\frac{\sigma_f}{\sigma_r}+\frac{\sigma_r}{\sigma_f}\right)^2(1+R_0)^2}, \tag{10}$$

where S is the skill score. R is the spatial correlation coefficient in climatological mean values between the simulations and observations. R_0 is the maximum correlation coefficient attainable (here we use 0.999). σ_f and σ_r are the standard deviations of the simulated and observed spatial patterns in climatological means, respectively.

The higher skill scores indicate the higher performance of GCM projections. Figure 3 shows the relationship between the GCMs importance and Taylor skill score for temperature and rainfall. The relative importance of GCMs for temperature were highly related with Taylor score ($r = 0.95$), which indicated high concordance for these two evaluation methods. Table 2 showed that 9 out of top 11 GCMs were consistent for both evaluation methods in temperature. However, the correlation coefficient r for rainfall was low with 0.54. Only 6 out of top 11 GCMs were consistent in rainfall for both evaluation methods.

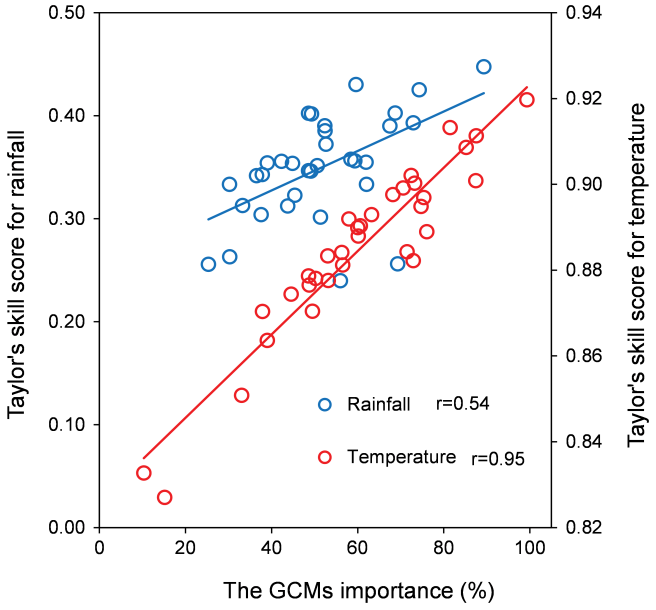

FIGURE 3 The correlation between the GCMs importance derived from the RF model and Taylor skill scores for temperature and rainfall [Colour figure can be viewed at wileyonlinelibrary.com]

3.3 | Ensemble size analysis

The model importance was an interesting outcome of the RF ensemble. Most of the GCMs never occupied the first three positions (Figure S6), which indicated that they were likely to be less important. This aspect offered the opportunity of selecting a subset of suitable GCMs that maintains certain key properties from the full ensemble (Mendlik and Gobiet, 2016; Herger *et al.*, 2017). To analyse the relationship between the number of models in an ensemble and the RMSE of subsets, we employed the GCM ranking derived from the RF to create different MMEs of size N (5, 10, 15, 20, 25 and 30). For each N, the RMSE was calculated using observed data and the RF ensemble simulations. Figure 4 displays that the RMSE for each N decreased with the increasing number of important climate models used for monthly temperature and rainfall. As we can find, for ensemble size 5–10, the RMSE of the performance decreased rapidly for temperature in each month (Figure 4a). This was because the top five important GCMs had less information than the full model ensemble. Interestingly, the performance results varied little with N when N was more than 25, especially for monthly rainfall.

To compare the RMSE of the different ensemble sizes with the full ensemble, a relative change in magnitude of RMSE was calculated by using the difference between each ensemble size and all 33 model runs as Equation (11),

$$\Delta\mathrm{RMSE}(\%)=\frac{\mathrm{RMSE}_N-\mathrm{RMSE}_{33}}{\mathrm{RMSE}_{33}}\times 100, \tag{11}$$

where N represented 5, 10, 15, 20, 25 and 30.

