A data-driven multi-model ensemble for deterministic and probabilistic precipitation forecasting at seasonal scale

Article *in* Climate Dynamics · April 2020

DOI: 10.1007/s00382-020-05173-x

CITATIONS
48

READS
990

4 authors:

Lei Xu
China University of Geosciences
49 PUBLICATIONS **1,778** CITATIONS
SEE PROFILE

Nengcheng Chen
Wuhan University
236 PUBLICATIONS **4,191** CITATIONS
SEE PROFILE

Zhang Xiang
China University of Geosciences
132 PUBLICATIONS **3,425** CITATIONS
SEE PROFILE

Zeqiang Chen
Wuhan University
82 PUBLICATIONS **1,676** CITATIONS
SEE PROFILE

Climate Dynamics
https://doi.org/10.1007/s00382-020-05173-x

A data-driven multi-model ensemble for deterministic and probabilistic precipitation forecasting at seasonal scale

Lei Xu[1] · Nengcheng Chen[1,2] · Xiang Zhang[1] · Zeqiang Chen[1]

Received: 9 September 2019 / Accepted: 8 February 2020

Abstract
Seasonal precipitation forecasting is valuable for regional water management and agricultural food security. Current numerical models have large uncertainty in model structure, parameterization and initial conditions. Here, a data-driven multi-model ensemble is constructed using a series of statistical and machine learning methods with varying inputs. Deterministic precipitation forecasts are produced by the weighting of ensemble members using Bayesian model averaging (BMA) and probabilistic forecasts are generated by sampling from BMA predictive probability density function (PDF). Three mathematical metrics are used to evaluate the performance of precipitation forecasts, including Pearson's correlation coefficient (PCC), root mean square error skill score (RMSESS) and continuous ranked probability skill score (CRPSS). The results demonstrate that the accuracy in the statistical ensemble is significantly higher than the North American multi-model ensemble (NMME) for both deterministic and probabilistic precipitation forecasts, especially at 1-month lead. Statistical models are considerably enhanced by incorporating wavelets, which decomposes the raw precipitation series into several different levels, potentially representing underlying precipitation patterns at different time-frequency scales. Selecting some good ensemble members can improve the ensemble performance, instead of including all the ensemble members with some inefficient models. Overall, the statistical ensemble can be considered as an effective complement of numerical models in both deterministic and probabilistic precipitation forecasts.

1 Introduction

Precipitation is a crucial atmospheric variable influencing global water cycle, reflecting changes and transitions of earth's energy budgets (Bosilovich et al. 2011; Trenberth et al. 2007). Excessive or deficient precipitation can lead to severe floods or droughts, which may cause enormous socioeconomic losses, such as the Yangtze floods in 1998 (Zong and Chen 2000) and the southern China drought in 2010 (Yang et al. 2012). Effective precipitation forecasts up to several months in advance can provide vital information for disaster early warning and preparations. For example, the climate smart agriculture (Lipper et al. 2014) utilizes seasonal precipitation forecasts to guide agricultural practices in Africa in order to reduce crop production loss and ensure food security. Therefore, accurate and reliable precipitation forecasts are invaluable for agricultural applications (Ingram et al. 2002). Seasonal precipitation forecasting is also helpful in socioeconomic decision-makings such as regional water management and waterway navigation alerts. Current seasonal precipitation forecasting models can be generally divided into statistical data-driven approaches and physically-based numerical models (Cuo et al. 2011; Hao et al. 2018).

Statistical methods relate hydrometeorological and auxiliary variables to precipitation through mathematical modeling based on historical data. The established relationship is then used for out-of-sample data to make forecasts. Numerous regression algorithms can be used to model the relationship but their performances vary. Traditional correlation-based methods or heuristic machine learning models are both widely examined in precipitation forecasting in existing studies (Choubin et al. 2016; Maldonado et al. 2013; Partal

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00382-020-05173-x) contains supplementary material, which is available to authorized users.

✉ Xiang Zhang
zhangxiangsw@whu.edu.cn

[1] State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430079, China

[2] Collaborative Innovation Center of Geospatial Technology, Wuhan 430079, China

Published online: 17 February 2020

and Kişi 2007; Shi et al. 2015), such as linear regression, artificial neural network (ANN), random forest (RF) and deep learning. In practice, the selection of suitable models is dependent on the specific applications and the characteristic of algorithms. Some models are easy to implement but have a low accuracy, while others may need sophisticated training before reaching good performance. The selection of models is dependent on the problems, data and corresponding requirements. In statistical models, different statistical models may have comparable forecasting performance (Xu et al. 2018b) towards a specific variable. This is related to the intrinsic mathematical mechanisms within different statistical models as they come from different theorems or theories. Combining the strengths of different algorithms is difficult from a theoretical perspective.

In the statistical models, the predictors are selected based on the relevant variables from previous months. For example, the temperature and precipitation in the last month should have an effect on the precipitation formulation in the target month. Large-scale climate oscillations, such as EI Niño-Southern Oscillation (ENSO), are good indicators of global extreme precipitation. SST anomaly in remote tropical oceans has an impact on regional precipitation in China (Chan and Zhou 2005; Kripalani and Kulkarni 2001; Xiao et al. 2015). The Yangtze floods in 1998 are closely related with the 1997–1998 ENSO event (Lieting 2001). Four teleconnection indices, i.e. Nino 1 + 2, Nino 3.4, North Atlantic Oscillation (NAO) and Dipole Mode Index (DMI), are collected as potential predictors of precipitation as they are found to be relevant with seasonal precipitation in some areas of China. Temperature is considered another exogenous variable influencing precipitation. The monthly maximum, minimum and average temperature data are acquired from Climatic Research Unit Timeseries (CRU TS) Version 4.01 (Harris et al. 2014).

Numerical weather models are soundly based on physical mechanism and seem more convincing in interpretations than statistical models. Dynamical models use physical equations and physical models to simulate the atmosphere–ocean–land interactions. Therefore, numerical models are based on physical mechanism instead of data-driven. Statistical models perform forecasts based on historical data and are data driven. Statistical models do not model the movement and interactions of wind, clouds and moisture, while numerical models do simulate. In this aspect, physically-based dynamical models seem more convincing than data-driven statistical models because the former can simulate how the precipitation is generated. Numerical weather models simulate the atmosphere–ocean–land interactions using general circulation models (GCMs), quantifying earth's climate system using physical equations (Bauer et al. 2015; Stensrud 2009). A lot of climate models were developed for weather simulation and prediction in several major research centers of the world (McFarlane et al. 1992; Molteni et al. 1996; Roeckner et al. 2003; Saha et al. 2014). The recently developed North American multi-model ensemble (NMME) is a global climate forecasting system incorporating a number of climate models developed in several climate modeling centers and can provide seasonal precipitation forecasts for a few months in advance (Kirtman et al. 2014). The NMME has been widely used in forecasting sea surface temperature (SST), precipitation and soil moisture (Becker et al. 2014; Slater et al. 2017; Thober et al. 2015; Xu et al. 2019). A distinctive feature of the NMME lies in the advantage of multi-model ensemble over single models in climatic forecasts. This superiority in multiple models has also been examined in climate projections and hydrological applications, such as the Coupled Model Inter-comparison Project Phase 5 (CMIP5) (Taylor et al. 2012) and the Inter-Sectoral Impact Model Inter-comparison Project (ISIMIP) (Warszawski et al. 2014). Given the uncertainty in model structure, parameters and initial conditions in numerical models, it is promising to use multi-model ensembles than using a single model in precipitation forecasting (Khajehei et al. 2018; Khajehei and Moradkhani 2017; Krishnamurti et al. 2016).

Currently, statistical models seem to have better accuracy than numerical models in precipitation forecasting, as shown in some studies (Hao et al. 2018; Xu et al. 2018a; Xu et al. 2018b). This is probably because that large uncertainties exist in physically-based numerical models, while data-driven statistical models can fit historical data by elaborate mathematical modeling. Both data-driven and physically-based methods are extensively used in precipitation forecasting today (Darji et al. 2015; Pokhrel et al. 2016). Numerical models dominate in climate science because they have plausible physical meaning. However, empirical statistical relationships explored from the historical data may provide useful information beyond current physical models, as potential factors influencing global and regional climate systems may not be well considered. Therefore, data-driven statistical models can be regarded as a complement to physically-based models (Reichstein et al. 2019).

Weighting multi-model ensembles is a way to consider model uncertainty when generating deterministic precipitation forecasts. Simple ensemble mean can be an easy and effective way to combine ensemble members. Some other weighting schemes such as Bayesian model averaging (BMA) and machine learning, may provide more accurate results towards ensemble modeling (Slater et al. 2017; Xu et al. 2018b; Zaherpour et al. 2019). The data-driven ensemble-based approaches are becoming increasingly popular in water resources modeling and hydrometeorological forecasts in recent years (Li et al. 2019; Nourani et al. 2018; Quilty et al. 2019). For example, Li et al. (2019) considered the ensemble member selection and weighting uncertainties by a multiwavelet ensemble stochastic forecasting framework.

Berkhahn et al. (2019) used an ensemble of ANNs to forecast urban floods in real time. The data-driven ensemble-based methods outperform the single-model forecasts in the fact that the former can provide both deterministic and probabilistic forecasts, versus the deterministic forecasts in single models. The data-driven ensemble-based models also have advantages in improving forecasting accuracy and robustness (Berkhahn et al. 2019; Quilty et al. 2019) and can achieve equitable or even better performance than physical models (Ham et al. 2019; Xu et al. 2018b). Some researchers suggest that weighting only good members is a good choice while other studies think that bad members should also be included (Najafi et al. 2012). There is not a unique rule determining whether using all the ensemble members or only some of them. Generally, weighting multiple models to produce a multi-model ensemble-based combination can result in better forecasting performance than the single best model. This is usually the case. Theoretically, although ensemble-based model combinations may not achieve always better performance than a single best model because some ensemble members may deteriorate the ensemble results, the ensemble-based combinations can be used to reduce model overfitting and quantify the predictive model uncertainty and can improve the forecasting accuracy relative to the majority of the single models.

Ensembles of numerical models have been increasingly developed in meteorological science, while statistical ensembles are scarcely seen in precipitation forecasting. Theoretically, it is possible to construct statistical ensembles to combine model strengths and reduce the uncertainty. It remains to be seen whether statistical ensembles contrast with numerical ensembles equivalently. In this work, a data-driven statistical multi-model ensemble is developed to forecast seasonal precipitation. How well the statistical ensemble in predicting precipitation both deterministically and probabilistically is compared with an advanced numerical ensemble, i.e., the NMME. Relative to the booming development of numerical models, we highlight that statistical multi-model ensembles can also be developed as an alternative way of examining precipitation forecasting uncertainty and are a strong complement to numerical models. Using all the available ensemble members or only some good members for ensemble prediction is also investigated. The rest of this study is organized as follows. Section 2 describes the study area and used data. Section 3 introduces the methodology, including statistical models, numerical ensembles and the weighting scheme. Section 4 demonstrates the result. Section 5 performs discussions and Sect. 6 makes the conclusion.

2 Study area and data

China is selected as the study area, which is located in the eastern part of Asia. Annually averaged rainfall decreases from south to north and from east to west (Fig. 1). Some arid areas in northwestern China have an annual rainfall of less than 30 mm, while areas in southern China may have an annual rainfall of as much as 3000 mm. The precipitation data is obtained from Global Precipitation Climatology Centre (GPCC) (Becker et al. 2011) version 7 from 1901 to 2013 and the monitoring product version 4 from 2014 to present at a monthly 1° resolution. Precipitation data in northwestern areas may suffer from large uncertainty due to less meteorological stations relative to southeastern China.

