Research Article

An Assessment of GCM Performance at a Regional Scale Using a Score-Based Method

Fangxin Shi ¹, Zhihui Wang ¹, Liang Qi,² and Rongxu Chen¹

¹Yellow River Institute of Hydraulic Research, Zhengzhou 450003, China

²The Office of Yellow River Flood Control and Drought Relief Headquarters, Zhengzhou 450003, China

Correspondence should be addressed to Zhihui Wang; wangzh@radi.ac.cn

Received 19 April 2018; Revised 5 August 2018; Accepted 19 September 2018; Published 8 November 2018

Academic Editor: Pedro Salvador

Copyright © 2018 Fangxin Shi et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

A multicriteria score-based method was developed to assess the performances of 18 general circulation models (GCMs) in the study region from 1970 to 2005. The results indicate the following. (1) GCMs simulate temperature better than rainfall. The temporal and spatial distributions of simulated temperature performed well compared with those from the observations. In comparison to temperature, the spatial distribution of simulated precipitation performed poorly. Most of the GCMs underestimated temperature and overestimated precipitation. (2) The Grubbs test was used to detect anomalous moving changes in the rank score (RS) results; the *inm-cm4* and *ipsl-cm5b-lr* models were rejected when simulating temperature, while the *bnu-esm* and *canesm2* models performed poorly when simulating precipitation. (3) Adding or removing any criterion does not significantly influence the RS result, which indicates that the multicriteria score-based method is robust. The advantages of using multicriteria score-based method to assess GCMs performance were demonstrated, and this method also provides a more comprehensive assessment when compared with the single-criterion method. The multicriteria method could replace other criteria as the research requirements and could be easily extended to different study regions; the results could be used for better informed regional climate change impact analyses.

1. Introduction

General circulation models (GCMs) are the most common tools for projecting future climate change. Errors and uncertainties in GCM metadata range in severity, specifically resulting in the inability to simulate observed meteorological events. GCM simulations are often characterized by biases and uncertainties that limit their direct application [1]. Different forcing scenarios, GCMs, and subgrid-scale forcings and processes cause uncertainties, revealing an abundance of information but also indicate that a large amount of work is required to identify useful information, which limits GCM applications [2]. Despite continuous efforts to improve the GCM simulation performance, the application of assessment methods is essential for climate change impact studies [3].

To improve the accuracy of GCM applications, GCMs have been assessed in many studies [4–6]. These assessments emphasize various aspects of GCMs according to their different applications. For example, in one study, where a long-term climate change analysis was the main focus, an

assessment of the GCM performance before its application only focused on its long-term temporal and spatial distribution simulations. However, the drawback of this assessment was that using a single criterion could only describe the temporal or spatial performances of the GCMs but may not meet the other requirements of the study [4]. A more comprehensive understanding of the advantages and disadvantages of GCMs is possible when more criteria are included in a GCM assessment.

To date, no assessment method used in the study of GCMs has been widely accepted. Assessing the performance of GCMs before using them is becoming an interesting issue. In this paper, a multicriteria score-based method was analyzed and the performances of all GCMs were quantitatively calculated and examined. We studied this method with the aim of comprehensively and accurately evaluating the performances of the GCMs.

The outline of the paper is as follows: the data and methods are presented in Sections 2 and 3, respectively. Section 4 describes the performance of each GCM. The GCM

simulations of temperature and precipitation are evaluated in the study region. The concluding remarks are provided in Section 5.

2. Study Region and Dataset

The performances of the GCMs in the Yellow-Huai-Hai region were assessed in this study. The Yellow-Huai-Hai region, which is located in north-central China between 30° and 42.5°N and 90° and 122.5°E (Figure 1), has the largest fluvial plain in China. Most parts of the study area are semiarid and semihumid (i.e., the Yellow River and Hai River basins, respectively), and only a small part of the region in the southeast of the study area has a humid climate (the area covering the Huai River basin). The Yellow-Huai-Hai region is an agricultural breadbasket and prime urban and industrial region in China. This region, therefore, plays an important role in the social and economic development of the country. Thus, the consequences of climate change seriously restrain economic growth [7, 8].

All GCM data are from the fifth phase of the Coupled Model Intercomparison Project (CMIP5), which is the most important tool for analyzing future climate change. The dataset from this project provides a framework for coordinating climate change experiments with the aim of evaluating climate simulations of the recent past to provide more accurate projections of climate change and quantifications of climate feedbacks compared to those from CMIP3 [9]. Details on the data can be found in Moss et al. [10] and Taylor et al. [9]. To simulate future climate change, 18 global GCMs from CMIP5 were considered in this paper (including access1-0, bcc-csm1, bnu-esm, canesm2, ccs4, cesm1-bgc, cnrm-cm5, giss-e2-h, csiro-mk3.6, fgoals-g2, gfdl-cm3, hadgem2, inm-cm4, noresm1-m, miroc-esm, ipsl-cm5b-lr, mri-cgcm3). More details on all of the models are available at http://cmip-pcmdi.llnl.gov/cmip5/docs/CMIP5_modeling_groups.docx. Since GCM horizontal resolutions vary, the GCM outputs were interpolated to a uniform resolution of 2.5° × 2.5°. The grid cell distributions over the study region are shown in Figure 1.

