The Optimal Multimodel Ensemble of Bias-Corrected CMIP5 Climate Models over China

XIAOLI YANG,[a,b] XIAOHAN YU,[b] YUQIAN WANG,[c] XIAOGANG HE,[d] MING PAN,[e] MENGRU ZHANG,[b] YI LIU,[b] LILIANG REN,[a] AND JUSTIN SHEFFIELD[f]

[a] *State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Hohai University, Nanjing, China*
[b] *College of Hydrology and Water Resources, Hohai University, Nanjing, China*
[c] *Northwest Electric Power Design Institute Co., Ltd. of China Power Engineering Consulting Group, Xi'an, China*
[d] *Water in the West, Woods Institute for the Environment, Stanford University, Stanford, California*
[e] *Department of Civil and Environmental Engineering, Princeton University, Princeton, New Jersey*
[f] *Geography and Environment, University of Southampton, Southampton, United Kingdom*

(Manuscript received 26 June 2019, in final form 13 January 2020)

ABSTRACT

A multimodel ensemble of general circulation models (GCM) is a popular approach to assess hydrological impacts of climate change at local, regional, and global scales. The traditional multimodel ensemble approach has not considered different uncertainties across GCMs, which can be evaluated from the comparisons of simulations against observations. This study developed a comprehensive index to generate an optimal ensemble for two main climate fields (precipitation and temperature) for the studies of hydrological impacts of climate change over China. The index is established on the skill score of each bias-corrected model and different multimodel combinations using the outputs from phase 5 of the Coupled Model Intercomparison Project (CMIP5). Results show that the optimal ensemble of the nine selected models accurately captures the characteristics of spatial–temporal variabilities of precipitation and temperature over China. We discussed the uncertainty of subset ensembles of ranking models and optimal ensemble based on historical performance. We found that the optimal subset ensemble of nine models has relative smaller uncertainties compared with other subsets. Our proposed framework to postprocess the multimodel ensemble data has a wide range of applications for climate change assessment and impact studies.

1. Introduction

The Fifth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC AR5; IPCC 2014) shows that, in recent decades, climate change has had a significant impact on natural and human systems over all continents and oceans, with increases in frequency and intensity of precipitation-related extremes (e.g., droughts and floods) and/or extreme temperature events (e.g., heat waves) in many regions (Wang and Zhou 2005; Sen Roy and Balling 2004; Li et al. 2015). For example, the frequency and the intensity of extreme climate events (e.g., heat waves and extreme precipitation) have increased significantly in China (Zhou and Ren 2011; Zhai et al. 2005). Research on climate change is urgent to understand how these changes may evolve in the future, which leads to an increased demand to develop more reliable and accurate climate change projection datasets (Fan et al. 2013).

The World Climate Research Program (WCRP) organized the development of the Coupled Model Intercomparison Project (CMIP), now in its fifth phase, whose outputs are widely used for climate change assessments (Covey et al. 2003; Meehl et al. 2000, 2005; Sheffield et al. 2013a,b; Taylor et al. 2012). The CMIP5 improves over the previous CMIPs with increased number of models, enhanced spatial resolution, and a larger set of experiments (Moss et al. 2010; Taylor et al. 2012). However, the spatial resolution of GCMs is far from enough to assess climate change impact at local and site-specific scales (Piani et al. 2010; Chen et al. 2018; Guo et al. 2019). Recently, many postprocessing methods have been developed to improve the spatial resolution of the GCM output (e.g., Mearns et al. 2003;

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JHM-D-19-0141.s1.

Corresponding author: Liliang Ren, rll@hhu.edu.cn

DOI: 10.1175/JHM-D-19-0141.1

Wilby et al. 2004; Maraun et al. 2010; Chen et al. 2013a, 2018, 2019; He et al. 2016; Cannon et al. 2015). Bias-correction methods can effectively correct the climate model output data (Teutschbein and Seibert 2012, 2013), such as monthly mean correction (Fowler et al. 2007), delta change (Hay et al. 2000), quantile mapping (Wood et al. 2002), and two-dimensional bias correction (Piani and Haerter 2012). Among these methods, the quantile mapping technique (Wood et al. 2004; Cannon 2016) has shown good performance for postprocessing climate model data to undertake climate change impact assessments (e.g., Piani et al. 2010; Cavazos and Arriaga-Ramírez 2012; Chen et al. 2013b, 2019). The quantile mapping method can correct the mean and the full marginal distribution and thereby the frequency and intensity of the target variable (Crochemore et al. 2016; Rajczak et al. 2016). This study uses the equal distance cumulative distribution function (EDCDF) method developed by Li et al. (2010) to bias correct the CMIP5 outputs.

To better improve the reliability of future projections from GCMs' outputs, multimodel ensemble (MME) methods have been proposed that distill the uncertainty across models in simulating the climate (Tebaldi and Knutti 2007; Herger et al. 2018). Different methods for developing MMEs exist, including simple model averaging (SMA; Miao et al. 2013), Bayesian model averaging (BMA; Miao et al. 2013; Katz and Ehrendorfer 2006), weighted ensemble averaging (WEA; Nohara et al. 2006), and reliability ensemble averaging (REA; Giorgi and Mearns 2002; Weiland et al. 2012). These indicate that the MME is usually superior to any individual model (Mote et al. 2011; Pierce et al. 2009; Reichler and Kim 2008; Jiang et al. 2016) and can overcome the systematic bias of a single model (Dong et al. 2015; Kim et al. 2012; Toh et al. 2018; Zhou et al. 2014).

However, due to the poor performance of some models for aspects of regional climate, an ensemble of all models equally will reduce the performance of the MME. For example, Gong et al. (2014) analyzed the output from 18 CMIP5 models and found that only half of the models can reasonably characterize the circulation pattern of the East Asian winter monsoon. Herger et al. (2018) developed a subset of the CMIP5 archive that minimizes regional biases in present-day climatology based on RMSE over space. Zhang and Soden (2019) found that the selected top five models for each continental subregion, the intermodel spread is reduced significantly in all cases compared to the full model ensemble. Therefore, selecting the most appropriate model and/or model ensembles based on statistical attributes to evaluate various impacts are important for specific study purposes (Aloysius et al. 2016; Ahmadalipour et al. 2017).

As a result, some researchers have focused on how to select a portion of models with better performance to compute the ensemble. For instance, Aloysius et al. (2016) classified models according to the differences in horizontal resolution, performance measures and causal mechanisms, and recommended a subset of models based on these criteria in areas without high-quality climate observations. Based on the pattern correlation coefficient (PCC) and signal-to-noise ratio of annual average precipitation, Lee and Wang (2014) found that the ensemble of the best four selected CMIP5 models has better skills than the multimodel ensemble average of all models. Thober and Samaniego (2014) developed computationally efficient methods for selecting sub-ensembles of RCMs to represent extreme precipitation and temperature indices over Germany and recommended a method that removes the worst performing models. Herrera-Estrada and Sheffield (2017) followed a similar path in identifying performance-based sub-ensembles over the continental United States but focused on the impact on uncertainty in future projections and showed via bootstrap sampling that small ensembles of models underestimated the variance in projections when limited to less than about 13 models out of 24 models. These studies demonstrate the challenge in identifying the optimal set of models as the goals, methods, and focus variables used by each study are different.

There are many model selection methods, such as simple ranking based on different performance metrics (Ahmadalipour et al. 2017; Salman et al. 2018; Pour et al. 2018). However, most studies focus on one climate field (precipitation or temperature) or give different ensemble for each climate fields (e.g., Gu et al. 2015; Ahmadalipour et al. 2017; Herger et al. 2018; Zhang and Soden 2019), rather than same members of the optimal ensemble. Normally, we need consistent members of optimal ensemble for both precipitation and temperature to ensure the physical consistency (relationship between precipitation and temperature should be maintained) and to further study the climate changes on hydrological process, which are the most important inputs of hydrological models (Woldemeskel et al. 2012; Teutschbein and Seibert 2012).

In this study, we aim to develop a novel method to generate an optimal subset that is representative of the range of precipitation and temperature indicated by the full ensemble, which could be adopted by hydrological model for climate change effects assessment across China. This paper is organized as follows. Sections 2 and 3 describe the datasets used in the study as well as the statistical downscaling methods, the bias-correction method, the performance evaluation metrics, and the

subensemble selection methods. Section 4 presents the main results, including evaluation of the model performance, optimization of the model subensemble, and its verification. Sections 5 and 6 discuss and summarize the main results of the study.

2. Datasets

The observed precipitation and temperature from 756 meteorological stations in China from 1961 to 2005 are obtained from the China Meteorological Administration Data Sharing Service System (http://data.cma.cn/). The distribution of the meteorological stations and the digital elevation model (DEM) are shown in Fig. 1. The observed data are interpolated to 0.5° × 0.5° using the inverse distance weighted method (IDW; Ma et al. 2016). The interpolation of temperature data takes into account the elevation, with a temperature decrease of 0.65°C for every 100 m increase in elevation (Choi and An 2010).

In this study, China is divided into four climate regions based on the long-term (1961–2005) mean precipitation: arid (precipitation < 200 mm), semiarid (200 mm ≤ precipitation < 400 mm), semihumid (400 mm ≤ precipitation < 800 mm), and humid (precipitation ≥ 800 mm) (Fig. 1). These regions are used to summarize the results.

We select monthly precipitation and temperature from 25 GCMs with r1i1p1 historical realization, which are archived at the Program on Climate Model Diagnosis and Intercomparison (PCMID) website (https://pcmdi.llnl.gov/index.html; Table 1). Due to the different spatial resolutions of the models, all of the climatic fields (monthly precipitation and temperature) are interpolated to the same 0.5° × 0.5° grid as the observations through the bilinear interpolation method (Kirkland 2010).

3. Methodology

a. Bias-correction method

The EDCDF method, developed by Li et al. (2010), is used to bias correct the output from the 25 climate models after interpolated by bilinear interpolation over China. Considering the differences between the two climatic variables (i.e., temperature and precipitation), the EDCDF constructs the cumulative distribution function (CDF) of the historical simulated value and the future simulated value of the climate elements, respectively. It improves on previous approaches based only on the historical CDF because it takes into account any changes in the future distribution (Aloysius et al. 2016; Yang et al. 2018). Equation (1) is used to bias correct the future GCM simulations of the temperature, adopting the beta distribution with four parameters [Eq. (2)]:

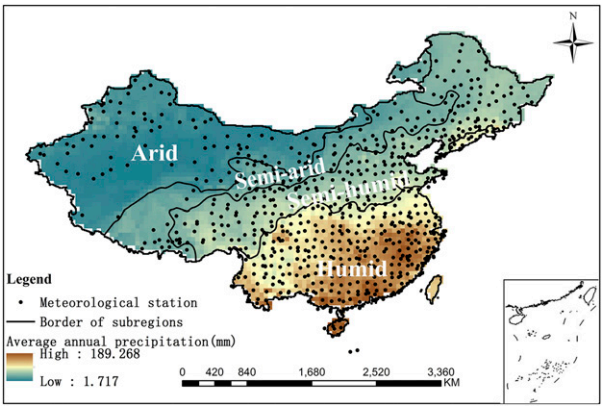


FIG. 1. Distribution of meteorological stations, the elevation, and classification of the four climate regions based on the long-term (1961–2005) averaged precipitation over China.

$$x_{m\text{–}p\_\text{adjust}} = x_{m\text{–}p} + F^{-1}_{o\text{–}t}[F_{m\text{–}p}(x_{m\text{–}p})] - F^{-1}_{m\text{–}t}[F_{m\text{–}p}(x_{m\text{–}p})], \tag{1}$$

$$f(x;a,b,p,q) = \frac{1}{B(p,q)(b-a)^{p+q-1}}(x-a)^{p-1}(b-x)^{q-1}, \tag{2}$$

$$a \le x \le b; \quad p,q > 0.$$

Equation (3) is used to bias-correct precipitation, with a two-parameter mixed gamma distribution [Eq. (4)] considering the intermittent nature of precipitation:

$$x_{m\text{–}p\_\text{adjust}} = x_{m\text{–}p}\frac{F^{-1}_{o\text{–}t}[F_{m\text{–}p}(x_{m\text{–}p})]}{F^{-1}_{m\text{–}t}[F_{m\text{–}p}(x_{m\text{–}p})]}, \tag{3}$$

$$f(x;k,\theta) = x^{k-1}\frac{e^{-x/\theta}}{\theta^k\Gamma(k)}, \quad \text{for} \quad x>0 \quad \text{and} \quad k,\theta>0, \tag{4}$$

where $x_{m\text{–}p}$ is the model projection value; $x_{m\text{–}p_{\text{–adjust}}}$ is the adjusted model projection value after bias correction; $F^{-1}_{o\text{–}t}$ and $F_{m\text{–}t}$ is the quantile function corresponding to the observations o and simulation m in the training period t; and $F_{m\text{–}p}$ is the CDF of the model simulated fields; B is the beta function, a and b are the range parameters as the extreme values from the data, σ is the standard deviation, and p and q are the shape parameters determined by the maximum likelihood estimation method. In the EDCDF method, the parametric distributions are fitted to both temperature and precipitation fields for each grid point. Meanwhile, the distribution range parameters are taken as the extreme values from the data extended by half of one standard deviation of each grid point. Further details

TABLE 1. Summary of the 25 climate models used in this study.