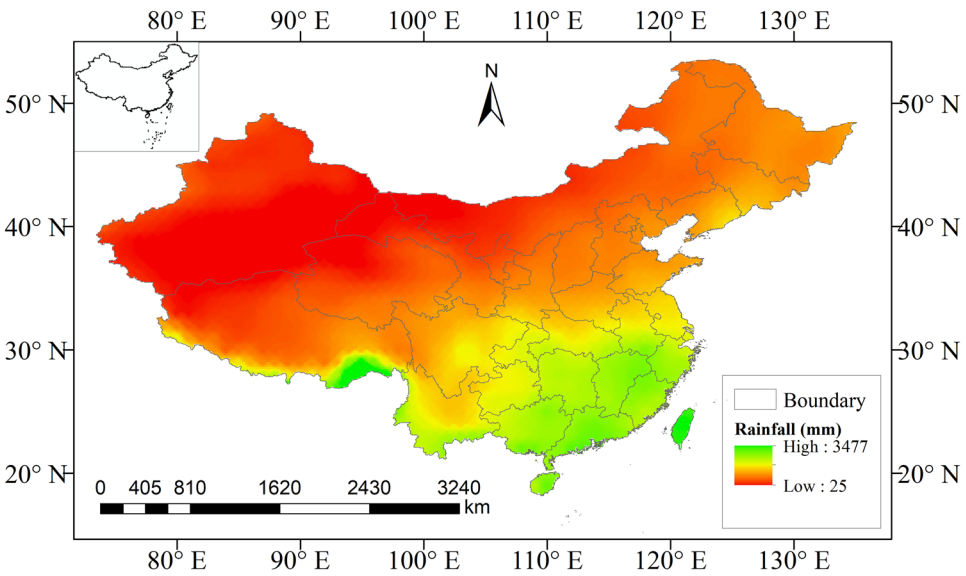

Fig. 1 Multi-year annually averaged rainfall during 1960–2016 in China

3 Methods

In this study, statistical and numerical ensembles are both constructed by a series of models (Table 1) based on different variable selection schemes, algorithms and preprocessing procedures. Statistical ensembles are made up of a combination of algorithms and variables. First, 6 models with only precipitation variable are considered as input and output. Then, the inclusion of climate indices and temperature, the inclusion of wavelet preprocessing and the inclusion of both of them are regarded as ensemble members. As a result, a total of 24 statistical models are generated. Deterministic precipitation forecasting is obtained by weighting ensemble members using BMA and probabilistic forecasts are generated by sampling from BMA predictive PDF described in the Sect. 3.3. Table 1 lists the statistical and numerical ensembles for 1-month lead forecast, while the forecasts for 2–6 months ahead are also conducted. The linear model refers to linear regression based on ordinary least squares algorithm. All the statistical and BMA based models are trained on each grid cell separately except convolutional long short-term memory network (ConvLSTM) based models (model 21–24), which are trained using all the grid cells over the study area.

There are four types of inputs in the forecasting experiments. The type I inputs only include precipitation in previous months. Specifically, the previous precipitation up to six months in advance are regarded as inputs in 1-month lead forecasts, i.e. {{P(t-6), P(t-5), P(t-4), P(t-3), P(t-2), P(t-1)}}. The type II inputs include precipitation, average temperature, minimum temperature, maximum temperature and climate indices. Here, only the lagged 1-month lead variables are included in 1-month lead forecasts, i.e. {{P(t-1), Tave(t-1), Tmin(t-1), Tmax(t-1), C(t-1)}}, due to consideration of the number of predictors. Type III models consider the decomposed precipitation components by wavelets as inputs. In 1-month lead forecasts, the decomposed precipitation components up to three months in advance, i.e. {{DP(t-3), DP(t-2), DP(t-1)}}, are regarded as inputs. One decomposed precipitation component includes one approximation series and three details, i.e. four variables. Type IV incorporates average temperature, minimum temperature, maximum temperature and climate indices apart from the decomposed precipitation components. The 1-month lead precipitation forecasts in type IV include corresponding 1-month lagged variables, i.e. {{DP(t-3), DP(t-2), DP(t-1), Tave(t-1), Tmin(t-1), Tmax(t-1), C(t-1)}}.

The predictor lags are determined by considering the number of variables. In 1-month lead forecasting, Type I experiments have precipitation from t-6 to t-1 {{P(t-6), P(t-5), P(t-4), P(t-3), P(t-2), P(t-1)}}, i.e. 6 variables. Type II experiments include 9 variables (4 climate indices), i.e. {{P(t-1), Tave(t-1), Tmin(t-1), Tmax(t-1), C(t-1)}}. If the lag of two months is used, a total of 18 variables will be included in Type II experiments, which may result in

Table 1 A summary of experimental design of statistical and numerical ensembles for 1-month lead precipitation forecasting

Number	Models	Inputs	Outputs
1	Linear	I	P(t)
2	Linear_CT	II	P(t)
3	W_Linear	III	P(t)
4	W_Linear_CT	IV	P(t)
5	SVM	I	P(t)
6	SVM_CT	II	P(t)
7	W_SVM	III	P(t)
8	W_SVM_CT	IV	P(t)
9	RF	I	P(t)
10	RF_CT	II	P(t)
11	W_RF	III	P(t)
12	W_RF_CT	IV	P(t)
13	ANN	I	P(t)
14	ANN_CT	II	P(t)
15	W_ANN	III	P(t)
16	W_ANN_CT	IV	P(t)
17	LSTM	I	P(t)
18	LSTM_CT	II	P(t)
19	W_LSTM	III	P(t)
20	W_LSTM_CT	IV	P(t)
21	ConvLSTM	I	P(t)
22	ConvLSTM_CT	II	P(t)
23	W_ConvLSTM	III	P(t)
24	W_ConvLSTM_CT	IV	P(t)
25	StBMA	Model 1–24	P(t)
26	CMC1-CanCM3	–	P(t)
27	CMC2-CanCM4	–	P(t)
28	COLA-RSMAS-CCSM4	–	P(t)
29	GFDL-CM2p1-AER04	–	P(t)
30	GFDL-CM2p5-FLOR-A06	–	P(t)
31	GFDL-CM2p5-FLOR-B01	–	P(t)
32	NASA-GMAO-062012	–	P(t)
33	NCEP-CFSV2	–	P(t)
34	NuBMA	Model 26–33	P(t)

Input of type I: {{P(t-6), P(t-5), P(t-4), P(t-3), P(t-2), P(t-1)}}; Input of type II: {{P(t-1), Tave(t-1), Tmin(t-1), Tmax(t-1), C(t-1)}}; Input of type III: {{DP(t-3), DP(t-2), DP(t-1)}}; Input of type IV: {{DP(t-3), DP(t-2), DP(t-1), Tave(t-1), Tmin(t-1), Tmax(t-1), C(t-1)}}*DP* decomposed precipitation, *Tave* average temperature, *Tmin* minimum temperature, *Tmax* maximum temperature, *C(t-1)* climate index at time t-1, *P(t)* precipitation at month t; Models with a prefix 'W_' represent that they are preprocessed by wavelets. Models with a suffix '_CT' indicate that they include climate indices and temperature in the inputs. StBMA refers to statistical BMA model and NuBMA refers to numerical BMA model

overfitting for small data samples. We also use 2-month lags to see if the predictive power can be improved. In fact, there is some improvement when incorporating more lags in the experiment using linear regression forecasts (not shown). It also has some improvement in Type I experiments when incorporating more lags. However, we focus on the model ensemble here and set the lag as 1 month in Type II currently.

Precipitation forecasts at all the leads are one-step forecast except long short-term memory network (LSTM) and ConvLSTM models which are multistep forecasts by taking the advantage of temporal memory effect. One-step forecast means that the forecasts at 1–6 months ahead are all conducted by one step. For example, P(t + 1) is directly trained and forecasted using {{P(t-6), P(t-5), P(t-4), P(t-3), P(t-2), P(t-1)}} for 2-month lead forecast in linear model. Multistep forecasts denote that the observed data {{P(t-6), P(t-5), P(t-4), P(t-3), P(t-2), P(t-1)}} is first used to forecast the precipitation at month t, i.e. P(t). Then, the observed and forecasted data {{P(t-5), P(t-4), P(t-3), P(t-2), P(t-1), P_f(t)}} are concatenated to forecast the precipitation at t + 1 month, i.e. P(t + 1).

In LSTM based models, the predictors are normalized to the range (0, 1) by a linear transformation using MinMaxScaler in scikit-learn. In SVM, the predictors are normalized by centering and scaling each column of the predictor data by the weighted column mean and standard deviation, respectively (implemented in MATLAB R2018a). For other models, the predictors are not normalized. No normalization may cause a longer time of training in optimization-based method like ANN. As a total of 1000 iteration times are used in ANN in our study, no normalization may not be an issue. Normalization is not needed in random forests as it is invariant to monotonic transformations of individual features. Currently, we haven't deal with the collinearity among the predictors using some preprocessing techniques like principal component analysis (PCA). The collinearity may be less an issue as we are only concerned about the predictive power of models instead of the causality relationship between variables or coefficient estimates.

Although different predictors are used in Table 1, the contribution of one predictor to the predictive performance may be replaced or compensated by another predictor. In Table 1, the lagged precipitation is used as predictors in the type I experiment. In the type II experiment, the lagged precipitation together with temperature and climate indices are used. It is possible that these two experiments can result in similar predictive performance because the predictors are used in a data-driven way. The forecasting results change a little (mean correlation from 0.68 to 0.67) when leaving P(t-1) predictor in Type II experiment based on linear regression method. This may be a result of the important contribution to the predictability by climate indices and temperature (Choubin et al. 2016; Mortensen et al. 2018; Xu et al. 2018a). That is, leaving a predictor in Type II inputs may cause little change in the predictive performance because other predictors can make up the effect of the lacked predictor. Although P(t-1) plays an important role in Type II (1-month lead) precipitation forecasts, its role can be replaced or compensated by other predictors such as climate indices and temperature from a data-driven view.

Precipitation forecasts are conducted at 1–6-month lead for both statistical and numerical models. Statistical forecasts start from 1960 and end at 2016, with the training period spanning from 1960 to 2010 and the rest as the validating period. Numerical forecasts have a hindcasting period during 1982–2010 and the validating period from 2011 to 2016.

3.1 Statistical models

3.1.1 Support vector machine

Support vector machine (SVM) is a machine learning method used for classification and regression based on Vapnik–Chervonenkis (*VC*) dimension theory and the rule of structural risk minimization (Cortes and Vapnik 1995; Drucker et al. 1997; Vapnik 2013). SVM can solve nonlinear problems by introducing kernel functions, mapping the data into a high-dimension linear space. SVM is widely used in forecasting science and has demonstrated good capability in nonlinear modeling, image classification and high-dimension representation (Tong and Koller 2001; Wang 2005), especially for small samples. Consider a linear regression model as follows

$$f(x) = \langle w, x \rangle + b \tag{1}$$

where $\langle w, x \rangle$ represents the dot product of w and x; w is support vectors and b denotes the constant; $x \in \mathbb{R}^n$. For ε-SVM, a margin of tolerance ε is allowed to make mistakes. In this case, an optimal hyperplane can be decided by maximizing margins under some constraints

$$\min \frac{1}{2} \| w \|^2, \quad s.t. \begin{cases} y_i - \langle w, x_i \rangle - b \le \epsilon \\ \langle w, x_i \rangle + b - y_i \le \epsilon \end{cases} \tag{2}$$

To prevent overfitting, slack variables can be added into Eq. (2)

$$\min \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{\iota} (\xi_i + \xi_i^*), \quad s.t. \begin{cases} y_i - \langle w, x_i \rangle - b \le \epsilon + \xi_i \\ \langle w, x_i \rangle + b - y_i \le \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases} \tag{3}$$

where ξ_i and ξ_i^* are slack variables; The parameter C controls the balance between forecasting errors and maximal

margins. To solve Eq. (3), this problem could be transformed into its Lagrange dual formulas. A detailed deduction process can refer to Smola and Schölkopf (2004). The solution of Eq. (3) can be expressed as

$$f(x) = \sum_{i=1}^{l} \left(a_i - a_i^*\right) K(x_i, x) + b \tag{4}$$

where a_i and a_i^* are Lagrange multipliers; K denotes kernel function, e.g. linear, sigmoid and radial basis function (RBF).