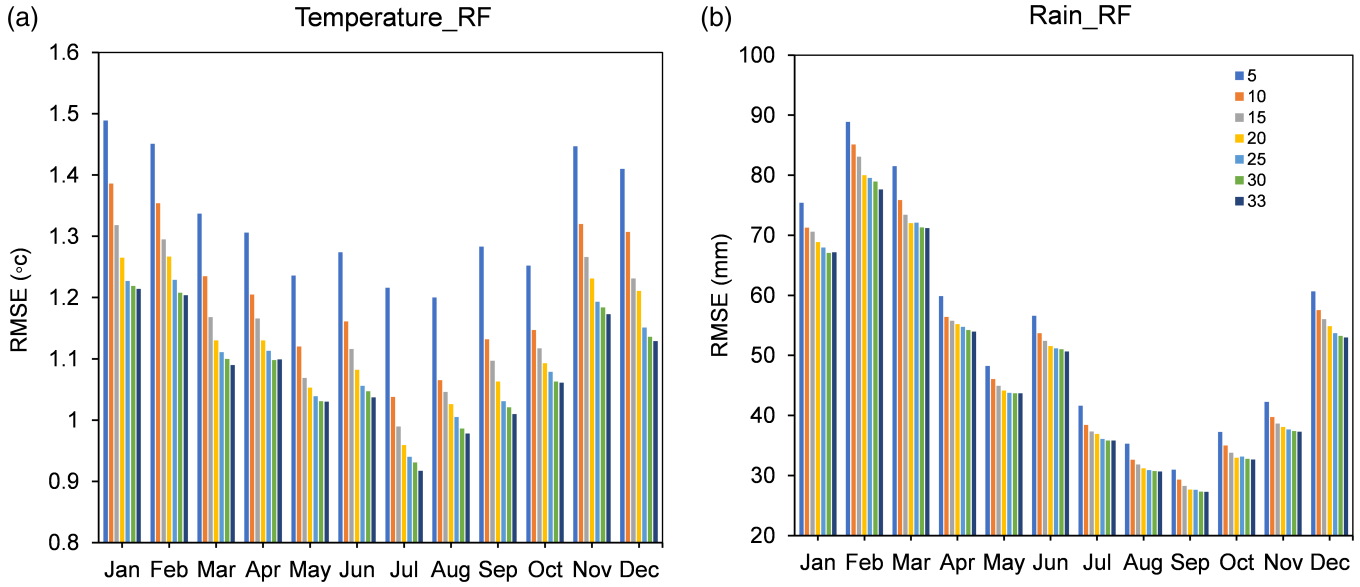

FIGURE 4 How the number of GCMs in an ensemble affects errors estimates. The RMSE of the RF ensemble simulations for reproducing monthly temperature and rainfall were calculated at different size of the CMIP5 subset. We selected 5, 10, 15, 20, 25, 30 of top-ranking GCMs derived from the RF model as benchmarks, respectively [Colour figure can be viewed at wileyonlinelibrary.com]

For a small ensemble size (e.g., $N = 5$), ΔRMSE was large (12% for rainfall and 23% for temperature). However, ΔRMSE decreased when more models were included until it reached a small difference (Figure 5). That was, ΔRMSE reduced by less than 2% when N was 25 for both temperature and rainfall. The question of how many simulations to actually select from the ensemble for different applications still remains and there is no unique and best solution to address this issue (Mendlik and Gobiet, 2016). However, our study provided some information for how to obtain an efficient number of simulations based on its importance. The use of GCMs importance as a criterion to select GCMs would make the ensemble model maintain its performance (RMSE in an acceptable range) and at the same time reduce the computational burden of downstream modelling that uses GCM outputs, such as regional climate modelling or impacts modelling.

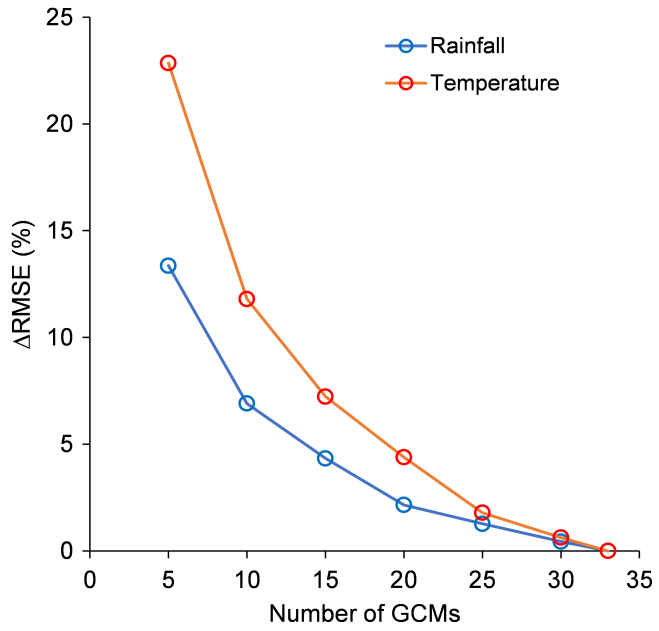

FIGURE 5 The relative change of RMSE for different size of the CMIP5 subset compared with full number of GCMs using the RF ensemble simulations [Colour figure can be viewed at wileyonlinelibrary.com]