No.	Model Name	Country	Institute	Resolution (lat × lon)
1	ACCESS1.0	Australia	Commonwealth Scientific and Industrial Research Organisation (CSIRO) and Bureau of Meteorology (BOM)	1.25° × 1.875°
2	ACCESS1.3	Australia	CSIRO and BOM	1.25° × 1.875°
3	BCC_CSM1.1	China	Beijing Climate Center	2.8° × 2.8°
4	CanESM2	Canada	Canadian Centre for Climate Modeling and Analysis	2.8° × 2.8°
5	CCSM4	United States	National Center for Atmospheric Research	0.9° × 1.25°
6	CMCC-CM	Italy	Centro Euro-Mediterraneo per i Cambiamenti Climatici	0.75° × 0.75°
7	CMCC-CMS	Italy	Centro Euro-Mediterraneo per i Cambiamenti Climatici	3.7° × 3.75°
8	CNRM-CM5	France	Centre National de Recherches Meteorologiques and Centre Européen de Recherche et Formation Avancée en Calcul Scientifique	1.4° × 1.4°
9	CSIRO Mk3.6.0	Australia	Queensland Climate Change Centre of Excellence and CSIRO	1.8° × 1.8°
10	GFDL CM3	United States	National Oceanic and Atmospheric Administration (NOAA)/Geophysical Fluid Dynamics Laboratory (GFDL)	2° × 2.5°
11	GFDL-ESM2G	United States	NOAA/GFDL	2° × 2.5°
12	GFDL-ESM2M	United States	NOAA/GFDL	2° × 2.5°
13	GISS-E2-R	United States	Goddard Institute for Space Studies	2° × 2.5°
14	HadGEM2-AO	United Kingdom/ South Korea	Met Office Hadley Centre/National Institute for Medical Research	1.25° × 1.875°
15	INM-CM4	Russia	Institute for Numerical Mathematics	1.5° × 2°
16	IPSL-CM5A-LR	France	Institute Pierre-Simon Laplace	1.9° × 3.75°
17	IPSL-CM5A-MR	France	Institute Pierre-Simon Laplace	1.3° × 2.5°
18	IPSL-CM5B-LR	France	Institute Pierre-Simon Laplace	1.9° × 3.75°
19	MIROC5	Japan	Atmosphere and Ocean Research Institute (The University of Tokyo), National Institute for Environmental Studies (NIES), and Japan Agency for Marine-Earth Science and Technology (JAMSTEC)	1.4° × 1.4°
20	MIROC-ESM	Japan	JAMSTEC, Atmosphere and Ocean Research Institute (The University of Tokyo), and NIES	2.8° × 2.8°
21	MIROC-ESM-CHEM	Japan	JAMSTEC, Atmosphere and Ocean Research Institute (The University of Tokyo), and NIES	2.8° × 2.8°
22	MPI-ESM-LR	Germany	Max Planck Institute for Meteorology	1.865° × 1.875°
23	MRI-CGCM3	Japan	Meteorological Research Institute	1.12° × 1.125°
24	NorESM1-M	Norway	Norwegian Climate Centre	1.9° × 2.5°
25	NorESM1-ME	Norway	Norwegian Climate Centre	1.9° × 2.5°

about this method can be found in Li et al. (2010) and Yang et al. (2018).

Reifen and Toumi (2009) and Li et al. (2010) found that bias-corrected time series are directly tied to the model's performance during the training period. In this study, the training period is from 1961 to 1990 for calibrating the model parameters and the validation period is from 1991 to 2005. We choose the period of 1961–90 as the training period for two reasons that this period is the reference period of many climate change studies (e.g., Giorgi and Mearns 2002) and have experienced warming and thus are more likely to resemble future projections (Yang et al. 2018).

b. *Performance evaluation methods*

We use the Taylor diagram (Taylor 2001) to quantify the degree of correspondence between the original and bias-corrected CMIP5 modeled and observed monthly climate fields. Both precipitation and temperature are smoothed by a 3-month running mean, which is a typical filter for examining interannual anomalies (Su and Neelin 2003). The diagram can effectively reflect the strength and weaknesses of each model's simulated results, especially for large ensembles. It utilizes the triangular conversion relationship among three indices, which are the correlation coefficient (CC), standard deviation (SD), and root-mean-square error (RMSE). The CC represents the strength and direction of the relationship between changes in two variables. The SD reflects the degree of dispersion of a dataset. The RMSE measures the deviation between the simulated and observed values. The normalized SD (NSD) is defined as the ratio of the standard deviation of simulated and observed climate fields. The closer the CC and NSD

are to 1, the better are the simulation values associated with the observed values. The smaller the RMSE, the smaller the errors between the simulated data and the observed data.

We apply the skill score (SS; Taylor 2001) to evaluate the quantitative performance of the individual models over the observed time period. The skill score is defined as

$$\mathrm{SS} = \frac{4(1+\mathrm{CC})^4}{\left(\mathrm{NSD}_m + \frac{1}{\mathrm{NSD}_m}\right)^2 (1+\mathrm{CC}_0)^4}, \tag{5}$$

where m indicates model simulation; NSD_m is the normalized standard deviation of the simulation; CC_0 is the maximum correlation coefficient; and CC is the correlation coefficient between the simulated and observed data. The closer SS is to 1, the better the ability of the individual model to represent the observations. The SS is used to identify the best performing models, which are then used later to form a multimodel ensemble (MME).

The MME average reduces the uncertainty derived from each individual model, while the performance of each individual model simulation directly affects the performance of the MME. Therefore, based on the performance of the individual models, a proportion of the models with better performances are combined to form the MME. By calculating different combinations of models (starting from single model, averaging two models, averaging three models, etc.), the skill score of each model combination is computed. The number of combined models with the highest skill score is the optimal number, and the corresponding models are the best model combination. For each model combination, all the models in the combination are averaged by simple averaging.

Once the MME average is calculated, we evaluate its performance of multimodel ensemble using the comprehensive index (CI) to determine the optimal multimodel:

$$\mathrm{CI} = \frac{\mathrm{SS}_{\mathrm{ME1\_pr}} + \mathrm{SS}_{\mathrm{ME1\_tas}}}{\mathrm{SS}_{\mathrm{ME2\_pr}} + \mathrm{SS}_{\mathrm{ME2\_tas}}}, \tag{6}$$

where ME1_pr, ME1_tas, ME2_pr, and ME1_tas are two selected multimodels of precipitation and temperature, respectively. Parameters $\mathrm{SS}_{\mathrm{ME1\_pr}}$ and $\mathrm{SS}_{\mathrm{ME1\_tas}}$ are the skill score of precipitation and temperature from selected multimodels of ME1, respectively. Similarly, $\mathrm{SS}_{\mathrm{ME2\_pr}}$ and $\mathrm{SS}_{\mathrm{ME2\_tas}}$ are the ME2's skill scores of precipitation and temperature, respectively. If $\mathrm{CI} > 1$, then ME1 is selected as the optimal ensemble; if $\mathrm{CI} < 1$, then ME2 is selected as the optimal ensemble; if $\mathrm{CI} = 1$, both ME1 and ME2 are the optimal ensembles.

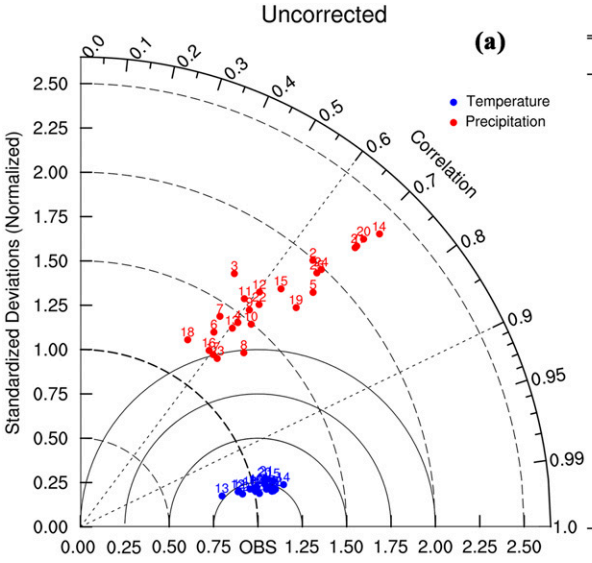
4. Results

a. *Performance assessment of each individual model*

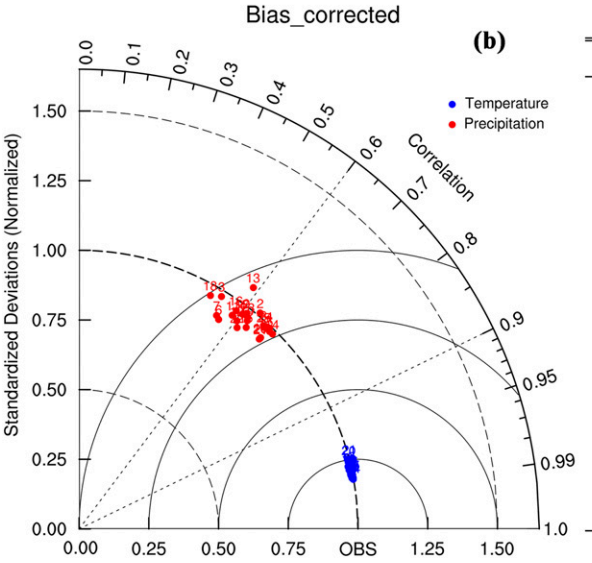
We calculate the NSD, RMSE, and CC of the 25 models (uncorrected models and bias-corrected models) and summarize the results in Taylor diagrams (Fig. 2), which are used to qualitatively reflect the models' performance. The polar plot of the Taylor diagram shows that, for both precipitation (red) and temperature (blue), the CC values of uncorrected models and observed data are similar to that of bias-corrected models and observed data. However, the uncorrected models show poor performance for the NSD, and the deviation of the NSD from 1 of the bias-corrected models is much less than that of the uncorrected models. For precipitation, the NSD of the uncorrected models can reach 2.3 and temperature can reach 1.1. The skill scores of the uncorrected models are generally lower than that of bias-corrected models. In general, although the CC of the uncorrected models is similar to that of bias-corrected models, their performance based on NSD is poor, resulting in their skill score being lower than bias-corrected models. Bias-corrected models therefore perform better than uncorrected models in overall performance. Therefore, the bias-corrected models are used hereinafter.

The Taylor diagram shows that the CC value of the simulated and observed precipitation (red dots) data is 0.4–0.7, showing a moderate correlation. The NSD of the simulated precipitation is far from the observed value, indicating a poor performance. The CC for temperature varies from 0.95 to 0.98, showing a strong correlation, and the NSD values are close to 1, which indicates that temperature simulation has a better skill than for precipitation. The SS of the 25 models (Fig. 2b) show that the temperature SS are higher than those of precipitation, which are similar to the results of the three indices of the Taylor diagram. The SS of precipitation is less than 0.54, with model CCSM4 (No. 5) having the highest SS (0.547) for precipitation, while IPSL-CM5B-LR (No. 18) has the lowest SS of 0.383. The overall SS of temperature is above 0.93. The highest SS is 0.968 for model CSIRO Mk3.6.0 (No. 9) and NorESM1-M (No. 24), and the lowest value is 0.939 for MIROC-ESM-CHEM (No. 21).