High-quality temperature and precipitation data were derived from the daily dataset on China's surface climate (V3.0) during the period 1970–2005 provided by the National Meteorological Information Center. These data are based on gauged data from 128 meteorological stations (Figure 1) and have been controlled for quality and accuracy by nearly 100%; for more details, see http://data.cma.cn/data/cdcdetail/dataCode/SURF_CLI_CHN_MUL_DAY_V3.0.html. To effectively assess the performances of the GCMs, the daily data observed by the meteorological stations were collected as monthly data and interpolated to 2.5° × 2.5° cells using the inverse distance weighted method. The hollow circle in Figure 1 represents the location of each GCM grid point, and the data for the GCM grid points, which are denoted with black circles, were selected for the assessment in this study.

3. Methods

In this study, a multicriteria score-based method was developed to assess the performances of GCM simulations at a regional scale. The criteria included mean annual data, standard deviation, annual climate cycle, normalized root mean square error (NRMSE), spatial distribution, climate change trend, empirical orthogonal function (EOF), and probability density function (PDF); these criteria are listed in Table 1.

In the assessment, the rank score (RS) values of 0–9, which are used to assess each individual assessment criterion, are written in the following form:

$$RS_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \times 9, \quad (1)$$

where x_i represents the relative error (RE) between the i th GCM result and observation or the related statistical value for the i th GCM. For the RE, a larger x_i indicates a larger RS in the GCM performance assessment. In addition, the total RS for each GCM is summated by the RS for all weighted criteria. This RS method is used to describe the fitting degree between observed and simulated sequential statistical characteristics. According to the fitting results, the score of each GCM was assigned a number between 0 and 9 to assess the performance of each GCM. The RS does not represent the actual simulation accuracy of the specific models but is suitable for comparison between different GCM performances. Several different criteria that have the same statistical purposes, such as the Mann-Kendall (M-K) test (Z) and trend magnitude (β), which are criteria for trend analyses; EOF1 and EOF2, which are criteria for EOF analyses; and Brier score (BS) and significance score (Sscore), which are criteria for PDFs (which will be described later), have weights of 0.5 each during this summation (Table 1), while the other individual criteria have weights of 1.0. If a GCM effectively simulates an observation, then the RS is small.

The RE was used to quantify the similarity between simulated and observed values for long-term monthly means and standard deviations:

$$RE = \frac{(1/n) \sum_{i=1}^n X_{Gi} - (1/n) \sum_{i=1}^n X_{Oi}}{(1/n) \sum_{i=1}^n X_{Oi}}, \quad (2)$$

where X_{Gi} and X_{Oi} represent the simulated and observed data of the time series, respectively, and n represents the duration of these samples (432 months from 1970 to 2005).

The GCM performances for time series are evaluated by the NRMSE [11, 12] and defined as

$$NRMSE = \frac{\sqrt{(1/n) \sum_{i=1}^n (X_{Gi} - X_{Oi})^2}}{\sqrt{(1/n - 1) \sum_{i=1}^n (X_{Oi} - \bar{X}_O)^2}}, \quad (3)$$

Based on historical data, X_{Gi} and X_{Oi} represent the GCM and observation results at historical time i , respectively; $\bar{X}_O$ represents the mean of the observations; and n represents the length of the time series. The advantage of the NRMSE is that it can consider the mean and standard deviation of the predictor. The NRMSE is essentially the root mean square error divided by the standard deviation in the corresponding

FIGURE 1: Locations of meteorological stations and selected GCM grid points (solid circles) in the study region.

TABLE 1: Statistics of climate variables and their weights.

Statistics of climate variables	Criteria	Weights
Mean	RE (relative error) (%)	1
Standard deviation	RE (%)	1
Temporal change	NRMSE (normalized root mean square error)	1
Monthly distribution	Correlation coefficient	1
Spatial distribution	Correlation coefficient	1
Trend and its magnitude	Mann-Kendall (M-K) test (Z)	0.5
	Sen's slope (β)	0.5
Spatiotemporal variability	EOF1 (first vector)	0.5
	EOF2 (second vector)	0.5
Probability density functions (PDFs)	BS (Brier score)	0.5
	Sscore (significance score)	0.5

observations. The lowest value of the NRMSE is always associated with the best results, and this lowest value is reliable for determining the best simulation. The range of the NRMSE varies from 0 to positive infinity, where 0 indicates that there is a perfect agreement between the GCM data and reference data.

The correlation coefficient of the annual cycle was calculated between the observed and modeled long-term monthly mean values. For the spatial distribution, the correlation coefficient was calculated between the observed and modeled long-term means for each individual grid cell.