4 | CONCLUSIONS

This study focused on developing improved MME schemes for the prediction of historical monthly rainfall and temperature in Australia. For this purpose, two ML methods namely RF and SVM were applied to rainfall and temperature data from 33 CMIP5 GCMs to develop more robust ensemble-based results. The performance of ML methods was compared with the traditional MME technique (e.g., EM and BMA) in terms of two skill metrics (e.g., R^2 and RMSE). The importance of individual GCMs from RF was also examined and compared with Taylor skill scores. In summary, the major findings of this study are:

1. The MMEs (SVM, RF, BMA and EM) obtained better results than any individual model for reproducing monthly temperature and rainfall. Of these four ensemble simulations, the SVM simulations performed better than RF, BMA and EM, which provided the most comparable results to the observations. In addition, MME worked less perfectly in reproducing rainfall than temperature, which might be due to the more stochastic nature of rainfall occurrence and magnitude.
2. The GCMs importance was assigned according to its performance in the RF ensemble simulations and it had a strong correlation with classic Taylor skill score in

temperature and rainfall, which showed a consistency in assessing model performance for both methods.

3. The RMSE of the RF ensemble declined with an increase number of ensemble members. However, the relative RMSE had a small decrease when the number of ensemble size was beyond 25. The smaller ensembles would reduce computational time when driving downstream models, such as regional climate models, compared to the full ensemble.

In summary, ML MME outperformed conventional MME and individual climate models, as demonstrated in the present study. Thus it has potential to be used in developing near/long-term scenarios of regional climate change for the future compared to traditional MME. In addition, it is worth noting that ranking ensemble members through a process-based analysis of RF model output is beneficial for understanding whether the models are adding value to the ensemble. It is suggested that more ML MME should be applied and tested for regional climate models, hydrological models or crop models in order to construct more reliable results in MME projections.

ACKNOWLEDGEMENTS

We acknowledge the modelling groups, Program for Climate Model Diagnosis and Intercomparison and the WCRP's Working Group on Coupled Modelling for their roles in making available the WCRP CMIP5 multi-model dataset. Support of this data set is provided by the Office of Science, US Department of Energy. Comments from Dr. Ian Macadam on an early draft of the manuscript are greatly appreciated. We sincerely thank the Editor and two anonymous reviewers for their constructive suggestions to enhance the quality of the manuscript.

ORCID

Bin Wang http://orcid.org/0000-0002-6422-5802
De Li Liu http://orcid.org/0000-0003-2574-1908

REFERENCES

Acharya, N., Shrivastava, N.A., Panigrahi, B.K. and Mohanty, U.C. (2014) Development of an artificial neural network based multi-model ensemble to estimate the northeast monsoon rainfall over south peninsular India: an application of extreme learning machine. *Climate Dynamics*, 43, 1303–1310.

Annan, J.D. and Hargreaves, J.C. (2017) On the meaning of independence in climate science. *Earth System Dynamics*, 8, 211–224.

Bishop, C.H. and Abramowitz, G. (2013) Climate model dependence and the replicate earth paradigm. *Climate Dynamics*, 41, 885–900.

Breiman, L. (2001) Random forests. *Machine Learning*, 45, 5–32.

Chen, S.-T., Yu, P.-S. and Tang, Y.-H. (2010) Statistical downscaling of daily precipitation using support vector machines and multivariate analysis. *Journal of Hydrology*, 385, 13–22.

Chiyuan, M., Qingyun, D., Qiaohong, S., Yong, H., Dongxian, K., Tiantian Y., Aizhong Y., Zhenhua, D. and Wei, G. (2014) Assessment of CMIP5 climate models and projected temperature changes over northern Eurasia. *Environmental Research Letters*, 9, 055007.

Cortes, C. and Vapnik, V. (1995) Support-vector networks. *Machine Learning*, 20, 273–297.

Deo, R.C. and Şahin, M. (2015) Application of the extreme learning machine algorithm for the prediction of monthly effective drought index in eastern Australia. *Atmospheric Research*, 153, 512–525.

Herger, N., Abramowitz, G., Knutti, R., Angélil, O., Lehmann, K. and Sanderson, B.M. (2017) Selecting a climate model subset to optimise key ensemble properties. *Earth System Dynamics: Discussion*, 9, 135–151.