The spatial distribution bias of the models with the highest and lowest SS is evaluated in Fig. 3. Model IPSL-CM5B-LR and CCSM4 have the worst and best performance for precipitation, respectively. For temperature,



	Skill Scroes	
	Precipitation	Temperature
1	0.423	0.941
2	0.416	0.946
3	0.384	0.947
4	0.383	0.934
5	0.450	0.945
6	0.374	0.943
7	0.370	0.940
8	0.452	0.944
9	0.419	0.948
10	0.404	0.940
11	0.406	0.937
12	0.416	0.935
13	0.373	0.885
14	0.435	0.926
15	0.430	0.916
16	0.410	0.931
17	0.422	0.939
18	0.345	0.942
19	0.443	0.940
20	0.396	0.920
21	0.401	0.917
22	0.416	0.945
23	0.402	0.943
24	0.418	0.949
25	0.415	0.945



	Skill Scroes	
	Precipitation	Temperature
1	0.523	0.961
2	0.492	0.967
3	0.436	0.962
4	0.474	0.954
5	0.547	0.967
6	0.427	0.957
7	0.423	0.957
8	0.525	0.961
9	0.496	0.968
10	0.484	0.954
11	0.481	0.955
12	0.494	0.954
13	0.426	0.955
14	0.545	0.960
15	0.504	0.949
16	0.451	0.953
17	0.482	0.956
18	0.383	0.960
19	0.540	0.955
20	0.526	0.941
21	0.525	0.939
22	0.496	0.961
23	0.489	0.965
24	0.542	0.968
25	0.538	0.964

FIG. 2. Taylor diagram and SS summarizing the performance of precipitation (red color) and temperature (blue color) from 25 CMIP5 models. (a) Uncorrected CMIP5 models and (b) bias-corrected CMIP5 models. The 14 models were selected (highlighted by the cyan color), whose SS values are higher than the mean of the 25 models.

the best performance model is CSIRO Mk3.6.0 but the worst model is MIROC-ESM-CHEM. Note that the spatial patterns of the simulated precipitation and temperature of the selected models are markedly different. For precipitation, both the highest and lowest SS models (IPSL-CM5B-LR and CCSM4) have lower bias in the northwest and southeast regions, while relatively higher bias in northeast and southwest China. Meanwhile, CCSM4 has a relatively lower bias over most parts of the southwest and northeast China. On the other hand, for temperature, the overall bias of CSIRO Mk3.6.0 is less than 0.2°C over most parts of China, except for some border areas in the southwest China. MIROC-ESM-CHEM has a higher bias over most parts of northern China.

We classify and compare precipitation and temperature based on the averaged SS value of 25 models over the four regions (arid, semiarid, semihumid, and humid regions) in Fig. 4. Results show that precipitation in the semihumid region has the highest average SS (0.639), and the semiarid region has the second highest SS (0.595), while the arid area has the lowest SS (0.260). Simulated precipitation in the four regions has a high consistency. For example, NorESM1-M (No. 24) has higher SS than other models in the semiarid and

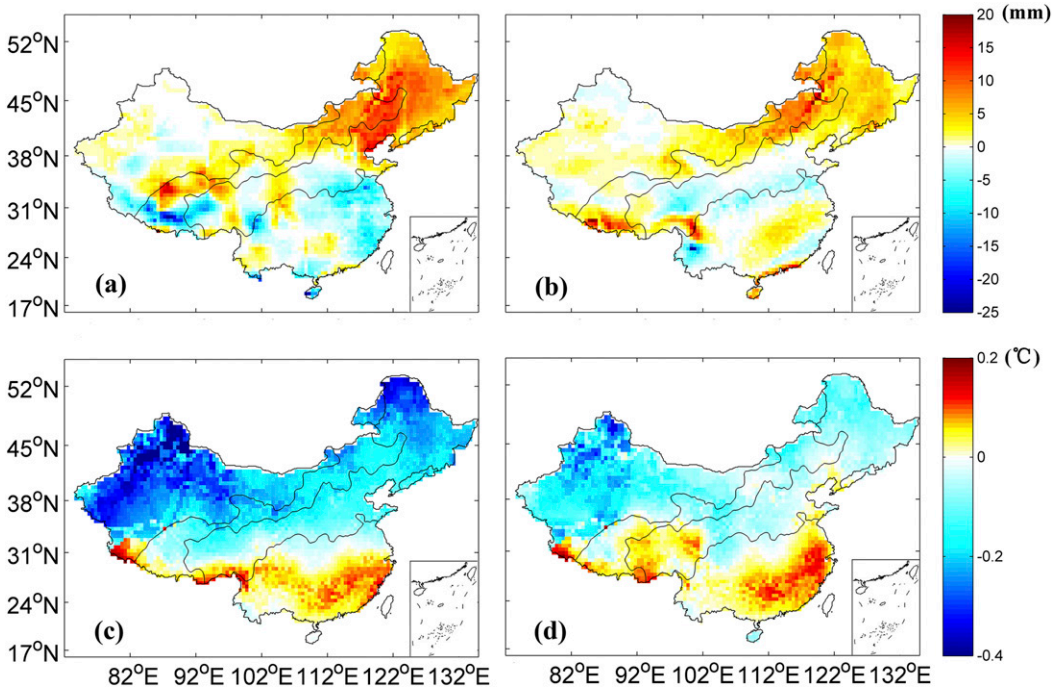


FIG. 3. Maps showing the spatial distribution of the bias of the model with the worst and best simulation capability. The (a) worst (IPSL-CM5B-LR) and (b) best (CCSM4) model performance for precipitation and the (c) worst (MIROC-ESM-CHEM) and (d) best (CSIRO Mk3.6.0) model performance for temperature.

semihumid areas, while the SS of IPSL-CM5B-LR (No. 18) is lower in all regions except for arid area. On the other hand, temperature shows the highest SS (0.966) in the semihumid area, while the lowest value (0.945) in the humid region. The simulation abilities in the semiarid area and arid area are slightly lower than that in the semihumid area, respectively. The simulations of temperature by the models are consistent in arid, semiarid, and semihumid regions. Models [i.e., CMCC-CM (No. 6), CMCC-CMS (No. 7), GFDL-ESM2G (No. 11), and GFDL-ESM2M (No. 12)] with a poor simulation accuracy in the humid areas have better simulations in the other regions.

In general, both temperature and precipitation have higher SS in the semiarid and semihumid regions than the other two regions. In other words, the 25 GCMs are more applicable in simulating the climatic fields in the semiarid and semihumid regions of China. At the same time, models with higher SS for the entire China scored relatively well in the four regions.

b. Optimization of multimodel compositions

Many former studies found that removing the models with obvious poor skills would certainly be better than assessing a correlation based on the entire archive and significantly reduce the range of projections sampled (e.g., McSweeney et al. 2015; Sanderson et al. 2015; Herger et al. 2018). Thus we define threshold for the performance of skill scores. The threshold selected is the mean value of the skill score. Precipitation simulation from the models is far less skillful than temperature over China, so the precipitation is regarded as a constraint for model selection in this study. The models are selected using a criterion that the precipitation SS is higher than 0.490 (mean SS of 25 CMIP5 models), resulting in 14 models selected, which are highlighted in cyan color in Fig. 2b. Based on the SS of each model, 14 models are systematically selected for combination with Eq. (6), yielding 16 383 combinations in total. Figure 5 shows the SS from these combinations for each group of C_{14}^{i} ($i = 1, 2, \ldots, 14$). It shows that the simulated abilities of the combined models stabilize or reach a maximum when the number of models reaches a certain value. As more model runs are included in the ensemble, the SS decreases again. For precipitation, the best performance over China is achieved with a combination of nine models (Fig. 5a): ACCESS1.0 (No. 1), ACCESS1.3 (No. 2), CCSM4 (No. 5), CNRM-CM5 (No. 8), HadGEM2-AO (No. 14), MIROC5 (No. 19), MIROC-ESM (No. 20), MIROC-ESM-CHEM (No. 21), and NorESM1-M (No. 24) out of the 2002 possible combinations of C_{14}^{9}. For temperature, the best performance of the multimodel combination is achieved with 11 models (Fig. 5b): ACCESS1.0 (No. 1), ACCESS1.3 (No. 2), BCC_CSM1.1 (No. 3), CCSM4 (No. 5), CMCC-CM (No. 6), CNRM-CM5

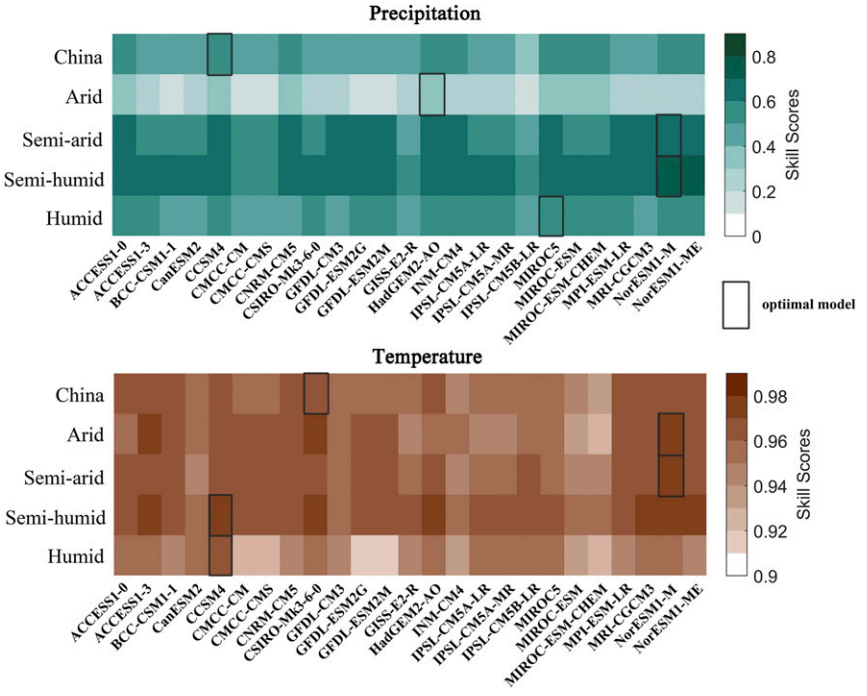


FIG. 4. SS of simulated precipitation and temperature from each GCM model over four regions (arid, semiarid, semihumid, and humid regions) in China. The black rectangle represents the best single model.