The M-K test and trend magnitude method were applied to determine the long-term monotonic annual trends and quantify their magnitudes [13]. The rank-based value of the nonparametric M-K test statistic (Z) for climate variables in the GCMs and observations was estimated by

$$Z = \begin{cases} \frac{s-1}{(\text{var}(s))^{1/2}}, & \text{if } S > 0, \\ 0, & \text{if } S = 0, \\ \frac{s+1}{(\text{var}(s))^{1/2}}, & \text{if } S < 0, \end{cases} \quad (4)$$

where

$$\text{var}(s) = \frac{\{n(n-1)(2n+5) - \sum_t t(t-1)(2t+5)\}}{18},$$

$$S = \sum_{k=1}^{n-1} \sum_{j=k+1}^n \text{sgn}(x_j - x_k),$$

$$\text{sgn}(\theta) = \begin{cases} 1, & \text{if } \theta > 0, \\ 0, & \text{if } \theta = 0, \\ -1, & \text{if } \theta < 0, \end{cases} \quad (5)$$

where x represents the time series of the annual climate variable, n represents the length of year, t represents the extent of any given tie (length of consecutive equal values), and $\sum_t$ denotes the summation over all ties.

The trend magnitude β , for Sen's slope, which is a metric developed by Hirsch et al. [13] and proposed by Sen [14]; is defined as

$$\beta = \text{median} \frac{(X_j - X_i)}{(j - i)}, \quad (6)$$

where $1 < i < j < n$. The slope estimator, β , is equal to the median for all possible combinations of pairs for the whole dataset [7]. X represents the time series for the variable to be assessed in the study. Sen's slope analyzes the change trend of the data by analyzing the time series data to possibly avoid the adverse influence of lost data in the analysis. The RE in Equation (2) was used to assess how close the values of Z and β are for each GCM to the observed values.

An EOF analysis was used in this study to compare the spatial distribution differences between the modeled climate variables and observations [15]. An EOF can identify and quantify the spatial structures of correlated variabilities [16]. The two leading modes are selected in this assessment since they account for a majority of the total variance.

The BS and Sscore were used to assess the PDFs of the monthly climate variables in the GCMs.

$$\begin{aligned} \text{BS} &= \frac{1}{n} \sum_{i=1}^n (P_{mi} - P_{oi})^2, \\ \text{Sscore} &= \sum_{i=1}^n \text{minimum}(P_{mi}, P_{oi}), \end{aligned} \quad (7)$$

where P_{mi} and P_{oi} represent the simulated and observed i th probability values, respectively, in each bin and n represents the number of bins. According to the data ranges, we set the number of bins as 100; thus, we divided all of the data into 100 equal parts sequentially and then calculated the probability density of each size. In this study, the BS represents the mean square error measure for probability forecasts [17, 18] and the Sscore represents the calculated cumulative minimum value of the observed and simulated distributions for each bin, which can quantify the overlap between the observed and simulated data [19, 20]. Therefore, when the BS of a GCM is lower and the Sscore is higher, the performance of the GCM is better.

4. Results

4.1. Assessment of Temperature. Table 2 includes the assessment results of the GCM performance for temperature in the Yellow-Huai-Hai region. The observed mean temperature during the historical period in the study region is 8.49°C, while the simulated temperature via the GCMs is 3.62–8.09°C for the same period. Most GCMs underestimate the mean temperature by approximately 2°C. The standard deviation in the observations is 0.53°C, and most standard deviations in the GCMs are from 0.4–0.6°C. The NRMSE is always used to compare the difference between the observations and simulations. Therefore, if there were sets of data with very similar results for the means and standard deviations, smaller NRMSE results indicated a better simulation for the set of data. For the monthly mean temperature, the best NRMSE occurs with the mpi-esm-lr GCM (0.16), while the NRMSE result for the ipsl-cm5b-lr model was the largest of the GCMs. The simulated monthly distribution for the annual climate cycle for each GCM was relatively similar to that from the observed data, which can be seen from the correlation index (all values are larger than 0.995). Consequently, the correlation results for the monthly distribution of the annual cycle were rounded to 1. The correlation coefficients of the spatial temperature distribution between each GCM and the observations were also larger than 0.9. The simulated spatial temperature had a distribution similar to that of the observations, where the temperature increased from west to east, and the temperature was lowest in the source region of the Yellow River, while it was highest in the southern region of the Huai River basin (Figure 2).

According to the results of the M-K analysis in Table 2, the temperature in the Yellow-Huai-Hai region has increased over the past 36 years. Most GCMs show an increasing trend in temperature, excluding the giss-e2-h model. The performances of the different GCMs in

simulating the change trend differ. The Z value in the M-K test for observed temperature is 4.81, which means that the observed mean temperature significantly increases at the 0.05 significance level. However, the Z values for most GCMs are between 1.13 and 4.59 (excluding giss-e2-h and canesm2), which indicates that most GCMs underestimate the temperature change trend in this region. The trend magnitude, β , via Sen's slope shows similar results.

The results of the analysis of spatial temperature by using an EOF show that the first and second vectors of the EOF for monthly temperature via the observations account for 98.9% and 0.51% of the total variance (Table 2), respectively. The range of the first explained variance of GCMs is between 96.99% and 98.63, while that of the second explained variance is between 0.55% and 1.23%. This result simply evaluates the GCMs' performance by using two explained variance values and indicates that all GCMs simulate the variability well. According to the EOF results of GCMs, all the GCMs perform well in terms of the physical process of temperature variability. It should be noted that there are certain special cases where the spatial patterns could differ, while the spatial patterns and observations have similar values of variance. However, this situation is relatively rare and is, therefore, not discussed in this study.