SVM models are implemented using MATLAB R2018a in our study. Three major parameters can be adjusted to optimize the model, i.e. box constraint (C), tolerance (ε) and kernel width (σ). The box constraint C is set as $iqr(y)/1.349$, where $iqr(y)$ denotes the interquartile range of response variable y. The tolerance ε is set as $iqr(y)/1.349$, which is an estimation of standard deviation. The RBF kernel is used and the width σ is obtained from a heuristic subsampling process.

3.1.2 Artificial neural network

ANN is popular in nonlinear modeling through a series of neural layers and nodes (Haykin 1994). ANN can model sophisticated problems by adjusting network depth and the number of neurons. A well-trained ANN based on large amounts of data has potential in forecasting out-of-sample data. ANN has undergone rapid development with increasing computer processing capability in the last decades. Here, a three-layer feedforward ANN is used to establish the relationship between precipitation and relevant variables, including one input layer, one hidden layer and one output layer. The back propagation (BP) algorithm (Rumelhart et al. 1988) is used to train the network. The three-layer feedforward ANN (Kim and Valdés 2003) can be expressed as

$$\hat{y}_k = f_o\left[\sum_{j=1}^{N} w_{kj} z_j + w_{ko}\right] \tag{5}$$

$$z_j = f_h\left(\sum_{i=1}^{M} w_{ji} x_i + w_{jo}\right) \tag{6}$$

where f_o denotes the activation function in the output neuron; w_{kj} is the weight between kth neuron in the output layer and jth neuron in the hidden layer; N is the number of nodes in hidden layers; z_j denotes the intermediate input from hidden layer to output layer; w_{ko} is the bias for kth output neuron; f_h is the activation function in the hidden layer; w_{ji} is the weight between jth neuron in the hidden layer and ith neuron in the input layer; M is the number of nodes in input layers; x_i represents the input data; w_{jo} is the bias for jth hidden neuron.

In this study, the number of nodes in the input layers is determined by input data and the number of nodes in hidden layers is automatically selected between 5 and 10 by a trial and error procedure. The number of training epochs is set as 1000. The ANN is implemented using MATLAB R2018a. The activation function in the output neuron f_o is the linear transfer function (purelin) and the activation function in the hidden layer f_h is hyperbolic tangent sigmoid transfer function (tansig). The linear activation function is selected in the output neuron f_o due to the final mapping of the outputs from the hidden layer to precipitation. The linear transfer function can map the outputs between -1 and $+1$ from the hidden layer to the absolute values of precipitation. Therefore, the linear transfer function is suitable for the activation function from hidden layer to output layer. The tangent sigmoid transfer function is used in the hidden layer because of the suitable bound between -1 and $+1$ and continuous outputs characteristic by the tansig function. The magnitude of the gradient during the training process and the number of training epochs are used to stop the training. If the magnitude of the gradient is less than $1e-5$, the training process will stop. When the training epochs reach up to 1000, the training will also stop.

3.1.3 Random forest

RF is an ensemble learning algorithm (Breiman 2001) using a number of decision trees. RF can be regarded as a result of randomly selected samples and randomly selected variables in an ensemble of trees. Three major factors can be adjusted to change model structure, i.e. the number of trees (*ntree*) bootstrapped from training data, the number of predictors (*mtry*) randomly selected at each split, and the minimal size of terminal nodes (*nodesize*) in each tree. A typical RF algorithm can be described as follows. First, *ntree* samples are bootstrapped from the training data, corresponding approximately two thirds of the raw dataset. Then an unpruned tree is grown by selecting about one third of predictors and the best split is determined according to the predictors. The trained *ntree* trees are then used in classification or regression. Finally, the ensemble results obtained from *ntree* trees can be integrated by an average or major voting method.

In our experiment, all predictors are selected without randomly selection because they may be all useful for determining the splits. This is the same as bagging only using randomly selected samples. The number of trees is set as 20 and the minimal terminal node is set as 5. The RF method is implemented using MATLAB R2018a.

3.1.4 LSTM and convolutional LSTM

LSTM (Hochreiter and Schmidhuber 1997) was proposed to solve some drawbacks in recurrent neural networks (RNNs), e.g. decaying error backflow. LSTM incorporates temporal memory effect by using a number of recurrently connected memory blocks. A memory block includes three major units: input gate, output gate and forget gate (Gers et al. 1999). These gates can be activated or closed to control the flow of error. A typical LSTM memory block is demonstrated in Fig. 2.

The forget gate layer f_t, determines which information should be kept or forgotten using hidden state and current input and produces an output between 0 and 1. An output of 0 indicates the information is not kept and a value of 1 means the information is totally retained. σ is the sigmoid function. The input gate layer i_t determines which information should be updated and new information g_t is then added based on input data and hidden state. Cell state is then updated using input gate layer and new information. The network output is generated by the output layer gate o_t and the hidden state is updated using the output and cell state. A detailed formula derivation can be found in Hochreiter and Schmidhuber (1997).

LSTM has shown good capability in sequence modeling, such as speech recognition and water flow forecasting. Here, LSTM is examined in temporal precipitation prediction. The parameters influencing network performance are adjusted by trial and error. In our experiment, the number of hidden units is set as 50. Training epochs are limited at 1000 and the batch size is set as the length of all the training data at each grid.

As LSTM models temporal correlation of a dataset, the spatial connections are not well considered. To account for spatial features, ConvLSTM was developed (Shi et al. 2015) to incorporate spatial information. Convolutional operation can be utilized to extract spatial representations of multi-dimensional data by different kernel filters. The number and shape of convolutional kernels can be adjusted to fit the data and problem. Although precipitation may have low spatial correlation, however, the climate system is spatially continuous and mutually influenced. Thus, the spatial information can be incorporated to see if the forecasting performance can be improved. The training process of convolutional LSTM is not grid by grid but a multidimensional image. The relevant parameters are determined by manual adjustment. LSTM and ConvLSTM are implemented using Tensorflow 1.4.0 with Python programming language.

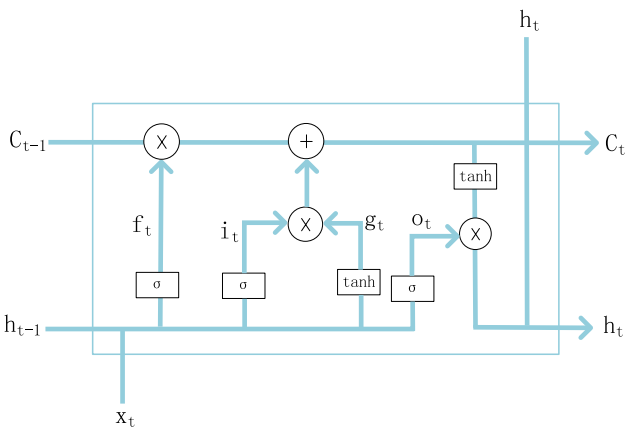

Fig. 2 A basic memory block of LSTM

3.1.5 Wavelet

Wavelet is an effective tool in signal decomposition, representation and processing (Heil and Walnut 1989; Mallat 1989). It can decompose a signal into multiresolution levels in time-frequency space by a modulated window. Compared to Fourier transform, wavelet can represent the time-frequency dependence at any location while Fourier only models frequency change. Wavelet decomposition can be used to identify the frequency, cycle and noise of a dataset at multiple scales. The continuous wavelet transform (CWT) of a sequence can be expressed as

$$\mathrm{W(a,b)} = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) \psi^* \left(\frac{t-b}{a} \right) dt \tag{7}$$

where a is the scale parameter and b is the translation; ψ is the mother wavelet and * denotes complex conjugate.

CWT is computationally more expensive than discrete wavelet transform (DWT). DWT is any kind of wavelet transform using a discrete set of wavelets. In DWT, mother wavelet is discretized at powers of two.

$$\psi_{(a,b)}(t) = \frac{1}{\sqrt{2^j}} \psi \left(\frac{t - k \times 2^j}{2^j} \right) \tag{8}$$

where j and k are scale and translation parameters, respectively.

In DWT, a signal can be decomposed and reconstructed using filters. A signal passes through low-pass and high-pass filters, resulting in an approximation series and a detail series. The approximation series represents the low-frequency part of the signal and the detail series describes the high-frequency part of the signal. The approximation can be decomposed iteratively into approximation and detail series when passing low-pass and high-pass filters. Finally, one approximation series and a number of details are generated. The original signal can be reconstructed using these approximations and details.

The Daubechies wavelet is a kind of commonly used wavelet transform, which exhibits good trade-off between parsimony and information richness and is suitable for precipitation series modeling (Abbaszadeh 2016; Ramana et al. 2013). In this study, the Daubechies wavelet (Daubechies 1992) order 5 (db5) is used to decompose the precipitation series into different levels. The decomposed approximations and details at antecedent months are regarded as predictors to forecast the precipitation in the target month. The decomposition may help identify the underlying precipitation patterns or cycles at different resolutions. Therefore, precipitation series is preprocessed to assist seasonal forecasts versus those unprocessed.

3.2 Numerical weather forecasting

The NMME (Kirtman et al. 2014) is a global seasonal climate forecasting system developed by a collaboration of several research centers in North America. The NMME can provide real-time operational forecasts of a number of climatic variables at a 1° × 1° resolution for a few months in advance. The NMME is used to compare with statistical ensembles. Eight NMME models are collected to be numerical forecasting ensembles (Table 2). Each NMME model has different ensemble members and the ensemble mean of those members is regarded as a single forecast for a specific model. Probabilistic prediction is generated by sampling from the eight NMME ensembles through BMA. Specifically, a total of 100 samples are drawn from the BMA predictive probability distribution function (PDF). Details are introduced in the following section. Here we focus more on model uncertainty and less on the uncertainty of initial conditions of each model to prevent a deviation towards a specific model that has more ensemble members than the others. Deterministic forecasts in numerical models are obtained through a weighting scheme described in Sect. 3.3. The lead times in the NMME are not integers and the 0.5-month lead is regarded as 1-month lead in order to compare with statistical models. The comparison is examined for 1–6-month lead time in this study. The eight NMME models are bias corrected using quantile mapping based on empirical probability distribution. Quantile mapping is a commonly used method to correct the systematic bias in climate model outputs relative to observations (Cannon et al. 2015).

$$y_{m,p}(t) = F_{o,h}^{-1}\left\{F_{m,h}\left[y_{m,p}(\mathrm{t})\right]\right\} \tag{9}$$

where $F_{o,h}$ and $F_{m,h}$ are the cumulative distribution functions (CDFs) of observed data (denoted by subscript o) and modeled data (denoted by subscript m), respectively. Equation (9) denotes the correction of the bias in modeled value at projected time (denoted by subscript p) t using observations at historical period (denoted by subscript h). F^{-1} represents the inverse of CDFs, i.e. quantile function. Here, the empirical CDF is used to map the distribution of numerical forecasts to observations, which is regarded as quantile–quantile transformation.