(No. 8), CSIRO Mk3.6.0 (No. 9), HadGEM2-AO (No. 14), IPSL-CM5B-LR (No. 18), MPI-ESM-LR (No. 22), and NorESM1-M (No. 24) out of the 364 possible combinations of C_{14}^{11}. We obtain two multimodel combinations: the 9 models ensemble (9ME) for precipitation and the 11 models ensemble (11ME) for temperature.

c. *Comprehensive index*

A single evaluation index cannot effectively and accurately describe the merits of a model's performance. Therefore, a CI is proposed to further evaluate the performance of two multimodel ensembles. The CI is obtained by a weighted average of the SS of two multimodel ensembles (9ME and 11ME). Table 2 shows the performance metrics for China and the four regions. Nationwide, the CI of the 9ME is slightly higher than that of the 11ME. The 9ME is superior in the arid (1.084), semiarid (1.007), and semihumid (1.006) regions than that of the 11ME but trivial lower in the humid region. It demonstrates that the 9ME is slightly superior to the 11ME, which is more suitable for China as a whole. Therefore, we choose the 9ME as the optimal multimodel ensemble for China.

d. *Evaluation of the multimodel ensemble*

Table 3 lists the SS of the single model optimal (SMO; the individual model with the highest SS, which is NorESM1-M for precipitation and CSIRO Mk3.6.0 for temperature), 25 models ensemble (25ME), 9ME, and 11ME of precipitation and temperature. It shows that the skills of the multimodel combinations for precipitation have been significantly improved by

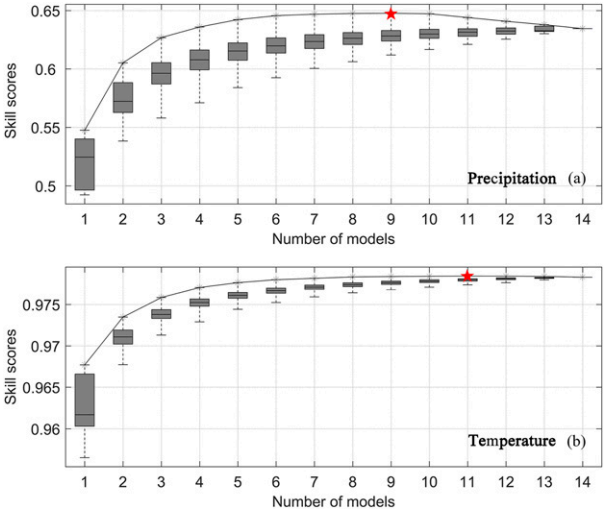


FIG. 5. The SS of each group of C_{14}^{i} (i = 1, 2, ... , 14) for (a) precipitation and (b) temperature. The asterisks represent the combination with the highest SS for each group. The red star represents the group with the highest SS for all combinations.

TABLE 2. Model performance indicators and the optimal model selection results for China and the four regions.

Areas	Models	Climatic variables	SS	Mean value of SS	CI	Optimal model
China	9ME	*P*	0.648	0.811	1.022	9ME
		T	0.975			
	11ME	*P*	0.608	0.793		
		T	0.978			
Arid	9ME	*P*	0.434	0.703	1.084	9ME
		T	0.973			
	11ME	*P*	0.320	0.649		
		T	0.978			
Semi-arid	9ME	*P*	0.760	0.869	1.007	9ME
		T	0.977			
	11ME	*P*	0.746	0.863		
		T	0.980			
Semi-humid	9ME	*P*	0.778	0.879	0.999	11ME
		T	0.981			
	11ME	*P*	0.777	0.880		
		T	0.983			
Humid	9ME	*P*	0.665	0.816	1.006	9ME
		T	0.967			
	11ME	*P*	0.650	0.811		
		T	0.971			

7%–10% in all regions except for arid areas, compared to the SMO. Although their performances show large variability, the spatial patterns of the SS of these four types of model combinations are similar with better performance in the southwest and northeast than other regions (Figs. 6a–d). Except for the SMO, the simulated effects of precipitation from other models in the semiarid and semihumid areas (SS above 0.68) are slightly better than those in humid areas (approximately 0.57), which is better than those in dry areas (less than 0.32).

The SMO shows lower skill in simulating temperature over most border areas of southwest China (Figs. 6e–h), which has the highest elevations and distinctive geographic features are likely to increase the errors in model simulations in this region (Song et al. 2013). Meanwhile, the SS for humid regions is lower than that for other regions. For example, the simulated accuracy of the humid area (approximately 0.95) is significantly lower than that in other areas. Meanwhile, the SS accuracies are similar in the arid and semiarid regions, both of which are lower than that in the semihumid region (0.975). It shows that the three multimodel ensembles have better performance than that of the SMO.

The relative absolute bias (ABIAS) is the ratio of the mean of the absolute deviations to the mean of observations, which can reflect the degree of deviation of the simulated sequence from the observed sequence. We compare the boxplots (maximum, the first quartile, the median quartile, the third quartile, and minimum) of the ABIAS of two model ensembles (9ME and 11ME) in Fig. 7. It shows that the 9ME has slightly lower ABIAS value of precipitation than that of the 11ME over China and the four regions. The 9ME and the 11ME have significantly higher ABIAS in the arid areas than other regions. On the other hand, the 11ME has lower ABIAS of temperature over China and the four regions. Among these, the ABIAS of temperature in the arid region is significantly higher than in other areas, which is lowest in the humid regions.

Figure 8 presents the bias of the multiyear averaged precipitation and temperature of the 9ME in each month. It demonstrates that there is a high consistency between the multiyear averaged precipitation simulated by the 9ME and the observed values, with the large seasonal variation. According to the seasonal pie chart, summer precipitation is the most abundant, accounting for approximately 50% of the annual precipitation, and winter precipitation is the lowest, less than 7% of the annual total, which is similar with the results of Liang et al. (2011) and Sui et al. (2013). However, the 9ME slightly underestimates the precipitation in JJA (about 2.6%) and SON (about 1%) and slightly overestimates the precipitation in MAM (about 3.5%). Similarly, the 9ME is accurate in simulating the seasonality of temperature. The monthly temperature values from

TABLE 3. The SS of the SMO and the three model combinations (25ME, 9ME, and 11ME) for the simulated precipitation and temperature in the five regions.

		China	Arid	Semiarid	Semihumid	Humid
Precipitation	SMO	0.547	0.392	0.682	0.728	0.573
	25ME	0.613	0.321	0.762	0.785	0.647
	9ME	0.648	0.434	0.760	0.778	0.665
	11ME	0.608	0.320	0.746	0.777	0.650
Temperature	SMO	0.968	0.971	0.971	0.974	0.961
	25ME	0.976	0.975	0.979	0.983	0.968
	9ME	0.975	0.973	0.977	0.981	0.967
	11ME	0.978	0.978	0.980	0.983	0.971

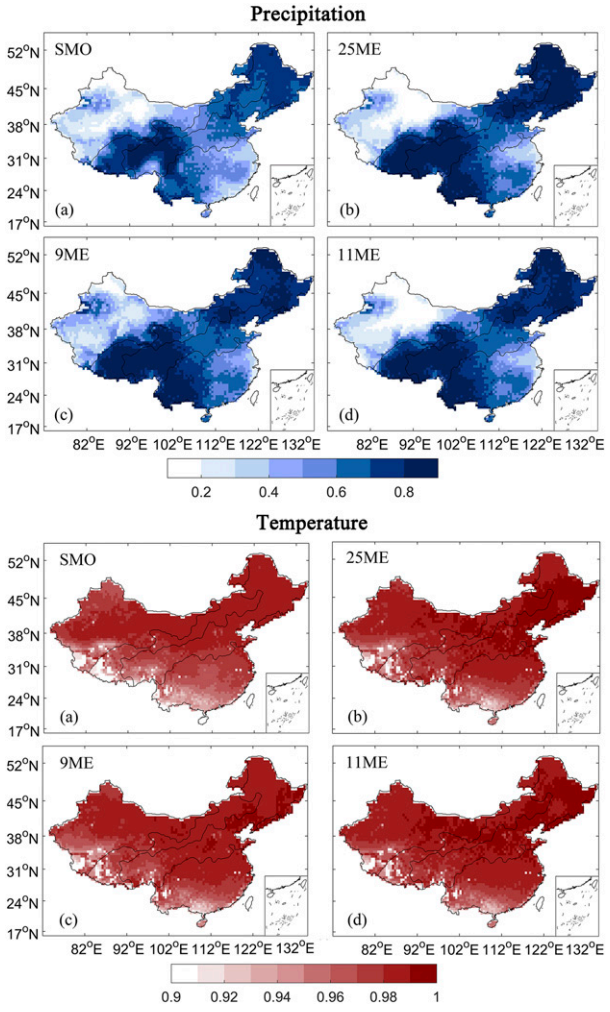


FIG. 6. Spatial distribution of the SS from the SMO, 25ME, 9ME, and 11ME during 1960–2005.

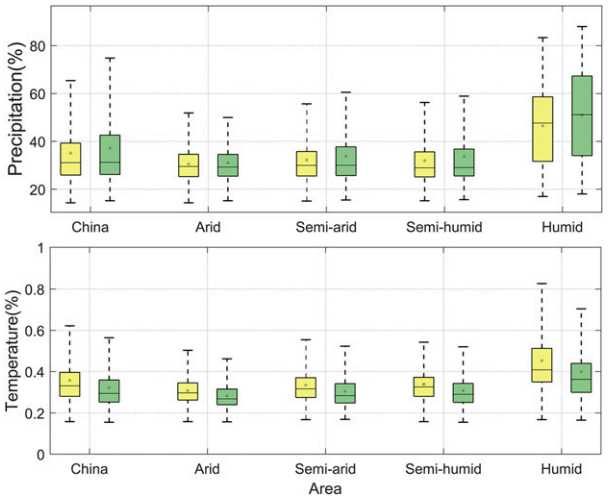


FIG. 7. ABIAS of the 9ME (yellow) and 11ME (green) in the four subregions and China. The asterisk represents the regional average.

the 9ME are lower than the observed temperatures in January, March, and December, and higher the observed values in July, August, and November.

The bias of the multiyear monthly average precipitation varies between −0.9 and 1 mm month^{-1} (Fig. 9). In the western part of the Tibetan Plateau, the deviation of precipitation is greater than 1 mm month^{-1}. The bias of the multiyear monthly average temperature is negative (from −0.4° to 0°C) in the northwest and northeast of China and positive in the southeast area (>0.05°C). Furthermore, the 9ME can well reproduce the spatial characteristics of the seasonal [MAM (March–May), JJA (June–August), SON (September–November) and DJF (December–February)] climatic fields (Fig. 10). In most parts of China, especially in the humid part, the precipitation of 9ME underestimates observed precipitation. The temperature from the 9ME has a larger bias in DJF, which is underestimated in the west and southwest but overestimated in other regions (Fig. 11). The SON average temperature shows a negative bias in the eastern region and a positive bias in other regions. The biases (from −2° to 2°C) of temperature in MAM and JJA in the semiarid and semihumid regions are smaller than other areas. In general, the 9ME can well simulate the spatial–temporal patterns of the observed precipitation and temperature over China.

5. Discussion

In this study, we apply the SS to assess the performance of 25 climate models over China. We select 14 models to optimize the combination of models with higher SS. The optimal combination is 9ME and 11ME for precipitation and temperature, respectively. According to the CI, we conclude that the optimal ensemble is the 9ME over China, which yields better performance in arid, semiarid and semihumid regions than that in humid region. Simulations of the 9ME are verified with the observed climatic fields. The 9ME has good skill in simulating the annual variation of precipitation and temperature.

However, there are three issues related to the performance of the MME to be further discussed: 1) the uncertainty of model selection and observational data, 2) the underestimation of interannual variability by the MME, and 3) the weighting of individual models when forming the ensemble.

a. The uncertainty of model selection and observational data

This study focuses on developing an optimal subset of climate models for making informed recommendations to

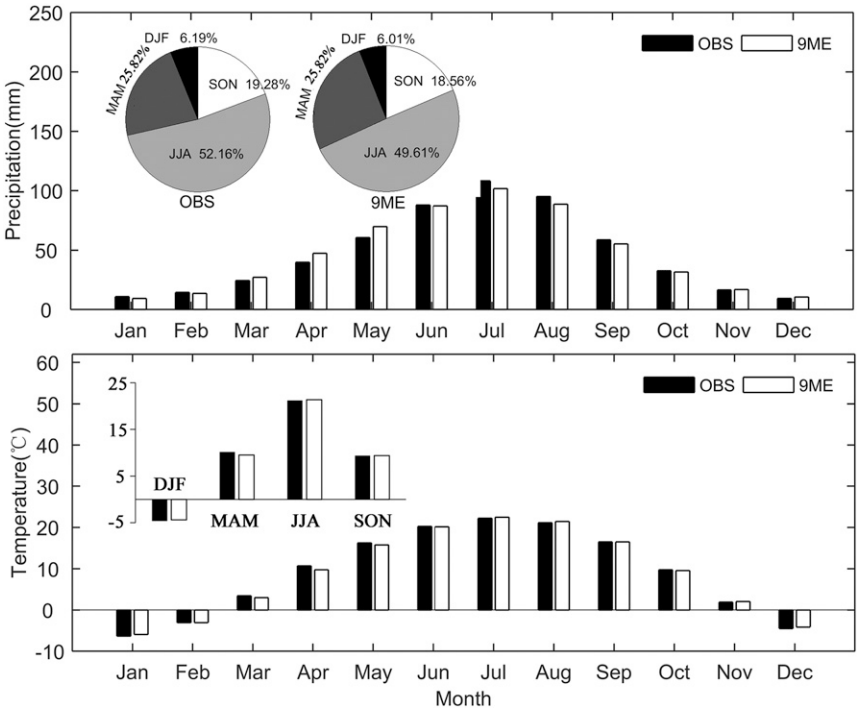


FIG. 8. Monthly climatology of (top) precipitation and (bottom) temperature.

those who may use these model outputs for climate changes effects on hydrology in China. Input uncertainties from climate forcing play an important role of the overall uncertainty quantification of climate change impact (Herrera-Estrada and Sheffield 2017; Ahmadalipour et al. 2017). Therefore, we need to pay particular attention to the uncertainty when arbitrarily choosing a small number of models.