The empirical cumulative probability distribution (Figure 3) shows that the empirical cumulative probability distributions for monthly mean temperatures that are simulated by most GCMs are quite close to the observations (excluding the inm-cm4 and ipsl-cm5b-lr models, which underestimate the ensemble temperature in the Yellow-Huai-Hai region). The results of the Sscore and BS across all 53 selected grid points are presented in Figure 4. The variations in the scores across all 53 grid points imply spatial differences. A high Sscore with a relatively low BS indicates excellent GCM performance in terms of probability distributions in the grid points. The mean Sscores in the grid points of most GCMs are over 80%. The results of the ipsl-cm5b-lr model are consistent with the empirical cumulative probability plots, which have larger BS and smaller Sscore values. The BS and Sscore of each model behave differently between each grid point, reflecting the spatial variability of climatic elements. For example, in some GCMs, the Sscore in some grid points is over 90%, and the BS value is close to 0, which means that the probability density distribution of the GCM temperature in these grid points is very similar to that of the observations. However, the Sscore does not exceed 50% when there is a high BS value in some grid points, which indicates that the probability distribution of temperature simulated by these GCMs is not quite as strong in these grid points. By using the RS assessment, the performances of the GCMs have been evaluated and the final score of each model has been calculated. The ccsm4 model has the highest score, while the inm-cm4 model has the lowest score. Figure 5 describes the differences in annual temperature changes between the observations and the best and worst performing models in the Yellow-Huai-Hai region. We can clearly see that even though the ccsm4 model has underestimated the mean temperature, the model simulates a change trend similar to that of the observations. In contrast, the inm-cm4

TABLE 2: Model performances for monthly mean temperature.

GCMs	GCM number	Mean value (°C)	StDev	NRMSE	Monthly corr.	Spatial corr.	Zc	M-K Slope (°C/10 a)	EOF1	EOF2	PDF BS	Sscore	Total RS
obs		8.49	0.53				4.81	0.37	98.90	0.51			
access1-0	1	6.81	0.47	0.23	0.999	0.95	3.64	0.28	97.46	0.74	0.006	89.22	16.89
bcc-csm1	2	5.67	0.38	0.33	0.999	0.95	4.40	0.23	97.66	1.19	0.011	80.84	33.22
bnu-esm	3	6.63	0.62	0.27	0.997	0.94	3.69	0.36	98.26	0.64	0.005	91.50	17.68
canesm2	4	7.77	0.51	0.18	0.998	0.93	5.38	0.35	97.66	1.15	0.002	91.19	17.99
ccsm4	5	6.65	0.54	0.27	0.999	0.96	2.98	0.27	98.40	0.63	0.006	89.16	13.73
cesm1-bgc	6	6.79	0.45	0.25	0.998	0.96	2.55	0.20	98.43	0.72	0.005	91.13	16.04
cnrm-cm5	7	5.07	0.46	0.41	0.996	0.95	1.38	0.10	98.03	0.76	0.011	83.78	31.09
giss-e2-h	8	7.12	0.42	0.25	0.995	0.94	-0.10	-0.01	97.95	0.99	0.006	83.48	32.58
csiro-mk3.6	9	6.76	0.42	0.26	0.997	0.93	3.69	0.25	98.52	0.68	0.006	88.92	22.95
fgoals-g2	10	7.49	0.58	0.19	0.996	0.95	4.59	0.44	97.73	1.19	0.004	86.06	17.73
gfdl-cm3	11	5.33	0.61	0.36	0.996	0.97	2.96	0.27	97.74	0.94	0.012	77.21	26.83
hadgem2	12	7.20	0.55	0.26	0.999	0.96	3.17	0.23	97.84	0.64	0.005	85.45	14.78
inm-cm4	13	3.90	0.47	0.53	0.998	0.93	1.13	0.09	96.99	1.08	0.017	78.21	49.67
noresm1-m	14	5.58	0.47	0.35	0.998	0.95	4.24	0.31	98.35	0.75	0.011	83.65	23.11
miroc-esm	15	7.33	0.29	0.22	0.997	0.95	2.60	0.13	97.80	1.23	0.004	90.79	29.35
mpi-esm-lr	16	8.09	0.36	0.16	0.999	0.96	2.90	0.19	98.06	0.93	0.001	91.01	15.53
ipsl-cm5b-lr	17	3.62	0.47	0.55	0.997	0.93	2.17	0.15	98.47	0.59	0.019	72.67	42.89
Mri-cgcm3	18	6.26	0.41	0.32	0.997	0.96	2.03	0.13	98.63	0.55	0.079	85.90	21.19

EOFs 1 and 2 are percentages of the explained variance for the first two leading modes of each EOF; PDF is the probability density function.

FIGURE 2: Spatial distribution of annual temperature: (a) observations; (b) simulations for all GCMs.

model vastly underestimates the temperature and simulates an incorrect temperature change in comparison to that of the observations.