Heung, B., Bulmer, C.E. and Schmidt, M.G. (2014) Predictive soil parent material mapping at a regional-scale: a random forest approach. *Geoderma*, 214–215, 141–154.

Hoeting, J.A., Madigan, D., Raftery, A.E. and Volinsky, C.T. (1999) Bayesian model averaging: a tutorial. *Statistical Science*, 14, 382–401.

Hou, Y.-K., Chen, H., Xu, C.-Y., Chen, J. and Guo, S.-L. (2017) Coupling a Markov chain and support vector machine for at-site downscaling of daily precipitation. *Journal of Hydrometeorology*, 18, 2385–2406.

Ines, A.V.M. and Hansen, J.W. (2006) Bias correction of daily GCM rainfall for crop simulation studies. *Agricultural and Forest Meteorology*, 138, 44–53.

IPCC. (2013) In: Stocker, T.F. , Qin, D. , Plattner, G.-K. , Tignor, M. , Allen, S. K. , Boschung, J. , Nauels, A. , Xia, Y. , Bex, V. and Midgley, P.M. (Eds.) *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge and New York, NY: Cambridge University Press.

Jeffrey, S.J., Carter, J.O., Moodie, K.B. and Beswick, A.R. (2001) Using spatial interpolation to construct a comprehensive archive of Australian climate data. *Environmental Modelling and Software*, 16, 309–330.

Kerkhoff, C., Künsch, H.R. and Schär, C. (2015) A Bayesian hierarchical model for heterogeneous RCM–GCM multimodel ensembles. *Journal of Climate*, 28, 6249–6266.

Kuhn, M. (2008) Building predictive models in R using the caret package. *Journal of Statistical Software*, 28, 26.

Kumar, A., Mitra, A.K., Bohra, A.K., Iyengar, G.R. and Durai, V.R. (2012) Multi-model ensemble (MME) prediction of rainfall using neural networks during monsoon season in India. *Meteorological Applications*, 19, 161–169.

Lambert, S.J. and Boer, G.J. (2001) CMIP1 evaluation and intercomparison of coupled climate models. *Climate Dynamics*, 17, 83–106.

Leduc, M., Laprise, R., Elía, R.D. and Šeparović, L. (2016) Is institutional democracy a good proxy for model independence? *Journal of Climate*, 29, 8301–8316.

Liu, D.L. and Zuo, H. (2012) Statistical downscaling of daily climate variables for climate change impact assessment over New South Wales, Australia. *Climatic Change*, 115, 629–666.

Madadgar, S. and Moradkhani, H. (2014) Improved Bayesian multimodeling: integration of copulas and Bayesian model averaging. *Water Resources Research*, 50, 9586–9603.

Mendlik, T. and Gobiet, A. (2016) Selecting climate simulations for impact studies based on multivariate patterns of climate change. *Climatic Change*, 135, 381–393.

Naghibi, S.A., Ahmadi, K. and Daneshi, A. (2017) Application of support vector machine, random forest, and genetic algorithm optimized random forest models in groundwater potential mapping. *Water Resources Management*, 31, 2761–2775.

Oh, S.-G. and Suh, M.-S. (2016) Comparison of projection skills of deterministic ensemble methods using pseudo-simulation data generated from multivariate Gaussian distribution. *Theoretical and Applied Climatology*, 129, 243–262.

Okkan, U. and Inan, G. (2015) Statistical downscaling of monthly reservoir inflows for Kemer watershed in Turkey: use of machine learning methods, multiple GCMs and emission scenarios. *International Journal of Climatology*, 35, 3274–3295.

Qi, Y., Qian, C. and Yan, Z. (2016) An alternative multi-model ensemble mean approach for near-term projection. *International Journal of Climatology.*, 37, 109–122.

Raftery, A.E., Gneiting, T., Balabdaoui, F. and Polakowski, M. (2005) Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133, 1155–1174.

Robertson, A.W., Lall, U., Zebiak, S.E. and Goddard, L. (2004) Improved combination of multiple atmospheric GCM ensembles for seasonal prediction. *Monthly Weather Review*, 132, 2732–2744.

Salcedo-Sanz, S., Deo, R.C., Carro-Calvo, L. and Saavedra-Moreno, B. (2016) Monthly prediction of air temperature in Australia and New Zealand with

machine learning algorithms. *Theoretical and Applied Climatology*, 125, 13–25.

Sanderson, B.M., Knutti, R. and Caldwell, P. (2015a) Addressing interdependency in a multimodel ensemble by interpolation of model properties. *Journal of Climate*, 28, 5150–5170.