To evaluate the uncertainty of subsets of multimodel ensemble, we follow the methodology and performance indices proposed by Herrera-Estrada and Sheffield (2017) to look at the uncertainty from different ensemble sizes for all combinations of models and comparing with ensembles of the best performance over China. We use bootstrap sampling to randomly select model runs from the pool

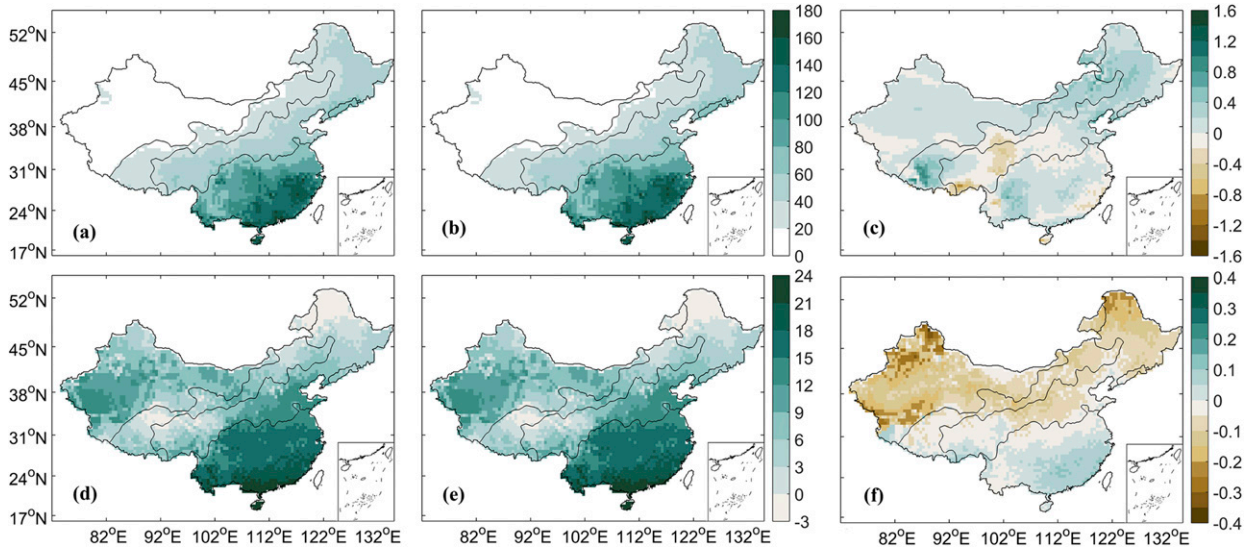


FIG. 9. Spatial distribution of (top) precipitation (mm month^{-1}) and (bottom) temperature (°C) for the (a),(d) observation, (b),(e) 9ME simulation, and (c),(f) bias during the 1960–2005.

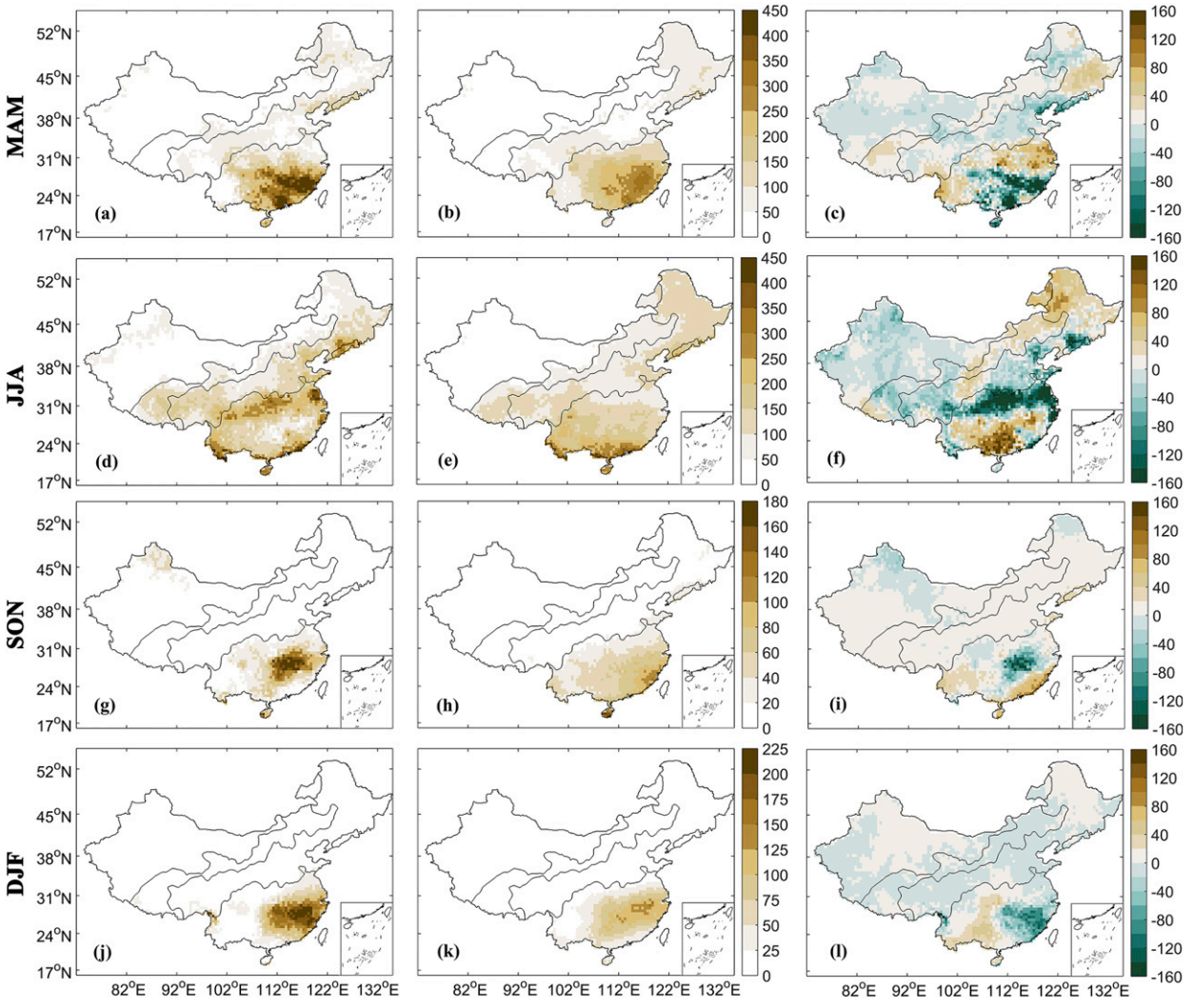


FIG. 10. Spatial distribution of the (left) observed precipitation (mm season^{-1}; three months), (center) the 9ME simulated precipitation (mm season^{-1}), and (right) bias (mm season^{-1}; 9ME-OBS) for (a)–(c) MAM, (d)–(f) JJA, (g)–(i) SON, and (j)–(l) DJF during 1960–2005.

(25 models) 1000 times without repetition for each ensemble size in order to gauge sampling uncertainty. The uncertainty range is found not to be very sensitive to the number of iterations. We present the ranges of the sampling uncertainty for ensemble sizes ranging from 5 to 25, 9ME, and 11ME in Fig. 12. The 25 models are ranked according to the skill score in precipitation and temperature separately. The interquartile range of the changes compared to the observed precipitation and temperature is calculated for each sample as a measure of uncertainty.

Results show that selecting a small sample of models using the rankings based on temperature yields overall larger uncertainties in precipitation than the median of the bootstrap analysis. In comparison, the ranking from precipitation consistently produces a lower uncertainty range. It is noticed that the 9ME has significantly lower uncertainty values for both precipitation and temperature. It indicates that selecting small subsets of the CMIP5 models will likely artificially reduce the uncertainty range of the projections (e.g., Gleckler et al. 2008; Knutti et al. 2010). The method developed in this study can be used for optimal subset in China for climate changes applications. It is similar to the suggestion of Gleckler et al. (2008) and Knutti et al. (2010).

This study uses meteorological station data interpolated to 0.5° grids as the reference data. Although the interpolated data fitted well with the observational data (Fig. S1 in the online supplemental material), we still need to pay attention to the quality of the observed

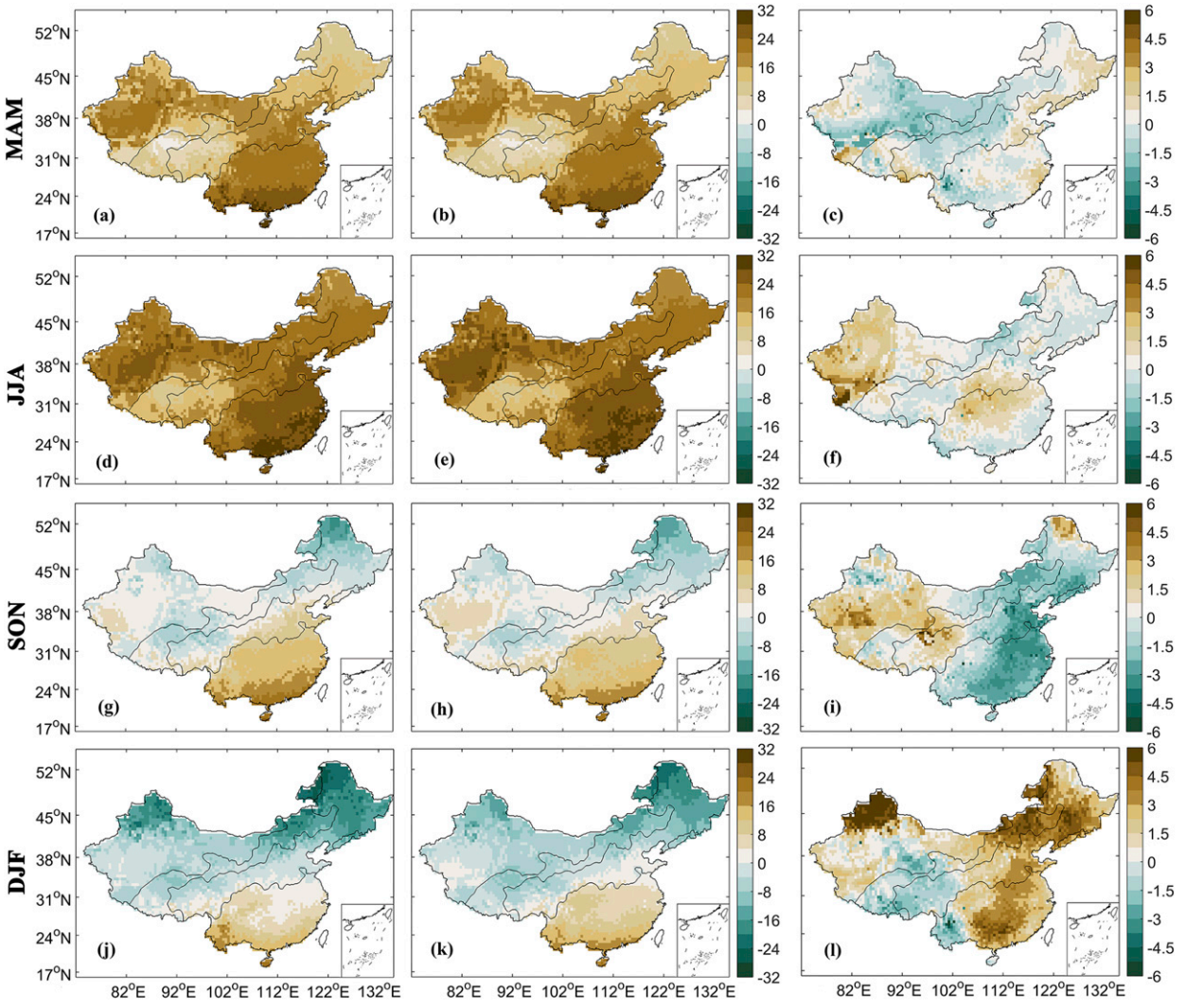


FIG. 11. As in Fig. 10, but for temperature (°C).

dataset. The measurement error can be largely depending on the variables (precipitation or temperature) and can thus result in a different optimal subset. Previous studies found that the optimal ensemble of climate models is sensitive to the metric, observational product, and preprocessing steps used (Yin et al. 2015; Herger et al. 2018). Therefore, the researchers need pay attention to the observational datasets for further application of the optimal ensemble method in this study.