3.3 Model weighting

Here, BMA is used to weight statistical and numerical ensemble members individually at each grid cell. BMA (Hoeting et al. 1999) is a statistical method to consider model uncertainty by posterior probability inference. Suppose y is variable of interest, the posterior distribution of y given data D is

$$\mathrm{P}(y|D) = \sum_{k=1}^{n} P\left(y|M_k, D\right) P\left(M_k|D\right) \tag{10}$$

where $M=\{\{M_1, M_2, \ldots, M_n\}\}$ are model sets and $P(M_k|D)$ is the posterior probability of model M_k, i.e. the weight in BMA. The posterior mean and variance can be expressed as

$$\mathrm{E}\left[\mathrm{y}|\mathrm{D}\right] = \sum_{k=1}^{n} P\left(M_k|D\right) \cdot E\left[P\left(y|M_k, D\right)\right] = w_k M_k \tag{11}$$

$$\mathrm{Var}\left[\mathrm{y}|\mathrm{D}\right] = \sum_{k=1}^{n} w_k (M_k - \sum_{i=1}^{n} w_i M_i)^2 + \sum_{k=1}^{n} w_k \sigma_k^2 \tag{12}$$

Table 2 The eight NMME models in this study

Model	Organization	Hindcast period	Ensemble size	Lead times (months)
CMC1-CanCM3	CMC	1981–2010	10	0.5–11.5
CMC2-CanCM4	CMC	1981–2010	10	0.5–11.5
COLA-RSMAS-CCSM4	NCAR	1982–2010	10	0.5–11.5
GFDL-CM2p1-AER04	GFDL	1982–2010	10	0.5–11.5
GFDL-CM2p5-FLOR-A06	GFDL	1982–2010	12	0.5–11.5
GFDL-CM2p5-FLOR-B01	GFDL	1981–2010	12	0.5–11.5
NASA-GMAO-062012	NASA	1981–2010	11	0.5–8.5
NCEP-CFSV2	NCEP	1982–2010	24/28	0.5–9.5

CMC Canadian Meteorological Center, *NCAR* National Center for Atmospheric Research, *GFDL* geophysical fluid dynamics laboratory, *NASA* national aeronautics and space administration, *NCEP* national centers for environmental prediction

where σ_k^2 is the variance in model M_k.

The posterior probability of a specific model can be obtained by maximizing the likelihood that the model is correct. To maximize the likelihood, log-likelihood function is commonly used.

$$\updownarrow(\theta) = \log\left(\sum_{k=1}^{n} w_k \cdot P(y|M_k, D)\right) \tag{13}$$

where $\theta = \{\{w_k, \sigma_k, k=1,2,\ldots,n\}\}$ is the parameters.

The expectation maximization (EM) algorithm (Moon 1996) can be used to solve the maximum likelihood problem. Let $z_{k,t}$ be a latent variable to record the state of kth model at time t. $z_{k,t}$ equals to 1 if kth model is the best at time t otherwise $z_{k,t}$ equals to 0. The EM algorithm includes two steps: E (expectation) step and M (maximization) step. At an initial state $\theta^{(0)}$, $z_{k,t}$ is calculated at E step.

$$\hat{z}_{k,t}^{j} = \frac{g\left(y_t|M_{k,t}, \sigma_k^{j-1}\right)}{\sum_{k=1}^{n} g\left(y_t|M_{k,t}, \sigma_k^{j-1}\right)} \tag{14}$$

where $g(\cdot)$ refers to gamma distribution (Sloughter et al. 2007) and j is jth iteration.

In the M step, θ is calculated based on $z_{k,t}$. This procedure is run repeatedly when a specific convergence threshold is satisfied.

$$w_k^j = \frac{1}{T}\sum_{t=1}^{T} z_{k,t}^j \tag{15}$$

$$\sigma_k^{2^j} = \frac{\sum_{t=1}^{T} \hat{z}_{k,t}^j \cdot \left(y_t - M_{k,t}\right)^2}{\sum_{t=1}^{T} \hat{z}_{k,t}^j} \tag{16}$$

The probabilistic forecasts are generated from the BMA predictive PDF $g_k(y_t|M_{k,t})$ according to posterior probability (Raftery et al. 2005). Specifically, an integer value k is generated from the numbers $\{\{1,\ldots,n\}\}$ based on probability $\{\{w_1,\ldots,w_n\}\}$. Then a forecasted value y_t is produced from PDF $g_k(y_t|M_{k,t})$. This process is repeated 100 times to obtain a forecasted ensemble, which is regarded as the probabilistic forecasts.

3.4 Evaluation metrics

Three mathematical metrics are used to measure the forecasting accuracy, i.e. Pearson's correlation coefficient (PCC), root mean square error (RMSE) skill score (RMSESS) and continuous ranked probability skill score (CRPSS) (Hersbach 2000). PCC and RMSESS are used to measure deterministic forecasts and CRPSS is utilized to measure probabilistic forecasts. PCC examines the linear correlation between two variables x and y. A PCC value of 1 indicates total correlation and a value of 0 means no correlation.

$$PCC_{x,y} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{y}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\sqrt{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}} \tag{17}$$

where x_i and y_i are data samples of x and y, respectively; $\bar{x}$ and $\bar{y}$ are arithmetic average of x and y, respectively; n denotes the sample size.

RMSESS (Jolliffe and Stephenson 2012) is an evaluation metric that measures the skill versus the reference forecasts (e.g. climatology) based on RMSE. RMSESS has a maximum value of 1, indicating perfect forecasts. A value of 0 in RMSESS means that the forecasting skill is equal to the climatological forecasts.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - y_i\right)^2}{n}} \tag{18}$$

$$\mathrm{RMSESS} = 1 - \frac{RMSE}{RMSE_{clim}} \tag{19}$$

where $\mathrm{RMSE_{clim}}$ refers to the RMSE obtained from a climatological forecast. The climatological forecast refers to the precipitation forecasts based on the multi-year averaged monthly precipitation in the training time period. For example, the climatological precipitation forecasts for January 2011 is obtained by averaging all the precipitation values in January from 1960 to 2010.

The continuous ranked probability score (CRPS) measures the difference between predicted and observed cumulative probability distributions. The associated skill score of CRPS, i.e. CRPSS, compares the forecasts to a climatological reference. A CRPSS close to 1 indicates good match between forecasts and observations. A value of CRPSS less than 0 suggests lower accuracy than climatology. CRPSS can be generated from continuous ranked probability score (CRPS).

$$\mathrm{CRPS} = \int_{-\infty}^{\infty}\left[P_f(x) - P_o(x)\right]^2 dx \tag{20}$$

$$\mathrm{CRPSS} = 1 - \frac{CRPS}{CRPS_{clim}} \tag{21}$$

where P_f and P_o are CDFs of forecasts and observations, respectively. $CRPS_{clim}$ refers to the reference CRPS obtained from a climatological forecast of predictand.

$$P_f(\mathrm{x}) = \int_{-\infty}^{x} \rho(y)dy \tag{22}$$

$$P_o(x) = H(x - x_o) \tag{23}$$

$$\mathrm{H(x)} = \begin{cases} 0 & for\ \ x < 0 \\ 1 & for\ \ x \geq 0 \end{cases} \tag{24}$$

where $\rho(y)$ is the PDF of forecasts. $H(\cdot)$ is the Heaviside function.

4 Results

4.1 Overall performance

The performance of statistical and numerical ensembles for deterministic and probabilistic precipitation forecasts are listed in Figs. 3 and 4 in terms of correlation and RMSESS, respectively. These scores are computed over space averaged in time. The PCC, RMSESS and CRPSS results are listed in tabular format (Tables 3 and 4). For deterministic forecasts, the correlations in statistical models seem much higher than numerical ensembles (Fig. 3 and Table 3) both at 1 and 3 month leads. Statistical models have an out-of-sample

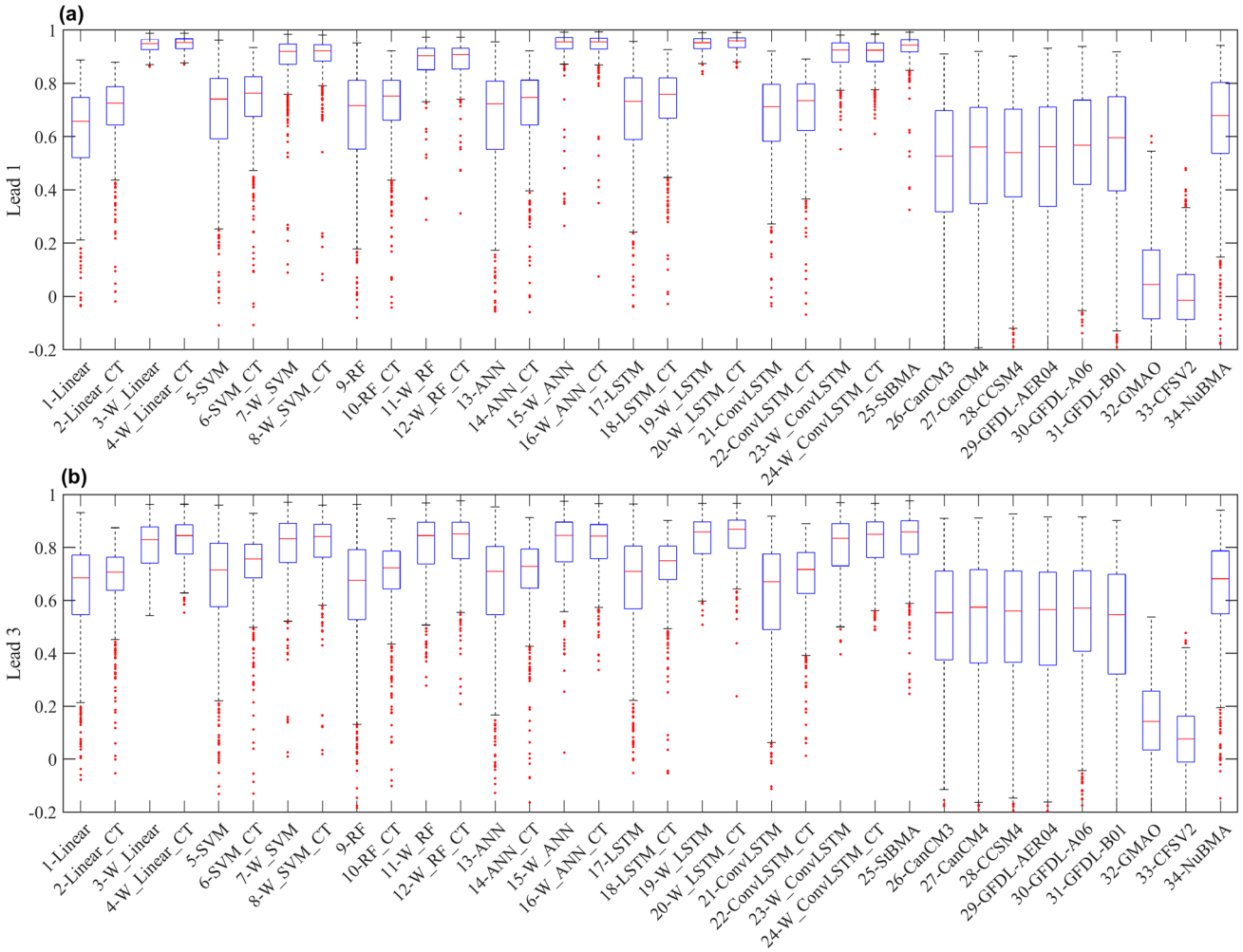

Fig. 3 The correlation between precipitation forecasts and observations in statistical and numerical ensembles at 1 (**a**) and 3 (**b**) month leads, respectively. The training data spans from 1960 to 2010 in statistical ensembles and from 1982 to 2010 in numerical ensembles, and the validating data ranges from 2011 to 2016. The boxplots represent the empirical quantile distribution of the correlation for each specific model. The horizontal line inside each box denotes the median, and the upper and lower margins of the box represent the upper quartile (Q_3) and lower quartile (Q_1), respectively. The upper whisker indicates the upper limit ($Q_3 + 1.5(Q_3 - Q_1)$) and the lower whisker indicates the lower limit ($Q_1 - 1.5(Q_3 - Q_1)$). The red dots outside the whiskers represent outliers. The boxplots are truncated to better show their differences