Sanderson, B.M., Knutti, R. and Caldwell, P. (2015b) A representative democracy to reduce interdependency in a multimodel ensemble. *Journal of Climate*, 28, 5171–5194.

Sarhadi, A., Burn, D.H., Johnson, F., Mehrotra, R. and Sharma, A. (2016) Water resources climate change projections using supervised nonlinear and multivariate soft computing techniques. *Journal of Hydrology*, 536, 119–132.

Sarhadi, A., Burn, D.H., Yang, G. and Ghodsi, A. (2017) Advances in projection of climate change impacts using supervised nonlinear dimensionality reduction techniques. *Climate Dynamics*, 48, 1329–1351.

Tao, Y., Yang, T., Faridzad, M., Jiang, L., He, X. and Zhang, X. (2018) Non-stationary bias correction of monthly CMIP5 temperature projections over China using a residual-based bagging tree model. *International Journal of Climatology*, 38, 467–482.

Taylor, K.E. (2001) Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres*, 106, 7183–7192.

Taylor, K.E., Stouffer, R.J. and Meehl, G.A. (2012) An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, 93, 485–498.

Tebaldi, C. and Knutti, R. (2007) The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 365, 2053–2075.

Trewin, B. (2013) A daily homogenized temperature data set for Australia. *International Journal of Climatology*, 33, 1510–1529.

Vu, M.T., Aribarg, T., Supratid, S., Raghavan, S.V. and Liong, S.-Y. (2016) Statistical downscaling rainfall using artificial neural network: significantly wetter Bangkok? *Theoretical and Applied Climatology*, 126, 453–467.

Wallach, D., Mearns, L.O., Ruane, A.C., Rötter, R.P. and Asseng, S. (2016) Lessons from climate modeling on the design and use of ensembles for crop modeling. *Climatic Change*, 139, 551–564.

Wang, B., Liu, D.L., Macadam, I., Alexander, L.V., Abramowitz, G. and Yu, Q. (2016a) Multi-model ensemble projections of future extreme temperature change using a statistical downscaling method in south eastern Australia. *Climatic Change*, 138, 85–98.

Wang, B., Waters, C., Orgill, S., Cowie, A., Clark, A., Li, L.D., Simpson, M., McGowen, I. and Sides, T. (2018) Estimating soil organic carbon stocks using different modelling techniques in the semi-arid rangelands of eastern Australia. *Ecological Indicators*, 88, 425–438.

Wang, L., Ranasinghe, R., Maskey, S., Van Gelder, P.H.A.J.M. and Vrijling, J. K. (2016b) Comparison of empirical statistical methods for downscaling daily climate projections from CMIP5 GCMs: a case study of the Huai River basin, China. *International Journal of Climatology*, 36, 145–164.

Wang, W., Shao, Q., Yang, T., Yu, Z., Xing, W. and Zhao, C. (2014) Multimodel ensemble projections of future climate extreme changes in the Haihe River basin, China. *Theoretical and Applied Climatology*, 118, 405–417.

Were, K., Bui, D.T., Dick, Ø.B. and Singh, B.R. (2015) A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. *Ecological Indicators*, 52, 394–403.

Yang, T., Hao, X., Shao, Q., Xu, C.-Y., Zhao, C., Chen, X. and Wang, W. (2012a) Multi-model ensemble projections in temperature and precipitation extremes of the Tibetan Plateau in the 21st century. *Global and Planetary Change*, 80, 1–13.

Yang, T., Li, H., Wang, W., Xu, C.Y. and Yu, Z. (2012b) Statistical downscaling of extreme daily precipitation, evaporation, and temperature and construction of future scenarios. *Hydrological Processes*, 26, 3510–3523.

Zeugner, S. and Feldkircher, M. (2015) Bayesian model averaging employing fixed and flexible priors: the BMS package for R. *Journal of Statistical Software*, 68(4), 1–37.

Zhuang, X.W., Li, Y.P., Huang, G.H. and Liu, J. (2016) Assessment of climate change impacts on watershed in cold-arid region: an integrated multi-GCM-based stochastic weather generator and stepwise cluster analysis method. *Climate Dynamics*, 47, 191–209.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Wang B, Zheng L, Liu DL, Ji F, Clark A, Yu Q. Using multi-model ensembles of CMIP5 global climate models to reproduce observed monthly rainfall and temperature with machine learning methods in Australia. *Int J Climatol*. 2018;38: 4891–4902. https://doi.org/10.1002/joc.5705