The results show that the optimal combination has relatively poor performance on both precipitation and temperature in northwest China. In addition, optimal combination of precipitation shows good skills in most climate zones, except for dry areas, which may be related to the difficulties in modeling the inherent spatial variability and characteristics of precipitation (Stephens and Ellis 2008; Yang et al. 2019). However, uncertainties from the optimal ensemble combination are expected to be larger at smaller scales (e.g., river basin), which are more relevant to hydrological applications and need to be investigated in future work.

b. *Interannual variability*

We compare the normalized standard deviation (standard deviations are normalized by the standard deviation of the observed fields) of selected individual model and the 9ME in Fig. 13. It shows that the multimodel ensemble underestimates the interannual variability over the time period. For each individual model, the standard deviation is similar to the observations. However, the 9ME shows a significant decrease for precipitation, indicating that it underestimates the interannual variations of precipitation compared to observations. These results are consistent with previous

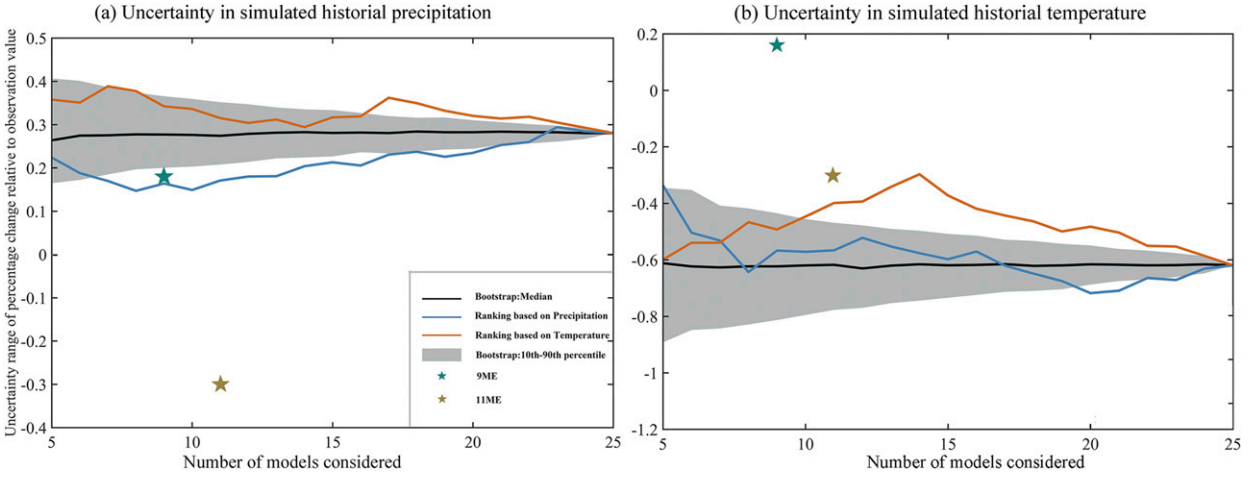


FIG. 12. Interquartile range of (a) precipitation and (b) temperature over China. The black line represents the median value when randomly sampling a subset of models, and the envelope corresponds to the 10th and 90th percentiles (derived from 1000 repetitions for each subset). Colored lines represent the range of uncertainty when using the models along the order of the rankings based on precipitation (blue) and temperature (orange). Colored stars represent the selected 9ME (green) and 11ME (brown) in this study.

studies (e.g., Cavazos and Arriaga-Ramírez 2012; Xu and Xu 2012; Chen and Frauenfeld 2014). Because each model is free running (no internal constraints on sea surface states, for example), they will have different peaks (wet or warm) and troughs (dry or cold), and so when averaged these peaks and troughs will tend to cancel each other out.

The interannual variability in each model is mainly driven by ENSO, one of the major drivers of interannual variability in China (Zou and Zhou 2015). Understanding how ENSO and its teleconnections with regional climate are represented in climate models is important for understanding the fidelity of models to simulate the year to year variations that have the most impacts (Guilyardi et al. 2009; Taylor et al. 2012; Zou and Zhou 2015). Comparing to CMIP3, the CMIP5 models have improved the physical credibility of their simulations of ENSO overall and its response to climate change, exhibiting a behavior qualitatively similar to that of the real-world ENSO (Guilyardi et al. 2009; Sperber et al. 2013; Bellenger et al. 2014). Nevertheless, many studies have shown that CMIP5 models still have some limitations in the simulation of ENSO phenomena (Jha et al. 2014; IPCC 2014). In the context of MME, the CMIP5 experimental design for the ''historical'' model runs is for initialization from a random point of a quasi-equilibrium control run (Taylor et al. 2012). Therefore, El Niño and La Niña years in the historical climate simulations will not necessarily coincide with the regular years in which they actually occurred and will differ across models. Further work, therefore, should evaluate the relationship of climate variability with ENSO in these models, and understand how to better represent the observed interannual variability in multimodel ensemble average.

c. *Impact of ensemble averaging method*

Another caveat in our study is that we combine the models using equal-weight averaging (EWA), without considering the individual skill of the models. Stott et al. (2006) suggested that the observed changes that are attributed to the anthropogenic activities can be used to restrain estimates of future climate warming, and they proposed a weighted climate probability prediction using a measure (e.g., decadal-mean temperature changes of observation) of the model performance evaluated

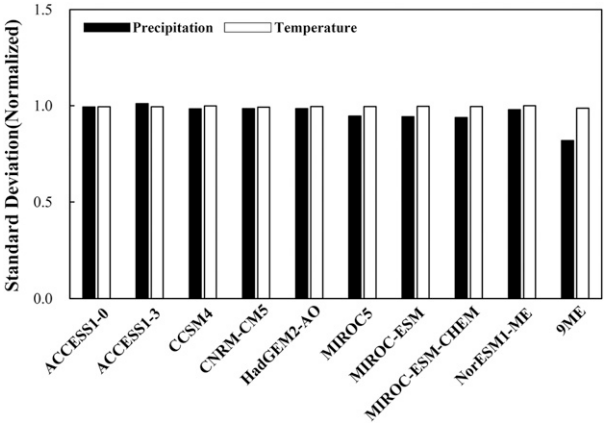


FIG. 13. Normalized standard deviation of each selected model and 9ME simulated annual precipitation and temperature in China from 1961 to 2005.

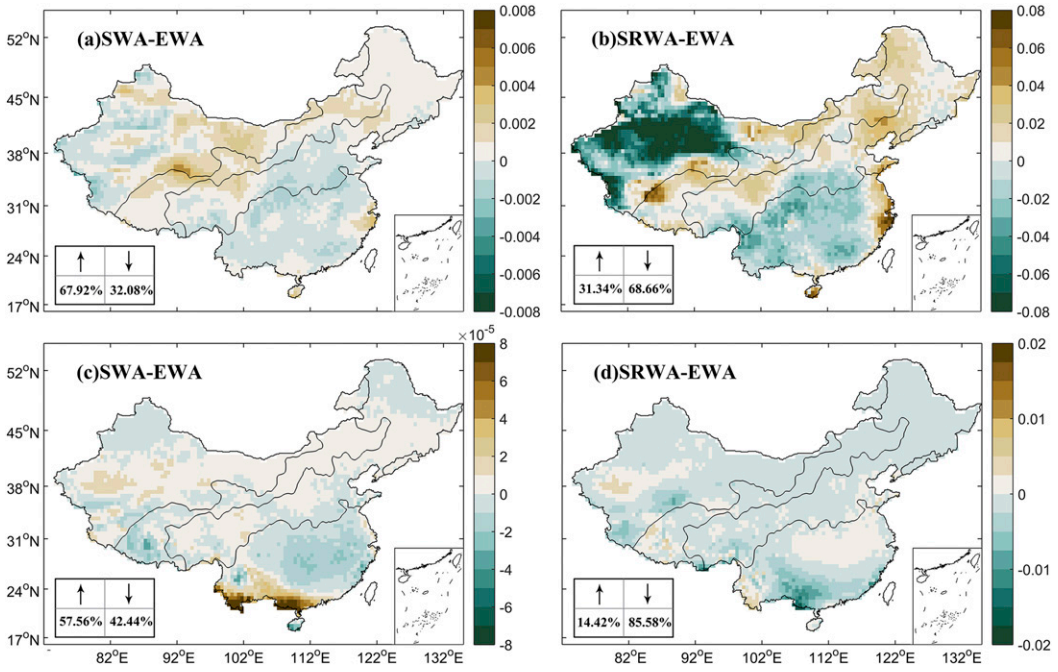


FIG. 14. Spatial distribution of the SS bias (decimation from EWA) for (a),(b) precipitation and (c),(d) temperature with two different weighting methods (SWA and SRWA).

by the observations. Several other studies have used weighted multimodel ensemble methods [e.g., reliability ensemble averaging (Giorgi and Mearns 2002; Xu et al. 2010), rank weighting method (Chen et al. 2011), and Bayesian model average (Raftery et al. 2005; Yang et al. 2011)] to improve the accuracy of the simulations compared with the equal-weighted multimodel ensemble.

Here we use the skill scores of the individual models to explore the impact of weighted ensemble averaging on the performance of the multimodel combination. We compare two different weighted ensemble methods: SS weighting average (SWA) and SS rank weighting average (SRWA) with EWA. SWA takes the value of SS as the weight of each model directly and SRWA ranks the models according to the SS and assigns weight based on the ranks.

Figure 14 presents the differences of SS for precipitation among three different weighting methods (EWA, SWA, and SRWA). It can be seen that the SS of SWA simulated precipitation is higher than EWA over 67.92% of China, but the magnitude of the improvement SS only accounts for 0–6 $\times 10^{-3}$. In contrast, the SRWA performs relatively worse, with only about 31.34% of China has higher SS (ranging from 0 to 0.08) compared with EWA (Figs. 14a,b). For temperature, the weighting effect is slightly better than that without considering the weighting, but the effect is not visually obvious. SWA is slightly better (less than 8×10^{-5}) than EWA in about 57.56% of areas of entire China. However, although the magnitude of deviation between SRWA with EWA is larger than that of SWA–EWA, the SS of 85.58% of the region are less than that of EWA. These results indicate that the weighted average methods have improved the SS in some regions, or for one climate field, than that of EWA, but they show lower skill in other regions/climate fields. We do not find a consistent pattern of better skill among these three multimodel ensembles averaging methods. Therefore, different weighted averaging methods should be used with care in different regions and for different climate fields.