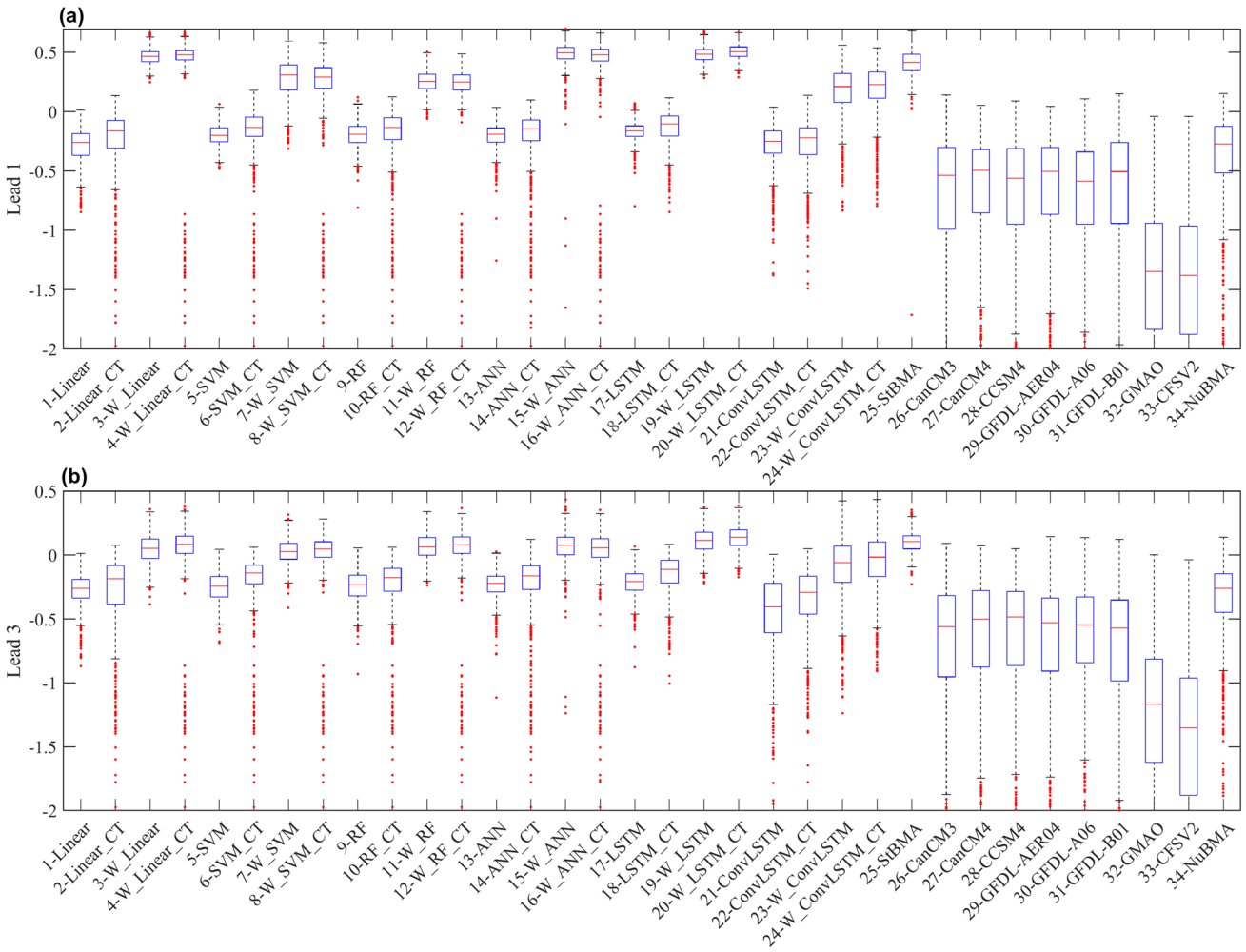

Fig. 4 The RMSESS between precipitation forecasts and observations in statistical and numerical ensembles at 1 (**a**) and 3 (**b**) month leads, respectively. The training and validating data are set the same as that in Fig. 3. The boxplots represent the empirical quantile distribution of the RMSESS for each specific model. The lines, whiskers and dots of the boxplots have the same meaning as that in Fig. 3

correlation coefficient over 0.9 at 1-month lead in models with wavelet preprocessing (Fig. 3a and Table 3). The correlation reduces to about 0.7 in models without wavelet processing, which is still better than that of numerical models with a correlation below 0.7. The forecasting performance decreases when the forecasting lead increases to 3 month (Fig. 3b and Table 3), with a PCC of about 0.85 (0.7) in the models with (without) wavelet preprocessing, which is still higher than that of the NMME ensembles with a PCC generally below 0.7. Model weighting through BMA produces a combination of statistical ensembles that has a performance level similar with that of wavelet-preprocessed models, resulting from larger weights in models with wavelet processing and smaller weights in those without wavelet. The weighted combination in numerical ensembles exhibits a better correlation than individual NMME ensemble members (Fig. 3 and Table 3), despite the low correlation with observations in GMAO and CFSV2 models.

For statistical models without wavelet processing, incorporating temperature and climate indices improves the performance, while this improvement is not that much in models with wavelet processing (Fig. 3 and Table 3). A two-sample *t*-test is performed on the precipitation forecasts of experimental pairs, i.e. (type I, type II) and (type III, type IV), to see whether they are statistically different in the population mean. The results are listed in the supplementary material (Table S1). The type I and type II models are statistically different from each other for nearly all the models at all the lead times. The type III and type IV pairs are statistically different from each other for the majority of the models except the wavelet LSTM and wavelet ConvLSTM pairs at some lead times. The statistical difference between type I (III) and type II (IV) models suggests that the two models are intrinsically different. Different predictors in these models lead to diverse distribution of the forecasts, indicating the importance of

Table 3 A summary of the median metrics for 1, 3 and 6-month lead precipitation forecasts

Model	Lead 1		Lead 3		Lead 6	
	PCC	RMSESS	PCC	RMSESS	PCC	RMSESS
Linear	0.66	− 0.26	0.69	− 0.26	0.66	− 0.26
Linear_CT	0.73	− 0.16	0.71	− 0.19	0.78	− 0.09
W_Linear	0.95	0.47	0.83	0.05	0.79	− 0.04
W_Linear_CT	0.95	0.48	0.85	0.08	**0.82**	0.00
SVM	0.74	− 0.20	0.72	− 0.24	0.71	− 0.21
SVM_CT	0.76	− 0.13	0.76	− 0.14	0.78	− 0.12
W_SVM	0.92	0.31	0.83	0.03	0.79	− 0.08
W_SVM_CT	0.92	0.29	0.84	0.05	0.80	− 0.07
RF	0.72	− 0.19	0.68	− 0.23	0.68	− 0.21
RF_CT	0.75	− 0.13	0.72	− 0.18	0.79	− 0.06
W_RF	0.90	0.25	0.85	0.06	0.80	− 0.04
W_RF_CT	0.91	0.25	0.85	0.08	**0.82**	− 0.01
ANN	0.72	− 0.19	0.71	− 0.22	0.69	− 0.21
ANN_CT	0.75	− 0.15	0.73	− 0.16	0.78	− 0.10
W_ANN	**0.96**	**0.50**	0.85	0.08	0.80	− 0.04
W_ANN_CT	**0.96**	0.48	0.84	0.06	0.81	− 0.04
LSTM	0.73	− 0.16	0.71	− 0.21	0.69	− 0.20
LSTM_CT	0.76	− 0.10	0.75	− 0.11	0.78	− 0.06
W_LSTM	0.95	0.48	0.86	0.11	0.81	− 0.02
W_LSTM_CT	**0.96**	**0.50**	**0.87**	**0.14**	**0.82**	0.01
ConvLSTM	0.71	− 0.25	0.67	− 0.41	0.67	− 0.50
ConvLSTM_CT	0.74	− 0.22	0.72	− 0.29	0.76	− 0.24
W_ConvLSTM	0.93	0.21	0.84	− 0.06	0.80	− 0.13
W_ConvLSTM_CT	0.92	0.23	0.85	− 0.02	0.81	− 0.11
StBMA	0.94	0.41	0.86	0.10	**0.82**	**0.02**
CMC1-CanCM3	0.53	− 0.54	0.55	− 0.56	0.55	− 0.52
CMC2-CanCM4	0.56	− 0.50	0.57	− 0.50	0.55	− 0.51
COLA-RSMAS-CCSM4	0.54	− 0.56	0.56	− 0.48	0.57	− 0.49
GFDL-CM2p1-AER04	0.56	− 0.50	0.57	− 0.53	0.55	− 0.54
GFDL-CM2p5-FLOR-A06	0.57	− 0.58	0.57	− 0.55	0.57	− 0.56
GFDL-CM2p5-FLOR-B01	0.60	− 0.51	0.55	− 0.57	0.53	− 0.68
NASA-GMAO-062012	0.04	− 1.35	0.14	− 1.17	0.06	− 1.31
NCEP-CFSV2	− 0.01	− 1.38	0.08	− 1.35	0.03	−1.34
NuBMA	0.68	− 0.27	0.68	− 0.26	0.68	− 0.25

These metrics are computed over each grid cell averaged in time. The best value is in bold over each column

Table 4 A summary of the median CRPSS for 1, 3 and 6-month lead precipitation forecasts

CRPSS	Lead 1	Lead 3	Lead 6
Statistical	0.22	0.05	0.04
NMME	− 0.24	− 0.24	− 0.24

predictor selection. Although using different predictors may result in similar predictive performance, the forecasted values based on different predictors are actually distinguishable. Therefore, the use of diverse predictors in multiple models can be utilized to quantify the model uncertainty in terms of predictor selection.

The RMSESS shares some similarities with PCC but exhibits some differences. Wavelet-based models have a median RMSESS much lower than the models without wavelet and numerical ensembles at 1-month lead (Fig. 4a), similar with that in PCC (Fig. 3a). The models with the highest PCC may not have the highest RMSESS (Table 3), such as the W_ANN_CT model. The BMA weighting of numerical ensembles has similar statistical quantiles with models without wavelets in terms of RMSESS, which is better than that of individual NMME ensemble members. The RMSESS

value decreases with the increase of lead time in statistical ensembles, with a RMSESS of 0.10 for the weighted ensemble at 3-month lead (Table 3). A higher RMSESS value is seen in the statistical ensemble weighting than that of numerical weighting (− 0.26). It is noticed that the correlation and RMSESS change little in numerical ensembles as the lead increases (Figs. 3, 4 and Table 3). However, this does not suggest more stable forecasts in numerical models than statistical models because the RMSESS values in numerical models is generally below 0, indicating poor capability in improving climatological forecasts. The numerical forecasts may simulate the seasonal variations to some extent, leading to similar correlation coefficients with observations at different leads.

The CRPSS demonstrated in Table 4 is calculated as the statistical quantiles of the median of CRPSS for each grid during the validating period. Statistical ensembles have a higher CRPSS than NMME ensembles at 1, 3 and 6-month lead, with positive score value in the statistical ensembles but negative scores in NMME. The CRPSS generally decreases with increasing lead time in statistical ensembles, with a median CRPSS of 0.22 at 1 month ahead and 0.04 at 6-month lead. A positive CRPSS indicates an added value over the climatology forecasts, while a negative CRPSS suggests less capability versus the climatology.