4.2. Assessment of Precipitation. Table 3 includes the assessment results of the GCM performances for precipitation. In comparison with temperature, the GCMs perform poorly

FIGURE 3: Empirical cumulative probabilities for monthly mean temperature via observations and GCMs.

FIGURE 4: Box plots of BS and Sscore values for the PDF-based performance scores for monthly mean temperatures across all grid points in the Yellow-Huai-Hai basin (unit: %).

FIGURE 5: Annual temperature changes in the Yellow-Huai-Hai region.

in terms of precipitation. The observed mean annual precipitation in the Yellow-Huai-Hai region is 568 mm, while most GCMs overestimate the value of precipitation (650 mm–1256 mm). Specifically, precipitation in the bnu-esm model reaches 1,256 mm, which is two times greater than the observed precipitation amount. The standard deviations in the bnu-esm and mri-cgcm3 are 83.4 mm and 33.58 mm, respectively, which is quite different from the observation (61.5 mm). The NRMSE for precipitation (only 0.54–1.5) is much larger than that for temperature (0.16–0.55). The correlation coefficients for monthly precipitation in the annual cycle via the GCMs are lower than those for temperature, but most correlation coefficient values are still greater than 0.9. However, when we analyze the performances of the GCMs in terms of the precipitation spatial distribution, the spatial correlation coefficients of the GCMs are 0.45–0.82, which indicates that the GCM simulations for the spatial distributions of precipitation, are much worse than those for the spatial distributions of temperature. Figure 6 shows that the bnu-esm model performs poorly in terms of simulating spatial precipitation and, specifically, it incorrectly estimates the high precipitation region in the study region.

The annual precipitation in the Yellow-Huai-Hai region experiences a nonsignificant decrease at the 0.05 significance level. According to the Z value and magnitude of Sen's slope, the change trends for most GCMs decrease less than those in the observations; specifically, some GCMs appear to have increasing trends during the study period. The M-K test shows that precipitation in the GCMs shows different change trends, which indicates that simulated precipitation via the GCMs is much more uncertain than the simulated temperature. Analyzing the EOFs reveals that the difference between the observations and GCMs is larger than that for temperature, which is consistent with other criteria assessment results, where precipitation is relatively poorly simulated compared to temperature. The physical mechanisms affecting precipitation are mainly influenced by large-scale circulation factors; the inconsistent spatial distribution of simulated precipitation indicates that some GCMs could not explain the influence of circulation factors.

The empirical cumulative probability distributions for the GCM monthly precipitation are compared with the

TABLE 3: Model performances for precipitation.

GCMs	GCM number	Mean Value (mm/y)	StDev	NRMSE	Monthly corr.	Spatial corr.	M-K		EOF		PDF		Total RS
							Zc	Slope (mm/10 a)	EOF1	EOF2	BS	Sscore	
obs	0	568.2	61.5				-0.86	-11.76	71.72	8.69			
access1-0	1	775.0	62.3	0.72	0.95	0.82	-0.34	-2.63	53.21	17.70	0.03	91.22	13.22
bcc-csm1	2	955.3	67.5	1.00	0.92	0.61	0.12	2.65	61.31	14.08	0.06	90.13	27.21
bnu-esm	3	1,255.9	83.4	1.50	0.94	0.45	0.12	1.78	60.56	16.00	0.10	81.30	48.52
canesm2	4	914.3	75.5	1.04	0.85	0.72	1.46	21.48	49.79	21.16	0.04	80.24	39.59
ccsm4	5	891.0	65.6	0.92	0.93	0.70	-1.10	-13.43	60.89	13.03	0.03	81.26	18.98
cesm1-bgc	6	888.2	50.7	0.90	0.93	0.70	-2.30	-19.45	58.82	13.27	0.03	81.75	23.99
cnrm-cm5	7	703.6	71.8	0.65	0.96	0.80	1.21	14.66	55.57	16.66	0.03	81.85	19.68
giss-e2-h	8	802.2	77.5	0.80	0.91	0.57	0.23	2.04	64.49	15.18	0.04	93.24	26.15
csiro-mk3.6	9	650.6	51.7	0.54	0.98	0.75	1.08	10.45	62.30	10.64	0.01	89.64	12.24
fgoals-g2	10	881.0	60.3	1.01	0.78	0.81	-0.59	-6.11	46.99	25.29	0.06	84.18	30.79
gfdl-cm3	11	879.7	65.72	0.88	0.92	0.52	-1.13	-16.57	57.97	19.90	0.05	85.45	27.14
hadgem2	12	725.4	64.31	0.70	0.91	0.76	1.16	9.07	49.51	18.78	0.02	82.22	21.85
inm-cm4	13	923.5	62.61	0.90	0.95	0.68	1.30	13.30	62.19	13.17	0.05	89.46	23.00
noresm1-m	14	1,035.7	65.13	1.12	0.94	0.54	-0.80	-5.93	70.20	10.14	0.04	74.48	26.06
miroc-esm	15	898.3	68.81	1.04	0.83	0.57	-0.18	-1.78	65.68	13.81	0.03	76.61	30.46
mpi-esm-lr	16	887.8	63.13	0.95	0.89	0.66	0.45	3.98	55.38	18.05	0.04	89.27	27.01
ipsl-cm5b-lr	17	674.1	66.00	0.71	0.90	0.71	2.39	22.50	49.97	14.25	0.02	73.81	27.14
mri-cgcm3	18	456.1	33.58	0.72	0.93	0.76	-0.15	-1.38	48.21	19.47	0.01	61.28	28.21

EOFs 1 and 2 are percentages of the explained variance for the first two leading modes of each EOF; PDF is the probability density function.