6. Conclusions

In this study, we present an evaluation of the performance of an optimal multimodel ensemble from 25 bias-corrected CMIP5 models over China. The EDCDF method is used to bias-correct and downscale the models, which are then shown to have good skills in simulating climate fields over China with substantial spatial heterogeneity and with relative smaller uncertainty. Furthermore, we identify an optimal multimodel ensemble based on historical performance that can provide future projections associated with model skill. The optimal selected nine member multimodel ensemble (ACCESS1.0, ACCESS1.3, CCSM4, CNRM-CM5, HadGEM2-AO, MIROC5, MIROC-ESM, MIROC-ESM-CHEM, and NorESM1-M) represents well the spatiotemporal

variability of precipitation and temperature over China, albeit with some regional differences and uncertainties. Although the choice of weighted averaging has some regional influence and is variable dependent, the equal weighting method in this study is generally superior across China. The optimal nine models averaged ensemble (9ME) could be applied to hydrologic mode for a range of uses related to climate assessments. The framework developed in this study can be extended and applied to other multimodel ensemble datasets, such as the recently developed sixth phase of the CMIP models (CMIP6; Pascoe et al. 2019), to investigate the robustness of the optimal combination with respect to improved physics and model parameterizations.

Acknowledgments. This work was funded by the National Key Research and Development Program under Grant 2016YFA0601504 approved by Ministry of Science and Technology of the People's Republic of China, the National Natural Science Foundation of China (51579066).

REFERENCES

Ahmadalipour, A., A. Rana, H. Moradkhani, and A. Sharma, 2017: Multi-criteria evaluation of CMIP5 GCMs for climate change impact analysis. *Theor. Appl. Climatol.*, **128**, 71–87, https://doi.org/10.1007/s00704-015-1695-4.

Aloysius, N. R., J. Sheffield, J. E. Saiers, H. Li, and E. F. Wood, 2016: Evaluation of historical and future simulations of precipitation and temperature in central Africa from CMIP5 climate models. *J. Geophys. Res. Atmos.*, **121**, 130–152, https://doi.org/10.1002/2015JD023656.

Bellenger, H., E. Guilyardi, J. Leloup, M. Lengaigne, and J. Vialard, 2014: ENSO representation in climate models: From CMIP3 to CMIP5. *Climate Dyn.*, **42**, 1999–2018, https://doi.org/10.1007/s00382-013-1783-z.

Cannon, A. J., 2016: Multivariate bias correction of climate model output: Matching marginal distributions and intervariable dependence structure. *J. Climate*, **29**, 7045–7064, https://doi.org/10.1175/JCLI-D-15-0679.1.

——, S. R. Sobie, and T. Q. Murdock, 2015: Bias correction of GCM precipitation by quantile mapping: How well do methods preserve changes in quantiles and extremes? *J. Climate*, **28**, 6938–6959, https://doi.org/10.1175/JCLI-D-14-00754.1.

Cavazos, T., and S. Arriaga-Ramírez, 2012: Downscaled climate change scenarios for Baja California and the North American monsoon during the twenty-first century. *J. Climate*, **25**, 5904–5915, https://doi.org/10.1175/JCLI-D-11-00425.1.

Chen, J., F. Brissette, D. Chaumont, and M. Braun, 2013a: Finding appropriate bias correction methods in downscaling precipitation for hydrologic impact studies over North America. *Water Resour. Res.*, **49**, 4187–4205, https://doi.org/10.1002/wrcr.20331.

——, ——, ——, and ——, 2013b: Performance and uncertainty evaluation of empirical downscaling methods in quantifying the climate change impacts on hydrology over two North American river basins. *J. Hydrol.*, **479**, 200–214, https://doi.org/10.1016/J.JHYDROL.2012.11.062.

——, C. Li, F. Brissette, H. Chen, M. Wang, and G. R. C. Essou, 2018: Impacts of correcting the inter-variable correlation of climate model outputs on hydrological modeling. *J. Hydrol.*, **560**, 326–341, https://doi.org/10.1016/j.jhydrol.2018.03.040.

——, F. P. Brissette, X. J. Zhang, H. Chen, S. Guo, and Y. Zhao, 2019: Bias correcting climate model multi-member ensembles to assess climate change impacts on hydrology. *Climatic Change*, **153**, 361–377, https://doi.org/10.1007/s10584-019-02393-x.

Chen, L., and O. W. Frauenfeld, 2014: A comprehensive evaluation of precipitation simulations over China based on CMIP5 multimodel ensemble projections. *J. Geophys. Res. Atmos.*, **119**, 5767–5786, https://doi.org/10.1002/2013JD021190.

Chen, W., Z. Jiang, and L. Li, 2011: Probabilistic projections of climate change over China under the SRES A1B scenario using 28 AOGCMs. *J. Climate*, **24**, 4741–4756, https://doi.org/10.1175/2011JCLI4102.1.

Choi, S. W., and J. S. An, 2010: Altitudinal distribution of moths (Lepidoptera) in Mt. Jirisan National Park, South Korea. *Eur. J. Entomol.*, **107**, 229–245, https://doi.org/10.14411/eje.2010.031.

Covey, C., K. M. AchutaRao, U. Cubasch, P. Jones, S. J. Lambert, M. E. Mann, T. J. Phillips, and K. E. Taylor, 2003: An overview of results from the coupled model intercomparison project. *Global Planet. Change*, **37**, 103–133, https://doi.org/10.1016/S0921-8181(02)00193-5.

Crochemore, L., M.-H. Ramos, and F. Pappenberger, 2016: Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts. *Hydrol. Earth Syst. Sci.*, **20**, 3601–3618, https://doi.org/10.5194/hess-20-3601-2016.

Dong, S., Y. Xu, B. Zhou, and Y. Shi, 2015: Assessment of indices of temperature extremes simulated by multiple CMIP5 models over China. *Adv. Atmos. Sci.*, **32**, 1077–1091, https://doi.org/10.1007/s00376-015-4152-5.

Fan, L., D. Chen, C. Fu, and Z. Yan, 2013: Statistical downscaling of summer temperature extremes in northern China. *Adv. Atmos. Sci.*, **30**, 1085–1095, https://doi.org/10.1007/s00376-012-2057-0.

Fowler, H. J., C. G. Kilsby, and J. Stunell, 2007: Modelling the impacts of projected future climate change on water resources in north-west England. *Hydrol. Earth Syst. Sci.*, **11**, 1115–1126, https://doi.org/10.5194/hess-11-1115-2007.

Giorgi, F., and L. O. Mearns, 2002: Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the "Reliability Ensemble Averaging" (REA) method. *J. Climate*, **15**, 1141–1158, https://doi.org/10.1175/1520-0442(2002)015<1141:COAURA>2.0.CO;2.

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

Gong, H., L. Wang, W. Chen, R. Wu, K. Wei, and X. Cui, 2014: The climatology and interannual variability of the East Asian winter monsoon in CMIP5 models. *J. Climate*, **27**, 1659–1678, https://doi.org/10.1175/JCLI-D-13-00039.1.

Gu, H., Z. Yu, J. Wang, G. Wang, T. Yang, Q. Ju, C. Yang, F. Xu, and C. Fan, 2015: Assessing CMIP5 general circulation model simulations of precipitation and temperature over China. *Int. J. Climatol.*, **35**, 2431–2440, https://doi.org/10.1002/joc.4152.

Guilyardi, E., A. Wittenberg, A. Fedorov, M. Collins, C. Wang, A. Capotondi, G. J. van Oldenborgh, and T. Stockdale, 2009: Understanding El Niño in ocean–atmosphere general circulation models: Progress and challenges. *Bull. Amer. Meteor. Soc.*, **90**, 325–340, https://doi.org/10.1175/2008BAMS2387.1.

Guo, Q., J. Chen, X. Zhang, M. Shen, H. Chen, and S. Guo, 2019: A new two-stage multivariate quantile mapping method for bias correcting climate model outputs. *Climate Dyn.*, **53**, 3603–3623, https://doi.org/10.1007/s00382-019-04729-w.

Hay, L. E., R. L. Wilby, and G. H. Leavesley, 2000: A comparison of delta change and downscaled GCM scenarios for three mountainous basins in the United States. *J. Amer. Water Resour. Assoc.*, **36**, 387–397, https://doi.org/10.1111/j.1752-1688.2000.tb04276.x.

He, X., N. W. Chaney, M. Schleiss, and J. Sheffield, 2016: Spatial downscaling of precipitation using adaptable random forests. *Water Resour. Res.*, **52**, 8217–8237, https://doi.org/10.1002/2016WR019034.

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

Herrera-Estrada, J. E., and J. Sheffield, 2017: Uncertainties in future projections of summer droughts and heat waves over the contiguous United States. *J. Climate*, **30**, 6225–6246, https://doi.org/10.1175/JCLI-D-16-0491.1.

IPCC, 2014: Summary for policymakers. *Climate Change 2014: Impacts, Adaptation, and Vulnerability*, C. B. Field et al., Eds., Cambridge University Press, 1–32.

Jha, B., Z.-Z. Hu, and A. Kumar, 2014: SST and ENSO variability and change simulated in historical experiments of CMIP5 models. *Climate Dyn.*, **42**, 2113–2124, https://doi.org/10.1007/s00382-013-1803-z.

Jiang, M., B. S. Felzer, and D. Sahagian, 2016: Predictability of precipitation over the conterminous U.S. based on the CMIP5 multi-model ensemble. *Sci. Rep.*, **6**, 29962, https://doi.org/10.1038/SREP29962.

Katz, R. W., and M. Ehrendorfer, 2006: Bayesian approach to decision making using ensemble weather forecasts. *Wea. Forecasting*, **21**, 220–231, https://doi.org/10.1175/WAF913.1.

Kim, H.-M., P. J. Webster, and J. A. Curry, 2012: Evaluation of short-term climate change prediction in multi-model CMIP5 decadal hindcasts. *Geophys. Res. Lett.*, **39**, L10701, https://doi.org/10.1029/2012GL051644.

Kirkland, E. J., 2010: *Advanced Computing in Electron Microscopy*. Springer Science & Business Media, 289 pp.

Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl, 2010: Challenges in combining projections from multiple climate models. *J. Climate*, **23**, 2739–2758, https://doi.org/10.1175/2009JCLI3361.1.

Lee, J. Y., and B. Wang, 2014: Future change of global monsoon in the CMIP5. *Climate Dyn.*, **42**, 101–119, https://doi.org/10.1007/s00382-012-1564-0.

Li, H., J. Sheffield, and E. F. Wood, 2010: Bias correction of monthly precipitation and temperature fields from Intergovernmental Panel on Climate Change AR4 models using equidistant quantile matching. *J. Geophys. Res.*, **115**, D10101, https://doi.org/10.1029/2009JD012882.

Li, X., W. Zhou, and Y. D. Chen, 2015: Assessment of regional drought trend and risk over China: A drought climate division perspective. *J. Climate*, **28**, 7025–7037, https://doi.org/10.1175/JCLI-D-14-00403.1.

Liang, L., L. Li, and Q. Liu, 2011: Precipitation variability in northeast China from 1961 to 2008. *J. Hydrol.*, **404**, 67–76, https://doi.org/10.1016/j.jhydrol.2011.04.020.

Ma, M., L. Ren, V. P. Singh, F. Yuan, L. Chen, X. Yang, and Y. Liu, 2016: Hydrologic model-based Palmer indices for drought characterization in the Yellow River basin, China. *Stochastic Environ. Res. Risk Assess.*, **30**, 1401–1420, https://doi.org/10.1007/s00477-015-1136-z.

Maraun, D., and Coauthors, 2010: Precipitation downscaling under climate change: Recent developements to bridge the gap between dynamical models and the end user. *Rev. Geophys.*, **48**, RG3003, https://doi.org/10.1029/2009RG000314.

McSweeney, C. F., R. G. Jones, R. W. Lee, and D. P. Rowell, 2015: Selecting CMIP5 GCMs for downscaling over multiple regions. *Climate Dyn.*, **44**, 3237–3260, https://doi.org/10.1007/s00382-014-2418-8.

Mearns, L. O., F. Giorgi, P. Whetton, D. Pabon, M. Hulme, and M. Lal, 2003: Guidelines for use of climate scenarios developed from regional climate model experiments. TGCIA-IPCC Rep., 38 pp., http://www.ipcc-data.org/guidelines/dgm_no1_v1_10-2003.pdf.

Meehl, G., G. J. Boer, C. Covey, M. Latif, and R. J. Stouffer, 2000: The Coupled Model Intercomparison Project (CMIP). *Bull. Amer. Meteor. Soc.*, **81**, 313–318, https://doi.org/10.1175/1520-0477(2000)081<0313:TCMIPC>2.3.CO;2.