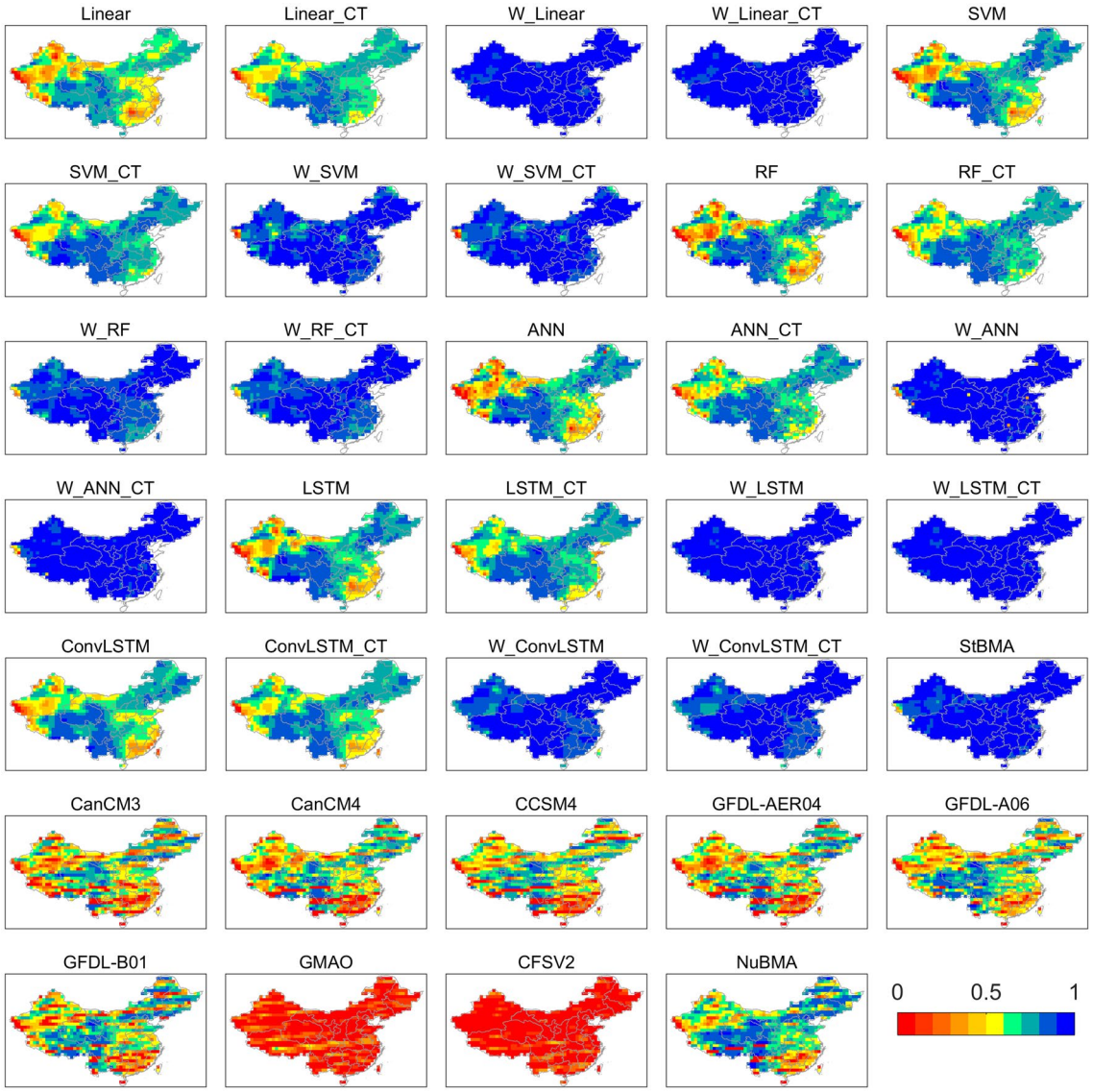

Fig. 5 Spatial patterns of PCC for precipitation forecasts at 1-month lead. The training and validating data are set the same as that in Fig. 3. The scale is truncated to highlight the difference of the correlations in respective models

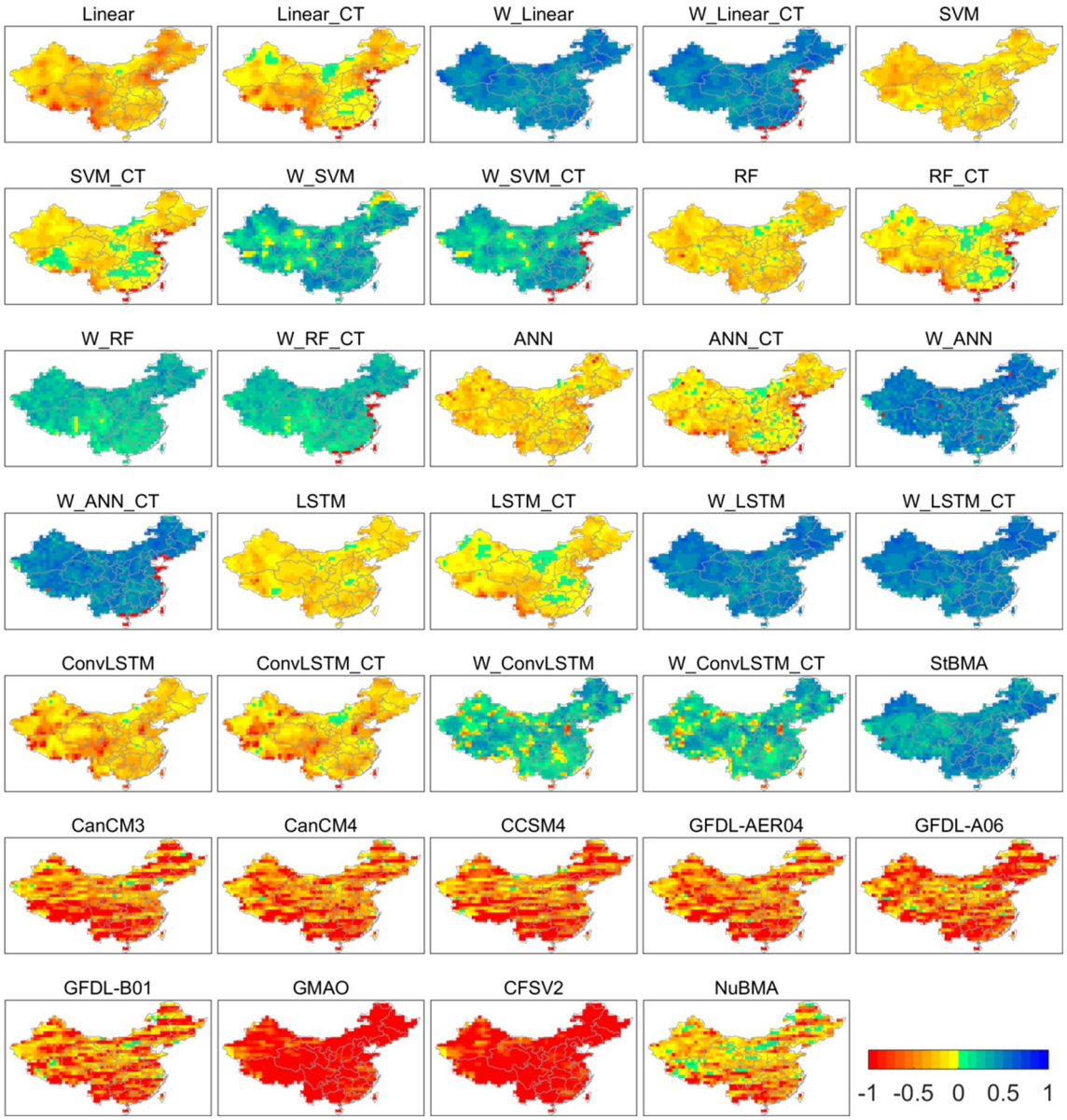

Fig. 6 Spatial patterns of RMSESS for precipitation forecasts at 1-month lead. The training and validating data are set the same as that in Fig. 3. The scale is truncated to highlight the difference of the RMSESS in respective models

4.2 Spatial patterns of forecasts

The spatial patterns of PCC and RMSESS for 1-month lead precipitation forecasts are demonstrated in Figs. 5 and 6, respectively. Wavelet-based models have high correlation with observations spatially (Fig. 5), such as W_Linear, W_ANN and W_LSTM, while the statistical models without incorporating wavelet have smaller PCC values. The NMME ensembles have low correlation coefficient compared to statistical ensembles, especially in GMAO and CFSV2 models. The BMA weighting of NMME members improves the correlation relative to single numerical models in some areas such as the northeastern China, central and western China.

As for the RMSESS, a similar conclusion can be drawn as that in Fig. 5, with higher RMSESS in statistical models with wavelet preprocessing, followed by models without wavelet processing and numerical ensembles. The RMSESS score is generally positive in wavelet based statistical models and is largely negative in statistical models without wavelet and numerical models (Fig. 6). The BMA weighted result gives a RMSESS value above zero for the majority of China in statistical models. The weighting of

numerical members exhibits some improvements over individual numerical models.

4.3 Selection of ensemble members

Here, three small areas (Fig. S1) in China (115°–116° E, 29°–30° N; 115°–116° E, 41°–42° N; 87°–88° E, 37°–38° N) are selected to examine the performance of different ensemble member selection strategies. These areas are located in different climate regimes of China. They are selected instead of the whole China is mainly due to the long running time when fitting the BMA predictive PDF over a long period and large areas. The chosen areas are demonstrated as an illustration to examine the performance of selected ensemble members. The correlation coefficient, RMSESS and CRPSS measurements are calculated for precipitation forecasts based on the single best model and based on some selected ensemble members (2, 4, 6, 8, 12, 16 and 24). The selected ensemble members are chosen based on the descending order of the correlation in the validating period. The best single model is determined based on maximum RMSESS, i.e. W_ANN, W_LSTM_CT, W_ANN, W_RF_CT, W_RF_CT and W_LSTM_CT for 1–6-month lead, respectively in the validating period. Weighting of some ensemble members may not lead to better forecasting performance than the best single model (Table 5), as suggested by All_24. Beginning with the best single model, adding some good ensemble members generally increases the ensemble performance in terms of correlation and RMSESS, such as selected 2, 4 and 6 ensemble members at 1-month lead. This improvement is also seen in some selected ensemble members at longer lead. However, the improvement over the best single model is limited when incorporating some ensemble members with medium or poor forecasting performance. A selection of 12, 16 and 24 ensemble members do not increase ensemble performance compared with selected 6 ensembles at 1-month lead (Table 5). This phenomenon that incorporating more ensemble members may decrease the performance is also shown at longer lead.

For probabilistic measurement, there is an improvement of CRPSS when selecting 6 ensemble members versus 2 and 4 members at 1-month lead (Table 5). However, increasing ensemble members beyond 6 does not increase the CRPSS score in 1-month lead forecast. This is probably related to the deterioration of ensemble forecasting performance when adding some ensemble members with medium or poor performance to the selected 6 members. In longer lead beyond 1-month, probabilistic forecasts benefit from incorporating over 6 ensemble members versus deterministic forecasts, which is probably because of the complementarity of added members. The CRPSS score is close to zero at longer lead (e.g. 4 months), indicating little added skill over

Table 5 A summary of the average metrics for 1 and 6-month lead precipitation forecasts at selected three locations

	Best single	Selected_2	Selected_4	Selected_6	Selected_8	Selected_12	Selected_16	All_24
PCC								
Lead 1	0.94	**0.95**	**0.95**	**0.95**	0.92	0.91	0.91	0.91
Lead 2	0.85	0.87	0.87	**0.88**	0.84	0.83	0.83	0.83
Lead 3	0.79	0.81	0.82	**0.83**	0.82	0.80	0.80	0.80
Lead 4	0.80	0.77	0.77	0.79	**0.81**	0.78	0.77	0.77
Lead 5	0.81	0.79	0.80	**0.82**	0.81	0.80	0.80	0.80
Lead 6	0.77	**0.79**	**0.79**	**0.79**	0.78	0.76	0.76	0.76
RMSESS								
Lead 1	0.48	0.53	0.53	**0.54**	0.37	0.37	0.37	0.37
Lead 2	0.20	0.24	0.25	**0.26**	0.12	0.14	0.14	0.14
Lead 3	0.08	0.11	0.13	**0.15**	0.03	0.08	0.08	0.08
Lead 4	0.08	0.03	0.04	**0.07**	− 0.01	0.05	0.04	0.04
Lead 5	0.09	0.03	0.05	**0.09**	− 0.02	0.07	0.07	0.07
Lead 6	0.02	0.04	0.04	**0.05**	− 0.06	0.02	0.02	0.02
CRPSS								
Lead 1		0.28	0.28	**0.29**	0.25	0.23	0.23	0.23
Lead 2		0.11	0.11	**0.12**	**0.12**	0.11	0.10	0.10
Lead 3		0.03	0.04	0.05	**0.07**	**0.07**	0.06	0.06
Lead 4		0.00	0.00	0.01	**0.04**	**0.04**	0.03	0.03
Lead 5		− 0.02	− 0.01	0.02	0.03	**0.05**	**0.05**	**0.05**
Lead 6		0.00	− 0.01	0.01	0.00	**0.02**	**0.02**	**0.02**

These metrics are computed over each point averaged in time. The best value is shown in bold over each row

the climatology. The incorporation of many more ensemble members may slightly balance the poor performance of less ensemble members.