FIGURE 6: Spatial distribution of annual precipitation (a) observations; (b) the bnu-esm model.

observations in Figure 7. The simulation of the empirical cumulative probability distribution for precipitation is not as accurate as that for temperature. Generally, most GCMs simulate a poorer result for high values of precipitation in

the empirical cumulative probability distribution. Most GCMs overestimate high precipitation values in the probability distribution, which is consistent with the results of the annual precipitation analysis. The BS for monthly

FIGURE 7: Empirical cumulative probabilities of monthly mean precipitation from observations and GCMs.

precipitation in the 53 grid points is much larger than that for monthly temperature, and the outliers also indicate an inconsistency between the GCMs and observations (Figure 8). Although the Sscore median values have almost the same magnitudes as those for monthly temperature, the result of a higher BS and lower Sscore also indicates that the temperature simulations in the GCMs are better than the simulations for precipitation (especially the bun-esm model).

The results of the GCM performances by using RS values are shown in Table 3. The csiro-mk3.6 model simulated precipitation better than the other models, and its RS was only 12.24. In addition, the bun-esm model performed the worst in terms of simulating precipitation, with the highest RS (48.52). Figure 9 describes the annual precipitation change in the Yellow-Huai-Hai region. The bun-esm model vastly overestimated annual rainfall in the study region, and the oscillations in the model seem to have a reversed phase compared with those in the observations. Even the csiro-mk3.6 model appears to slightly overestimate the annual precipitation in the study area and exhibits a different fluctuation change compared with the observations in the early 1970s; the model has a fluctuation change similar to that in the observed data after 1975.

Five GCMs, which have a total RS for temperature less than 17 (ccsm4, hadgem2, mpi-esm-lr, cesm1-bgc and access1-0), and 4 GCMs, which have a total RS for precipitation less than 20 (csiro-mk3.6, access1-0, ccsm4, cnrm-cm5, hadgem2, and cesm1-bgc), were chosen as good GCM groups (Figure 10). Compared with the observations, the good GCM groups are narrower in terms of their uncertainty

FIGURE 8: Box plots of BS and Sscore values for the PDF-based performance scores for monthly precipitation across all grid points in the Yellow-Huai-Hai basin (unit: %).

FIGURE 9: Annual precipitation changes in the Yellow-Huai-Hai region.

intervals, and the mean values are closer to those of the observations. Note that the errors in the GCM metadata affect the entire intensity spectrum, and bias correction is required to improve the GCM simulation capacity. After a simple bias correction, the models in the good GCM groups could be effectively applied in future studies.

4.3. Overall Performance of the GCMs and Sensitivity Analysis.

The RSs for temperature and precipitation are used for assessing the performances of all GCMs (the last columns in Tables 2 and 3). In ascending order, the difference between two successive ranking scores (i.e., the moving range (MR)) is used to detect the presence of any change points [21–23]. In addition, the Grubbs test [24] is used to test whether there is

FIGURE 10: Comparison between the simulation results for all GCMs and those in the good GCM groups (the black solid circles represent the mean value of the group).

an anomalous value in a univariate dataset. If the tests indicate significant differences, then we have evidence to reject the GCMs within the larger ranking score group. The results of the MR analysis and Grubbs test are shown in Table 4.

Two change points in the temperature were detected, while the result of the Grubbs test indicates that these change points are outliers at the 95% significance level (Figure 11). Thus, these GCMs (inm-cm4 and ipsl-cm5b-lr) should be rejected because their RSs are significantly different from those of the other GCMs. The differences in the first 8 GCMs are not significant. Some GCMs with high-ranking scores could not be rejected by the test because these GCMs could capture one or more characteristics in the temporal or spatial distributions for monthly temperature.

For precipitation, two change points were detected (Figure 12). The result of the Grubbs test indicates that the bnu-esm and canesm2 models should be rejected due to their poor performances in simulating precipitation. The RSs of the last two GCMs are remarkably different compared to those for the other GCMs, while the differences in the GCM RSs among the other models are small.

According to Table 4, the simulations in the GCMs for temperature perform better than those for precipitation. This result is consistent with the study of global-scale AR4 GCMs, which revealed that most GCMs could capture the characteristics of monthly temperature but not those of precipitation [12]. We should note that the same GCM model performs differently for different climate variables. For example, the bnu-esm model is the 6th best model for temperature, but it is the worst model for precipitation. In addition, the csiro-mk3.6 model simulates precipitation the best, while it ranks only 10th in the RS assessment.

In addition, the GCMs perform differently in different regions. For example, the bnu-esm model is not suitable for projecting future climate changes in the Yellow-Huai-Hai

TABLE 4: The scores of the GCMs for monthly temperature and precipitation.