——, C. Covey, B. McAvaney, M. Latif, and R. J. Stouffer, 2005: Overview of the coupled model intercomparison project. *Bull. Amer. Meteor. Soc.*, **86**, 89–93, https://doi.org/10.1175/BAMS-86-1-89.

Miao, C., Q. Duan, Q. Sun, and J. Li, 2013: Evaluation and application of Bayesian multi-model estimation in temperature simulations. *Prog. Phys. Geogr.*, **37**, 727–744, https://doi.org/10.1177/0309133313494961.

Moss, R. H., and Coauthors, 2010: The next generation of scenarios for climate change research and assessment. *Nature*, **463**, 747–756, https://doi.org/10.1038/nature08823.

Mote, P., L. Brekke, P. B. Duffy, and E. Maurer, 2011: Guidelines for constructing climate scenarios. *Eos, Trans. Amer. Geophys. Union*, **92**, 257–258, https://doi.org/10.1029/2011EO310001.

Nohara, D., A. Kitoh, M. Hosaka, and T. Oki, 2006: Impact of climate change on river discharge projected by multimodel ensemble. *J. Hydrometeor.*, **7**, 1076–1089, https://doi.org/10.1175/JHM531.1.

Pascoe, C., B. N. Lawrence, E. Guilyardi, M. Juckes and K. E. Taylor, 2019: Designing and documenting experiments in CMIP6. *Geosci. Model Dev. Discuss.*, https://doi.org/10.5194/GMD-2019-98.

Pour, S. H., S. Shahid, E. S. Chung, and X. J. Wang, 2018: Model output statistics downscaling using support vector machine for the projection of spatial and temporal changes in rainfall of Bangladesh. *Atmos. Res.*, **213**, 149–162, https://doi.org/10.1016/j.atmosres.2018.06.006.

Piani, C., and J. O. Haerter, 2012: Two dimensional bias correction of temperature and precipitation copulas in climate models. *Geophys. Res. Lett.*, **39**, L20401, https://doi.org/10.1029/2012GL053839.

——, ——, and E. Coppola, 2010: Statistical bias correction for daily precipitation in regional climate models over Europe. *Theor. Appl. Climatol.*, **99**, 187–192, https://doi.org/10.1007/s00704-009-0134-9.

Pierce, D. W., T. P. Barnett, B. D. Santer, and P. J. Gleckler, 2009: Selecting global climate models for regional climate change studies. *Proc. Natl. Acad. Sci. USA*, **106**, 8441–8446, https://doi.org/10.1073/pnas.0900094106.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174, https://doi.org/10.1175/MWR2906.1.

Rajczak, J., S. Kotlarski, and C. Schär, 2016: Does quantile mapping of simulated precipitation correct for biases in transition probabilities and spell lengths? *J. Climate*, **29**, 1605–1615, https://doi.org/10.1175/JCLI-D-15-0162.1.

Reichler, T., and J. Kim, 2008: How well do coupled models simulate today's climate? *Bull. Amer. Meteor. Soc.*, **89**, 303–312, https://doi.org/10.1175/BAMS-89-3-303.

Reifen, C., and R. Toumi, 2009: Climate projections: Past performance no guarantee of future skill? *Geophys. Res. Lett.*, **36**, L13704, https://doi.org/10.1029/2009GL038082.

Salman, S. A., S. Shahid, T. Ismail, K. Ahmed, and X. J. Wang, 2018: Selection of climate models for projection of spatiotemporal changes in temperature of Iraq with uncertainties. *Atmos. Res.*, **213**, 509–522, https://doi.org/10.1016/j.atmosres.2018.07.008.

Sanderson, B. M., R. Knutti, and P. Caldwell, 2015: A representative democracy to reduce interdependency in a multimodel ensemble. *J. Climate*, **28**, 5171–5194, https://doi.org/10.1175/JCLI-D-14-00362.1.

Sen Roy, S., and R. C. Balling, 2004: Trends in extreme daily precipitation indices in India. *Int. J. Climatol.*, **24**, 457–466, https://doi.org/10.1002/joc.995.

Sheffield, J., and Coauthors, 2013a: North American climate in CMIP5 experiments. Part I: Evaluation of historical simulations of continental and regional climatology. *J. Climate*, **26**, 9209–9245, https://doi.org/10.1175/JCLI-D-12-00592.1.

——, and Coauthors, 2013b: North American climate in CMIP5 experiments. Part II: Evaluation of historical simulations of intraseasonal to decadal variability. *J. Climate*, **26**, 9247–9290, https://doi.org/10.1175/JCLI-D-12-00593.1.

Song, Y., F. Qiao, Z. Song, and C. Jiang, 2013: Water vapor transport and cross-equatorial flow over the Asian-Australia monsoon region simulated by CMIP5 climate models. *Adv. Atmos. Sci.*, **30**, 726–738, https://doi.org/10.1007/s00376-012-2148-y.

Sperber, K. R., H. Annamalai, I. S. Kang, A. Kitoh, A. Moise, A. Turner, B. Wang, and T. Zhou, 2013: The Asian summer monsoon: An intercomparison of CMIP5 vs. CMIP3 simulations of the late 20th century. *Climate Dyn.*, **41**, 2711–2744, https://doi.org/10.1007/s00382-012-1607-6.

Stephens, G. L., and T. D. Ellis, 2008: Controls of global-mean precipitation increases in global warming GCM experiments. *J. Climate*, **21**, 6141–6155, https://doi.org/10.1175/2008JCLI2144.1.

Stott, P. A., J. A. Kettleborough, and M. R. Allen, 2006: Uncertainty in continental-scale temperature predictions. *Geophys. Res. Lett.*, **33**, L02708, https://doi.org/10.1029/2005GL024423.

Su, H. and J. D. Neelin, 2003: The scatter in tropical average precipitation anomalies. *J. Climate*, **16**, 3966–3977, https://doi.org/10.1175/1520-0442(2003)016<3966:TSITAP>2.0.CO;2.

Sui, Y., D. Jiang, and Z. Tian, 2013: Latest update of the climatology and changes in the seasonal distribution of precipitation over China. *Theor. Appl. Climatol.*, **113**, 599–610, https://doi.org/10.1007/s00704-012-0810-z.

Taylor, K. E., 2001: Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.*, **106**, 7183–7192, https://doi.org/10.1029/2000JD900719.

——, R. J. Stouffer, and G. A. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.*, **93**, 485–498, https://doi.org/10.1175/BAMS-D-11-00094.1.

Tebaldi, C., and R. Knutti, 2007: The use of the multi-model ensemble in probabilistic climate projections. *Philos. Trans. Roy Soc. London*, **A365**, 2053–2075, https://doi.org/10.1098/rsta.2007.2076.

Teutschbein, C., and J. Seibert, 2012: Bias correction of regional climate model simulations for hydrological climate change impact studies: Review and evaluation of different methods. *J. Hydrol.*, **456-457**, 12–29, https://doi.org/10.1016/j.jhydrol.2012.05.052.

——, and ——, 2013: Is bias correction of regional climate model (RCM) simulations possible for nonstationary conditions? *Hydrol. Earth Syst. Sci.*, **17**, 5061–5077, https://doi.org/10.5194/hess-17-5061-2013.

Thober, S., and L. Samaniego, 2014: Robust ensemble selection by multivariate evaluation of extreme precipitation and temperature characteristics. *J. Geophys. Res. Atmos.*, **119**, 594–613, https://doi.org/10.1002/2013JD020505.

Toh, Y. Y., A. G. Turner, S. J. Johnson, and C. E. Holloway, 2018: Maritime Continent seasonal climate biases in AMIP experiments of the CMIP5 multimodel ensemble. *Climate Dyn.*, **50**, 777–800, https://doi.org/10.1007/s00382-017-3641-x.

Wang, Y., and L. Zhou, 2005: Observed trends in extreme precipitation events in China during 1961–2001 and the associated changes in large-scale circulation. *Geophys. Res. Lett.*, **32**, L09707, https://doi.org/10.1029/2005GL023769.

Weiland, F. S., L. P. H. Beek, A. Weerts, and M. F. P. Bierkens, 2012: Extracting information from an ensemble of GCMs to reliably assess future global runoff change. *J. Hydrol.*, **412–413**, 66–75, https://doi.org/10.1016/J.JHYDROL.2011.03.047.

Wilby, R. L., S. P. Charles, E. Zorita, B. Timbal, P. Whetton, and L. O. Mearns, 2004: Guidelines for use of climate scenarios developed from statistical downscaling methods. TGCIA-IPCC Rep., 27 pp., https://www.ipcc-data.org/guidelines/dgm_no2_v1_09_2004.pdf.

Woldemeskel, F. M., A. Sharma, B. Sivakumar, and R. Mehrotra, 2012: An error estimation method for precipitation and temperature projections for future climates, *J. Geophys. Res.*, **117**, D22104, https://doi.org/10.1029/2012JD018062.

Wood, A. W., E. P. Maurer, A. Kumar, and E. P. Lettenmaier, 2002: Longrange experimental hydrologic forecasting for the eastern United States. *J. Geophys. Res.*, **107**, 4429, https://doi.org/10.1029/2001JD000659.

——, L. R. Leung, V. Sridhar, and D. P. Lettenmaier, 2004: Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. *Climatic Change*, **62**, 189–216, https://doi.org/10.1023/B:CLIM.0000013685.99609.9e.

Xu, Y., and C. H. Xu, 2012: Preliminary assessment of simulations of climate changes over China by CMIP5 multi-models. *Atmos. Ocean. Sci. Lett.*, **5**, 527–533, https://doi.org/10.1080/16742834.2012.11447041.

——, X. Gao, and F. Giorgi, 2010: Upgrades to the reliability ensemble averaging method for producing probabilistic climate-change projections. *Climate Res.*, **41**, 61–81, https://doi.org/10.3354/cr00835.

Yang, T., and Coauthors, 2011: Changes of climate extremes in a typical arid zone: Observations and multimodel ensemble projections. *J. Geophys. Res.*, **116**, D19106, https://doi.org/10.1029/2010JD015192.

Yang, X., E. F. Wood, J. Sheffield, L. Ren, M. Zhang, and Y. Wang, 2018: Bias correction of historical and future simulations of precipitation and temperature for China from CMIP5 models. *J. Hydrometeor.*, **19**, 609–623, https://doi.org/10.1175/JHM-D-17-0180.1.

——, X. Yu, Y. Wang, Y. Liu, M. Zhang, L. Ren, F. Yuan, and S. Jiang, 2019: Estimating the response of hydrological regimes to future projections of precipitation and temperature over the upper Yangtze River. *Atmos. Res.*, **230**, 104627, https://doi.org/10.1016/j.atmosres.2019.104627.

Yin, H., M. G. Donat, L. V. Alexander, and Y. Sun, 2015: Multi-dataset comparison of gridded observed temperature and precipitation extremes over China. *Int. J. Climatol.*, **35**, 2809–2827, https://doi.org/10.1002/joc.4174.

Zhai, P., X. Zhang, H. Wan, and X. Pan, 2005: Trends in total precipitation and frequency of daily precipitation extremes over China. *J. Climate*, **18**, 1096–1108, https://doi.org/10.1175/JCLI-3318.1.

Zhang, B., and B. J. Soden, 2019: Constraining climate model projections of regional precipitation change. *Geophys. Res. Lett.*, **46**, 10 522–10 531, https://doi.org/10.1029/2019GL083926.

Zhou, B., Q. H. Wen, Y. Xu, L. Song, and X. Zhang, 2014: Projected changes in temperature and precipitation extremes in China by the CMIP5 multimodel ensembles. *J. Climate*, **27**, 6591–6611, https://doi.org/10.1175/JCLI-D-13-00761.1.

Zhou, Y., and G. Ren, 2011: Change in extreme temperature event frequency over mainland China, 1961-2008. *Climate Res.*, **50**, 125–139, https://doi.org/10.3354/cr01053.

Zou, L., and T. Zhou, 2015: Asian summer monsoon onset in simulations and CMIP5 projections using four Chinese climate models. *Adv. Atmos. Sci.*, **32**, 794–806, https://doi.org/10.1007/s00376-014-4053-z.