4.4 Comparison of BMA and RMSE weighting

The inverse of forecasting error in the training period is commonly used as weights to combine multiple models apart from BMA. A comparison of the BMA weights and the weights obtained from RMSE may be useful to demonstrate their differences. The RMSE weights are determined by the inverse of error variance (squared RMSE) during 1960–2010. A summary of the quantiles of weights of different statistical models obtained from BMA and RMSE for 1-month lead precipitation forecasts is listed in Fig. 7. The RMSE weights seem to be more concentrative than that of BMA because the forecasting error variances in individual statistical models are not very large. However, the BMA determines the weight from the probability that a model is the best for a specific time period. The BMA method tends to assign larger weights to good models and lower weights to bad models, in contrast to the way of absolute error assessment in RMSE. Overall, insufficient models are

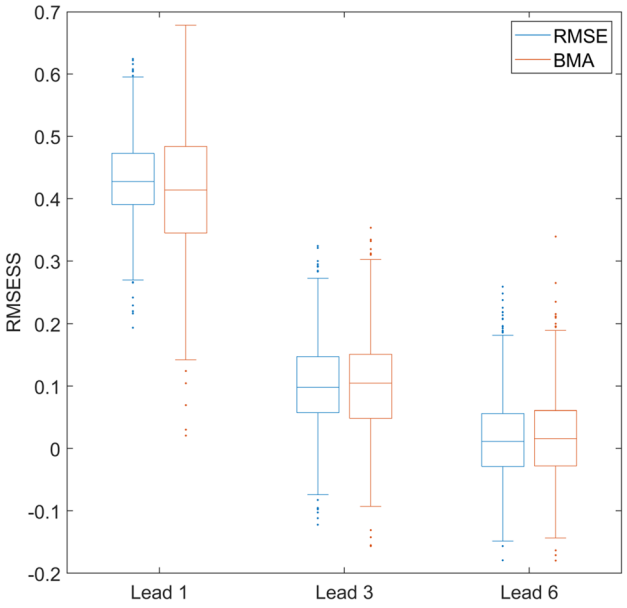

Fig. 8 A comparison of the RMSESS between RMSE weighting and BMA weighting methods for 1, 3 and 6-month lead precipitation forecasts

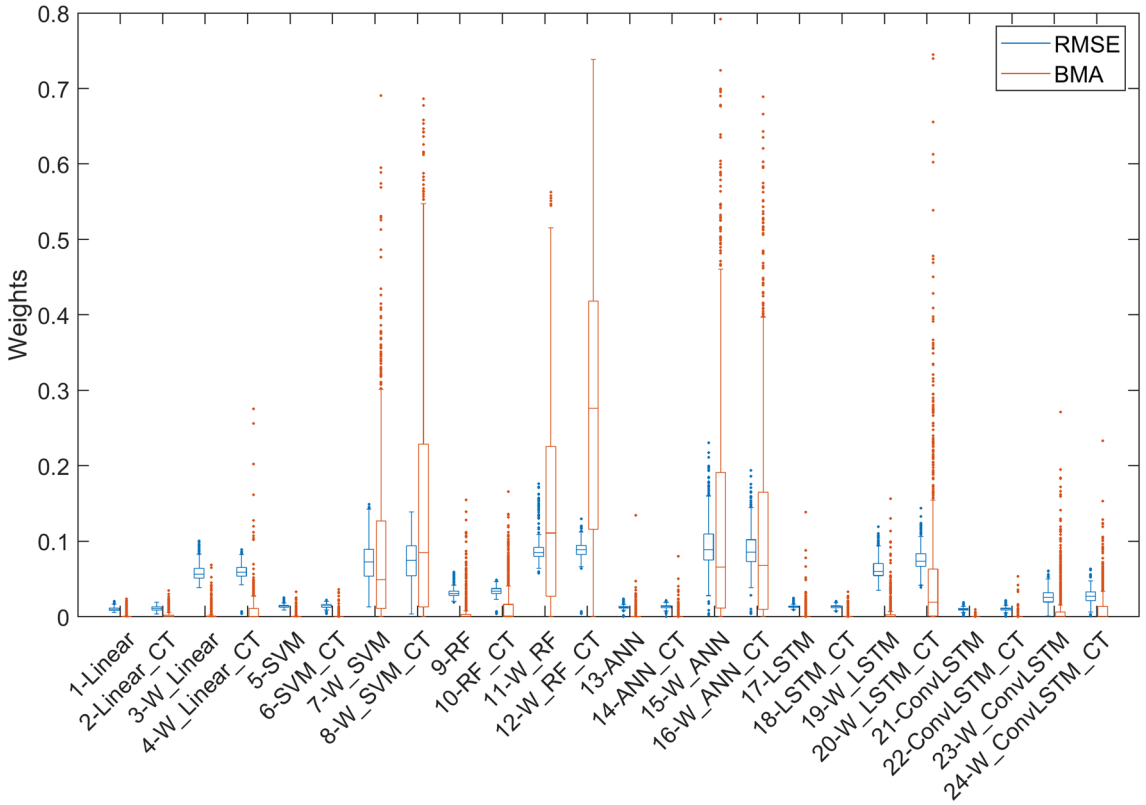

Fig. 7 The weights of 1-month lead precipitation forecasts obtained from RMSE and BMA using a series of statistical models. The training data are set the same as that in Fig. 3. The boxplots represent the empirical quantile distribution of the weights for each specific model. The lines, whiskers and dots of the boxplots have the same meaning as that in Fig. 3

given lower median weights in BMA weighting than that of RMSE weighting.

The weighted deterministic precipitation forecasts exhibit little difference between RMSE weighting and BMA weighting methods at 1, 3 and 6-month leads (Fig. 8). The BMA weighted performance shows a wider range of RMSESS values than that of the RMSE weighting at 1-month lead precipitation forecasting. When the forecasting lead increases to 3–6 month, the BMA weighted performance exhibits slightly higher median RMSESS value than the RMSE weighting. In addition, BMA can be used to generate probabilistic precipitation forecasts, while the RMSE weighting approach is not able to model the probability of an ensemble of members.

5 Discussion

Data-driven multi-model ensembles exhibit much better accuracy than numerical NMME, both in deterministic and probabilistic precipitation forecasts. This is related to the flexible adaptability of statistical modeling based on data directly in statistical models versus complex simulations in physical models. Large uncertainties inevitably exist in current numerical simulations, including data, initial conditions and model structures. Although these uncertainties are partly quantified and considered in numerical models or some analysis studies (Anthes et al. 1989; Bauer et al. 2015), they can be further reduced by the improvement of data quality, models and algorithms. Large gaps exist between statistical and numerical models, especially at 1-month lead. The NMME models have not reached a level of climatology forecasts, while statistical models surpass the climatology in precipitation forecasting for a few months ahead. However, numerical models have advantages in interpretation and attribution over statistical models. Therefore, statistical models can serve as an enhancement of precipitation forecasting to improve simulation accuracy of numerical models. For example, statistical models can be incorporated into numerical ensembles to construct hybrid combinations.

Wavelet preprocessing plays a significant role in enhancing statistical models. The accuracy of statistical models exhibits a large improvement relative to unprocessed ones. Wavelet decomposes precipitation into different levels, including low-frequency and high-frequency components. Low-frequency components can be regarded as the long-term changes, while high-frequency components are considered as transient oscillations. Suppose precipitation is modulated by numerous periodic and stochastic factors (Kai et al. 1983), they may influence precipitation intensity and timing at different temporal scales. The decomposed low-frequency and high-frequency components may correspond to different influences analogously. In this case, precipitation forecasting using multiple influencing factors seems more reasonable than using a hybrid of them, i.e. the raw precipitation time series. A mixture of these components may have difficulty finding intrinsic patterns influencing precipitation because they are blended. Separating individual patterns seems to be more promising in forecasting precipitation than using a hybrid pattern.

Selecting good ensemble members can increase the ensemble accuracy in some degree (Raftery et al. 2005). However, the number of good ensemble members influences the ensemble performance. Starting with the best single model, adding some good ensemble members can improve the forecasting performance. However, the performance may begin to decrease when excess ensemble members are incorporated. The suitable number of ensemble members to produce better or even the best performance depends on the data and performance of individual models. It may be a good choice to select good ensemble members based on a specific rule, while the best number of ensemble members and which models need to be selected among all the model combinations remain to be investigated. Bad ensemble members can reduce ensemble performance, although BMA considers model uncertainty by posterior probability and assigns smaller weights to them.

Multi-model ensembles are often used to quantify the uncertainty of simulations of a specific variable, which is widely adopted in current hydrometeorological applications (Knutti and Sedláček 2013; Tebaldi and Knutti 2007). Bad ensemble members may inflate the uncertainty because they bring more forecasting uncertainty than good ensemble members. Selecting good ensemble members can improve probabilistic prediction towards observed distribution by discarding some bad ones. It needs to be cautious when quantifying uncertainty using multi-model ensembles because good ensembles are not constant. Another concern in a recent study mentions that multi-model ensemble does not have systematic relationships with uncertainty in asymptotic and consistent properties but it can be used in sensitivity analysis (Nearing and Gupta 2018).

6 Conclusion

In this study, a statistical multi-model ensemble is constructed to forecast precipitation for 1 to 6 months in advance. The performance of the statistical ensemble seems much better than an advanced numerical ensemble (i.e. NMME) in terms of correlation, RMSESS and CRPSS, especially at 1-month lead. The statistical ensemble exhibits a decreasing trend of forecasting accuracy with the increase of lead time. Despite the decreasing performance at longer lead, it is still better than that of numerical models in NMME. Given current gap of the forecasting performance between statistical and numerical

models, it seems promising to consider data-driven ensembles as a strong complement to numerical models. Moreover, it is likely that statistical and numerical ensembles may be combined to generate a blended ensemble in order to improve seasonal precipitation forecasts.

It seems that linear and nonlinear models have roughly the same performance at 1 and 3-month lead times. We plotted the PCC and RMSESS for 6-month lead precipitation forecasts in the supporting information (Fig. S2 and S3) and found that the relationship that linear and nonlinear models have roughly the same performance still holds. The best model is nonlinear model. However, nonlinear models are not significantly better than linear model. Precipitation forecasts are nonlinear in nature because numerous factors influence the precipitation. Wavelet based models are significantly better than the models without wavelet preprocessing. The advantage of the nonlinear transformation by wavelet can provide some evidence that nonlinear effects exist in precipitation forecasts. In this aspect, the linear and nonlinear models after wavelet processing can all be considered nonlinear. However, if we only look at the difference between linear and nonlinear models with or without wavelet processing, the discrepancy is not that much. This phenomenon may be due to the variable selection, data generation process and model itself. In fact, nonlinear model does not seem to be significantly superior over linear models in previous studies (Choubin et al. 2016). It can be seen from Figs. 3 and 4 that the results of Type II experiments are generally better than Type I experiments. Therefore, the predictor selection may be more important than model selection. The little improvement in nonlinear models over linear models is probably a combined result of incomplete selection of predictors, model inefficiency and the data generation process (e.g. interpolation) in GPCC precipitation.

Selecting good ensemble members or discarding some bad ones can increase the ensemble performance, especially for an ensemble that has some worse models than others. However, selecting too many models may lead to the degradation of ensemble performance in probabilistic forecasts, especially at longer lead. An empirical rule is that models with very limited performance should be discarded while retaining the diversity and equitability among an ensemble. Further work should focus on the rules to select optimal models using statistical approaches such as principal component analysis and maximum entropy.

Acknowledgements This work was supported by grants from the National Key Research and Development Projects (2018YFB2100500), National Natural Science Foundation of China program (no. 41890822), the National Nature Science Foundation of China program (41801339, 41971351, 41771422, 41601406) and the Nature Science Foundation of Hubei Province (2017CFB616).