	ID	Temperature	Precipitation	Total
access1-0	1	16.89	13.22	30.11
bcc-csm1	2	33.22	27.21	60.43
bnu-esm	3	17.68	48.52*	66.21
canesm2	4	17.99	39.59*	57.57
ccsm4	5	13.73	18.98	32.71
cesm1-bgc	6	16.04	23.99	40.03
cnrm-cm5	7	31.09	19.68	50.77
giss-e2-h	8	32.58	26.15	58.73
csiro-mk3.6	9	22.95	12.24	35.19
fgoals-g2	10	17.73	30.79	48.52
gfdl-cm3	11	26.83	27.14	53.97
hadgem2	12	14.78	21.85	36.63
inm-cm4	13	49.67*	23.00	72.67
noresm1-m	14	23.11	26.06	49.17
miroc-esm	15	29.35	30.46	59.81
mpi-esm-lr	16	15.53	27.01	42.53
ipsl-cm5b-lr	17	42.89*	27.14	70.03
mri-cgcm3	18	21.19	28.21	49.41

*GCM was statistically rejected.

region, but it may potentially perform well for another study region.

To analyze each individual assessment criterion in the ranking results, the overall results were compared with the results by removing an individual statistical criterion. Based on Figure 13, adding or removing any assessment criterion does not obviously influence the overall ranking. The RS score may change after a criterion is removed, but the better-performing mode still performs well after adding or removing a criterion. The results indicate that this RS method robustly assesses the GCM performances. This robust assessment provides an advantage when using the multicriteria

FIGURE 11: Moving test and breaking points of temperature via GCMs: (a) moving range for temperature and (b) breaking point of ranking scores for temperature.

FIGURE 12: Moving test and breaking points of precipitation via GCMs: (a) moving range for precipitation and (b) breaking point of ranking scores for precipitation.

method to assess the performances of GCMs rather than using an individual assessment criterion because a GCM may simulate individual statistical factors well but not provide good simulations for other factors.

Each of the RSs for the statistics that were produced by a single criterion were used individually and compared with

FIGURE 13: Sensitivity analysis for each individual criterion corresponding to the sums of the temperature and precipitation rankings.

the overall results (Figure 14). According to the correlation analysis, no single criterion produced exactly the same result as the overall ranking, which also affirmed that the multi-criteria method produced more information than the single-criterion assessment. The assessments of a single criterion provided different results, such as the RS of the NRMSE criterion being close to the overall ranking and the correlation coefficient being 0.75, while the correlation coefficient of the spatial distribution was only 0.08. Thus, if there is a GCM that can simulate the spatial and seasonal distribution well in the Yellow-Huai-Hai region, this does not mean this model would also have better results in simulating

FIGURE 14: Correlations between single criterion and overall ranking scores.

other statistics (e.g., long-term means, trend magnitude, or probability density).

5. Conclusion

In this paper, a multicriteria score-based method is developed to assess GCM performance in the Yellow-Huai-Hai region from 1970 to 2005. The RSs of these criteria are applied to comprehensively assess the temporal and spatial performances of 18 GCMs when simulating precipitation and temperature in the study region.

All GCMs perform well when simulating temperature. Although all of the models underestimated the mean temperature, the results of the temporal and spatial distributions are quite close to those from the observations. The GCMs did not simulate precipitation as well as temperature, especially in terms of simulating precipitation spatial distributions. Most GCMs overestimate mean precipitation in the study area. The good performing models are selected to comprise the good GCM groups, where the means of the good GCM groups are closer to the observations.

By analyzing the sensitivity of the criteria, we found that removing or adding a criterion does not obviously influence the results of the assessment, which indicates that the multicriteria score-based method is a robust method for assessing GCMs. This study provides a different method from a single evaluation criterion to assess the GCMs' simulation ability. Researchers could specify the criteria relevant to their specific application and research requirements to select appropriate GCMs for their study. This method could be easily applied to different study regions and guide the selection of GCMs for use in regional climate change impact studies.

Data Availability

The GCMs data used to support the findings of this study have been deposited in the <http://www.pemdi.lnl.gov>.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

We are very grateful to the anonymous reviewers for their constructive comments, which greatly improved the manuscript. We are grateful to the climate modeling groups for producing and making their model outputs available to the public. Financial support from the National Key Research and Development Program (grant 2018YFC0407403), National Natural Science Foundation of China (grants 51809103 and 41701509), Special Research Fund of the Yellow River Institute of Hydraulic Research (grants HKY-JBYW-2018-06 and HKY-JBYW-2017-08), China Postdoctoral Science Foundation (grant 2017M610458), and the Technology Development Foundation of the Yellow River Institute of Hydraulic Research (grant HKF201604) is gratefully acknowledged.

References

- [1] D. Sharma, A. Das Gupta, and M. S. Babel, "Spatial disaggregation of bias-corrected GCM precipitation for improved hydrologic simulation: ping River Basin, Thailand," *Hydrology and Earth System Sciences*, vol. 11, no. 4, pp. 1373–1390, 2007.
- [2] F. Giorgi and R. Francisco, "Evaluating uncertainties in the prediction of regional climate change," *Geophysical Research Letters*, vol. 27, no. 9, pp. 1295–1298, 2000.
- [3] S. Seneviratne, N. Nicholls, D. Easterling et al., "Changes in climate extremes and their impacts on the natural physical environment: an overview of the IPCC SREX report," *EGU General Assembly Conference Abstracts*, vol. 14, p. 12566, 2012.
- [4] G. Fu, Z. Liu, S. P. Charles, Z. Xu, and Z. Yao, "A score—based method for assessing the performance of GCMs: a case study of southeastern Australia," *Journal of Geophysical Research: Atmospheres*, vol. 118, no. 10, pp. 4154–4167, 2013.
- [5] J. A. Marengo, I. F. A. Cavalcanti, P. Satyamurty et al., "Assessment of regional seasonal rainfall predictability using the CPTEC/COLA atmospheric GCM," *Climate Dynamics*, vol. 21, no. 5-6, pp. 459–475, 2003.
- [6] W. Wang and A. Kumar, "A GCM assessment of atmospheric seasonal predictability associated with soil moisture anomalies over North America," *Journal of Geophysical Research: Atmospheres*, vol. 103, no. 22, pp. 28637–28646, 1998.
- [7] G. Fu, S. Chen, C. Liu, and D. Shepard, "Hydro-climatic trends of the Yellow River basin for the last 50 years," *Climatic Change*, vol. 65, no. 1-2, pp. 149–178, 2004.
- [8] Y. Zhang, A. H. Arthington, S. E. Bunn, S. Mackay, J. Xia, and M. Kennard, "Classification of flow regimes for environmental flow assessment in regulated rivers: the Huai River basin, China," *River Research and Applications*, vol. 28, no. 7, pp. 989–1005, 2012.
- [9] K. E. Taylor, R. J. Stouffer, and G. A. Meehl, "An overview of CMIP5 and the experiment design," *Bulletin of the American Meteorological Society*, vol. 93, no. 4, pp. 485–498, 2012.
- [10] R. H. Moss, J. A. Edmonds, K. A. Hibbard et al., "The next generation of scenarios for climate change research and assessment," *Nature*, vol. 463, no. 7282, pp. 747–756, 2010.
- [11] S. R. Hanna and D. W. Heinold, *Development and Application of a Simple Method for Evaluating Air Quality Models*, American Petroleum Institute, Washington, DC, USA, 1985.
- [12] D. A. Randall, R. A. Wood, S. Bony et al., "Climate models and their evaluation," in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth*

- Assessment Report of the IPCC (FAR)*, pp. 589–662, Cambridge University Press, Cambridge, UK, 2007.
- [13] R. M. Hirsch, J. R. Slack, and R. A. Smith, “Techniques of trend analysis for monthly water quality data,” *Water Resources Research*, vol. 18, no. 1, pp. 107–121, 1982.
 - [14] P. K. Sen, “Estimates of the regression coefficient based on Kendall’s tau,” *Journal of the American Statistical Association*, vol. 63, no. 324, pp. 1379–1389, 1968.
 - [15] L. D. D. Harvey and T. M. L. Wigley, “Characterizing and comparing control-run variability of eight coupled AOGCMs and of observations. Part 1: temperature,” *Climate Dynamics*, vol. 21, no. 7–8, pp. 619–646, 2003.
 - [16] Q. Mu, C. S. Jackson, and P. L. Stoffa, “A multivariate empirical-orthogonal-function-based measure of climate model performance,” *Journal of Geophysical Research: Atmospheres*, vol. 109, no. D15, p. 101, 2004.
 - [17] G. W. Brier, “Verification of forecasts expressed in terms of probability,” *Monthly Weather Review*, vol. 78, no. 1, pp. 1–3, 1950.
 - [18] K. Fraedrich and L. M. Leslie, “Evaluation of techniques for the operational, single station, short-term forecasting of rainfall at a midlatitude station (Melbourne),” *Monthly Weather Review*, vol. 115, no. 8, pp. 1645–1654, 1987.
 - [19] S. E. Perkins, A. J. Pitman, N. J. Holbrook, and J. McAneney, “Evaluation of the AR4 climate models’ simulated daily maximum temperature, minimum temperature, and precipitation over Australia using probability density functions,” *Journal of Climate*, vol. 20, no. 17, pp. 4356–4376, 2007.
 - [20] I. G. Watterson, “Calculation of probability density functions for temperature and precipitation change under global warming,” *Journal of Geophysical Research: Atmospheres*, vol. 113, no. D12, p. 106, 2008.
 - [21] M. Natrella, *NIST/SEMATECH e-Handbook of Statistical Methods*, NIST/SEMATECH, Washington, DC, USA, 2003.
 - [22] NIST/SEMATECH, *Engineering Statistics Handbook*, NIST/SEMATECH, Washington, DC, USA, 2006.
 - [23] D. J. Wheeler, *Understanding Variation: The Key to Managing Chaos*, SPC Press, Knoxville, TN, USA, 1993.
 - [24] F. E. Grubbs, “Sample criteria for testing outlying observations,” *Annals of Mathematical Statistics*, vol. 21, no. 1, pp. 27–58, 1950.