References

Abbaszadeh P (2016) Improving hydrological process modeling using optimized threshold-based wavelet de-noising technique. Water Resour Manage 30:1701–1721

Anthes RA, Kuo YH, Hsie EY, Low-Nam S, Bettge TW (1989) Estimation of skill and uncertainty in regional numerical models. Q J R Meteorol Soc 115:763–806

Bauer P, Thorpe A, Brunet G (2015) The quiet revolution of numerical weather prediction. Nature 525:47

Becker A, Finger P, Meyer-Christoffer A, Rudolf B, Ziese M (2011) GPCC full data reanalysis Version 6.0 at 1.0: monthly land-surface precipitation from rain-gauges built on GTS-based and historic Data. Global Precipitation Climatology Centre (GPCC): Berlin, Germany

Becker E, den D Hv, Zhang Q (2014) Predictability and forecast skill in NMME. J Clim 27:5891–5906

Berkhahn S, Fuchs L, Neuweiler I (2019) An ensemble neural network model for real-time prediction of urban floods. J Hydrol 575:743–754

Bosilovich MG, Robertson FR, Chen J (2011) Global energy and water budgets in MERRA. J Clim 24:5721–5739

Breiman L (2001) Random forests. Mach Learn 45:5–32

Cannon AJ, Sobie SR, Murdock TQ (2015) Bias correction of GCM precipitation by quantile mapping: how well do methods preserve changes in quantiles and extremes? J Clim 28:6938–6959

Chan JC, Zhou W (2005) PDO, ENSO and the early summer monsoon rainfall over south China. Geophys Res Lett 32:1

Choubin B, Khalighi-Sigaroodi S, Malekian A, Kişi Ö (2016) Multiple linear regression, multi-layer perceptron network and adaptive neuro-fuzzy inference system for forecasting precipitation based on large-scale climate signals. Hydrol Sci J 61:1001–1009

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297

Cuo L, Pagano TC, Wang QJ (2011) A review of quantitative precipitation forecasts and their use in short- to medium-range streamflow forecasting. J Hydrometeorol 12:713–728. https://doi.org/10.1175/2011JHM1347.1

Darji MP, Dabhi VK, Prajapati HB Rainfall forecasting using neural network: A survey. In: 2015 international conference on advances in computer engineering and applications (2015) IEEE, pp 706–713

Daubechies I (1992) Ten lectures on wavelets

Drucker H, Burges CJ, Kaufman L, Smola AJ, Vapnik V Support vector regression machines. In: Advances in neural information processing systems, 1997. pp 155–161

Gers FA, Schmidhuber J, Cummins F (1999) Learning to forget: continual prediction with LSTM

Ham Y-G, Kim J-H, Luo J-J (2019) Deep learning for multi-year ENSO forecasts. Nature 573:568–572

Hao Z, Singh VP, Xia Y (2018) Seasonal drought prediction: advances, challenges, and future prospects. Rev Geophys 56:108–141

Harris I, Jones PD, Osborn TJ, Lister DH (2014) Updated high-resolution grids of monthly climatic observations—the CRU TS3.10 Dataset. Int J Climatol 34:623–642. https://doi.org/10.1002/joc.3711

Haykin S (1994) Neural networks: a comprehensive foundation. Prentice Hall PTR, NJ

Heil CE, Walnut DF (1989) Continuous and discrete wavelet transforms. SIAM Rev 31:628–666

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

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9:1735–1780

Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999) Bayesian model averaging: a tutorial. Stat Sci 1:382–401
Ingram K, Roncoli M, Kirshen P (2002) Opportunities and constraints for farmers of West Africa to use seasonal precipitation forecasts with Burkina Faso as a case study. Agric Syst 74:331–349
Jolliffe IT, Stephenson DB (2012) Forecast verification: a practitioner's guide in atmospheric science. Wiley, UK
Kai S, Mueller SC, Ross J (1983) Periodic precipitation patterns in the presence of concentration gradients. 2. Spatial bifurcation of precipitation bands and stochastic pattern formation. J Phys Chem 87:806–813
Khajehei S, Ahmadalipour A, Moradkhani H (2018) An effective post-processing of the North American multi-model ensemble (NMME) precipitation forecasts over the continental US. Clim Dyn 51:457–472
Khajehei S, Moradkhani H (2017) Towards an improved ensemble precipitation forecast: a probabilistic post-processing approach. J Hydrol 546:476–489
Kim T-W, Valdés JB (2003) Nonlinear model for drought forecasting based on a conjunction of wavelet transforms and neural networks. J Hydrol Eng 8:319–328
Kirtman BP et al (2014) The North American multimodel ensemble: phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. B Am Meteorol Soc 95:585–601
Knutti R, Sedláček J (2013) Robustness and uncertainties in the new CMIP5 climate model projections. Nat Clim Chang 3:369
Kripalani RH, Kulkarni A (2001) Monsoon rainfall variations and teleconnections over South and East Asia. Int J Climatol 21:603–616
Krishnamurti T, Kumar V, Simon A, Bhardwaj A, Ghosh T, Ross R (2016) A review of multimodel superensemble forecasting for weather, seasonal climate, and hurricanes. Rev Geophys 54:336–377
Li Y, Liang Z, Hu Y, Li B, Xu B, Wang D (2019) A multi-model integration method for monthly streamflow prediction: modified stacking ensemble strategy. J Hydroinf
Lieting C (2001) The Role of the Anomalous Snow Cover over the Qinghai-Xizang Plateau and ENSO in the Great Floods of 1998 in the Changjiang River Valley. Chin J Atmos Sci 2
Lipper L et al (2014) Climate-smart agriculture for food security. Nat Clim Chang 4:1068
Maldonado T, Alfaro E, Fallas-López B, Alvarado L (2013) Seasonal prediction of extreme precipitation events and frequency of rainy days over Costa Rica, Central America, using Canonical Correlation Analysis. Adv Geosci 33:41–52
Mallat SG (1989) A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans Pattern Anal Mach Intell 1:674–693
McFarlane NA, Boer G, Blanchet J, Lazare M (1992) The Canadian Climate Centre second-generation general circulation model and its equilibrium climate. J Clim 5:1013–1044
Molteni F, Buizza R, Palmer TN, Petroliagis T (1996) The ECMWF ensemble prediction system: methodology and validation. Q J R Meteorol Soc 122:73–119
Moon TK (1996) The expectation–maximization algorithm. IEEE Signal Process Mag 13:47–60
Mortensen E et al (2018) Regression-based season-ahead drought prediction for southern Peru conditioned on large-scale climate variables. Hydrol Earth Syst Sc 22:287
Najafi MR, Moradkhani H, Piechota TC (2012) Ensemble streamflow prediction: climate signal weighting methods vs. climate forecast system reanalysis. J Hydrol 442:105–116
Nearing GS, Gupta HV (2018) Ensembles vs. information theory: supporting science under uncertainty. Front Earth Sci 1:1–8
Nourani V, Baghanam AH, Gokcekus H (2018) Data-driven ensemble model to statistically downscale rainfall using nonlinear predictor screening approach. J Hydrol 565:538–551
Partal T, Kişi Ö (2007) Wavelet and neuro-fuzzy conjunction model for precipitation forecasting. J Hydrol 342:199–212
Pokhrel S et al (2016) Seasonal prediction of Indian summer monsoon rainfall in NCEP CFSv2: forecast and predictability error. Clim Dyn 46:2305–2326
Quilty J, Adamowski J, Boucher MA (2019) A stochastic data-driven ensemble forecasting framework for water resources: a case study using ensemble members derived from a database of deterministic wavelet-based models. Water Resour Res 55:175–202
Raftery AE, Gneiting T, Balabdaoui F, Polakowski M (2005) Using Bayesian model averaging to calibrate forecast ensembles. Mon Weather Rev 133:1155–1174
Ramana RV, Krishna B, Kumar S, Pandey N (2013) Monthly rainfall prediction using wavelet neural network analysis. Water Resour Manage 27:3697–3711
Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N (2019) Deep learning and process understanding for data-driven Earth system science. Nature 566:195
Roeckner E et al (2003) The atmospheric general circulation model ECHAM 5. Model description, PART I
Rumelhart DE, Hinton GE, Williams RJ (1988) Learning representations by back-propagating errors. Cognit Model 5:1
Saha S et al (2014) The NCEP climate forecast system version 2. J Clim 27:2185–2208
Shi X, Chen Z, Wang H, Yeung D-Y, Wong W-K, Woo W-C, Convolutional LSTM network (2015): A machine learning approach for precipitation nowcasting. In: Advances in neural information processing systems, pp 802–810
Slater LJ, Villarini G, Bradley AA (2017) Weighting of NMME temperature and precipitation forecasts across Europe. J Hydrol 552:646–659
Sloughter JML, Raftery AE, Gneiting T, Fraley C (2007) Probabilistic quantitative precipitation forecasting using Bayesian model averaging. Mon Weather Rev 135:3209–3220
Smola AJ, Schölkopf B (2004) A tutorial on support vector regression. Stat Comput 14:199–222. doi:https://doi.org/10.1023/b:stco.0000035301.49549.88
Stensrud DJ (2009) Parameterization schemes: keys to understanding numerical weather prediction models. Cambridge University Press, Cambridge
Taylor KE, Stouffer RJ, Meehl GA (2012) An overview of CMIP5 and the experiment design. B Am Meteorol Soc 93:485–498
Tebaldi C, Knutti R (2007) The use of the multi-model ensemble in probabilistic climate projections. Philos Trans R Soc A Math Phys Eng Sci 365:2053–2075
Thober S, Kumar R, Sheffield J, Mai J, Schäfer D, Samaniego L (2015) Seasonal soil moisture drought prediction over Europe using the North American Multi-Model Ensemble (NMME). J Hydrometeorol 16:2329–2344
Tong S, Koller D (2001) Support vector machine active learning with applications to text classification. J Mach Learn Res 2:45–66
Trenberth KE, Smith L, Qian T, Dai A, Fasullo J (2007) Estimates of the global water budget and its annual cycle using observational and model data. J Hydrometeorol 8:758–769
Vapnik V (2013) The nature of statistical learning theory. Springer, Berlin
Wang L (2005) Support vector machines: theory and applications. Springer, Berlin
Warszawski L, Frieler K, Huber V, Piontek F, Serdeczny O, Schewe J (2014) The inter-sectoral impact model intercomparison project (ISI–MIP): project framework. Proc Natl Acad Sci 111:3228–3232
Xiao M, Zhang Q, Singh VP (2015) Influences of ENSO, NAO, IOD and PDO on seasonal precipitation regimes in the Yangtze River basin, China. Int J Climatol 35:3556–3567
Xu L, Chen N, Zhang X (2018a) A comparison of large-scale climate signals and the North American Multi-Model Ensemble (NMME)

for drought prediction in China. J Hydrol 557:378–390. doi:https://doi.org/10.1016/j.jhydrol.2017.12.044

Xu L, Chen N, Zhang X, Chen Z (2018) An evaluation of statistical, NMME and hybrid models for drought prediction in China. J Hydrol 566:235–249. https://doi.org/10.1016/j.jhydrol.2018.09.020

Xu L, Chen N, Zhang X, Chen Z, Hu C, Wang C (2019) Improving the North American multi-model ensemble (NMME) precipitation forecasts at local areas using wavelet and machine learning. Clim Dyn 1:1–15

Yang J, Gong D, Wang W, Hu M, Mao R (2012) Extreme drought event of 2009/2010 over southwestern China. Meteorol Atmos Phys 115:173–184

Zaherpour J et al (2019) Exploring the value of machine learning for weighted multi-model combination of an ensemble of global hydrological models. Environ Model Softw 114:112–128. https://doi.org/10.1016/j.envsoft.2019.01.003

Zong Y, Chen X (2000) The 1998 flood on the Yangtze, China. Nat Hazards 22:165–184