Evaluation and comparison of CMIP6 and CMIP5 models performance in simulating the runoff

Hai Guo

Chesheng Zhan (zhancs@igsnrr.ac.cn)

Institute of Geographic Sciences and Natural Resources Research Chinese Academy of Sciences

https://orcid.org/0000-0001-5014-1723

Like Ning

Zhonghe Li

Research Article

Keywords: CMIP6, CMIP5, runoff, model evaluation, uncertainty

Posted Date: April 1st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1380289/v1

Evaluation and comparison of CMIP6 and CMIP5 models performance in simulating the runoff

Hai Guo[a, b], Chesheng Zhan[a, *], Like Ning[a], Zhonghe Li[a, b]

[a] Key Laboratory of Ecosystem Network Observation and Modeling, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China;

[b] University of Chinese Academy of Sciences, Beijing 100049, China

* Correspondence: zhancs@igsnrr.ac.cn; Tel.: +86-010-64889069 (Chesheng Zhan)

Abstract

This study evaluates and compares the performance of Coupled Model Intercomparison Project Phase 6 (CMIP6) and CMIP5 in simulating the runoff on global scale and eight large-scale basins, over the period 1981–2005 using percent bias (PBIAS), correlation coefficient (CC), root mean square error (RMSE), Theil-Sen median trend, and the Taylor diagram. The CMIP models are ranked by comprehensive rating index (MR), which is determined by PBIAS, CC and RMSE three metrics. LORA, GRUN and ERA5-Land were selected as reference data sets. LORA was used as the main reference data to evaluate the historical runoff results of CMIP from 1981 to 2012 for three aspects: trend, PBIAS and uncertainty. Results reveal that: (i) CMIP6 models have obviously overvalued on the global and basins (except Amazon and Lena basin), this phenomenon was more prominent in arid and semi-arid areas (Murray-Darling and Nile basin). (ii) Compared with CMIP5 models, CMIP6 models have less uncertainty on the global scale, but it has not made outstanding progress on the basin scale. (iii) CMIP6 multi-model ensemble mean (CMIP6_MMEs) has better simulation effect than most individual models, which reduces the uncertainty among different models to some extent. (iv) There were differences in trends and PBIAS between the three reference data sets at both the global and basin scale. However, the interannual fluctuations of the three data sets were basically the same and have high correlation coefficient (except for ERA5 in the world and Nile basin), which shows that LORA data set has high reliability. The global comprehensive rating metric (GR) of CMIP6_MMEs was better than CMIP5_MMEs in all metrics, but this result was not found in eight basins. This shows that CMIP6 models has better effect in simulating global runoff and related diagnostic indicators. Implying further improvements are needs for the runoff simulation capability at the basin scale.

Keyword

CMIP6, CMIP5, runoff, model evaluation, uncertainty

Evaluation and comparison of CMIP6 and CMIP5 models performance in simulating the runoff

1.Introduction

With global warming, the water crisis is further aggravated and the changes in runoff may result in many environmental and hydrological problems (Gosling and Arnell, 2013; Padrón et al., 2020). Simulation and prediction of runoff is the key to cope with water crisis and adapt to global warming, which is also one of the research hotspots in the climate change community (Adnan et al., 2017; Seibert and Beven, 2009; Wen et al., 2019).

In recent years, with the improvement of global climate models (GCMs), product quality and usability, many researchers have started to use GCMs products to simulate and predict runoff (Dobrovolski et al., 2019; Kooperman et al., 2018; Wen et al., 2018). GCMs have been the primary tools for the simulation and prediction of global runoff, which provide an alternative way to achieve large-scale runoff data (Gain et al., 2013; Teklesadik et al., 2017; Vaze et al., 2010). The Coupled Model Intercomparison Project (CMIP) has become a central element of national and international climate change assessment. CMIP Phase 6 (CMIP6) aims to solve new scientific problems in the field of climate change, and 33 research institutions around the world have registered to participate (Eyring et al., 2016a). Compared with CMIP5, the atmospheric and ocean resolution of CMIP6 seems to be improved, it also includes new and more complex processes, including more complex land surface processes, ice fields, and permafrost, etc. (Simpkins, 2017), which improve the hydrological processes.

Although each phase of CMIP has made progress, GCMs have uncertainties owing to imperfect boundary conditions, poor parameterization, misrepresentation of physical processes, etc. (Giuntoli et al., 2015; Knutti and Sedláček, 2012; Mockler et al., 2016; Wang et al., 2014). To simulate and predict the climate change, and to understand some factors that lead to the uncertainty of GCMs, model evaluation is a key step in the development and application of any model of the environment (Chen et al., 2012; Dankers and Kundzewicz, 2020; Eyring et al., 2016b). A number of previous studies have assessed the effectiveness of runoff simulations using global model output archived in the CMIP3 and CMIP5. Milly et al. (2005) compared the output of CMIP3 models with observational runoff over 165 basins, finding that the correlations between trends computed from individual models and the observed trends are all positive, ranging from 0.05 to 0.28. Alkama et al. (2013) examined the simulation of runoff in 14 CMIP5 models at the global scale during 1958-2100. The results show that CMIP5 model can well simulate the average state of runoff (simulated runoff = observed runoff ±25%) on a global scale. With the advent of CMIP6, more and more studies are using runoff data from CMIP. Gao et al. (2021) used CMIP6 to project future glacier variation and its impact on runoff under two climate scenarios (RCP2.6 and RCP8.5). And Yin et al. (2021) used a high-end emission scenario (RCP 8.5) of CMIP6 simulated future rain-induced runoff extremes in future warming climates. However, the accuracy of CMIP6 runoff simulation has not been verified. Moreover, there is great uncertainty in the simulation of CMIP data at global and watershed scales. Dobrovolski et al. (2019) compared observational data with 28 CMIP5 models and found that although there were differences between models, reanalysis and observations, such differences were much smaller than differences between basins. It is very important to understand the improvement of runoff simulation of the existing CMIP6 models at global and basin scales and to evaluate their performance, which will provide strong support for the runoff simulation results of CMIP6 models.

In this study, multi-model ensemble is used to analyze the runoff simulation of CMIP6 model in the world and eight basins. The paper is organized as follows: Section 2 describes the reference data sets, CMIP5 model, CMIP6 models, and the methodology used. Section 3 shows results of the CMIP6 models evaluation, which are the main results of this study. In Section 4, our results are discussed and analyzed, while conclusions are drawn in Section 5.

2. Study area and data

2.1 Study area

To further evaluate the adaptability of CMIP6 model in basin scale, this study selects eight basins for evaluation while evaluating the global runoff characteristics. The eight basins (Fig. 1) located in different hydrologic and climatic regions were selected: Amazon basin in Af, Am, Aw climate region, Lena basin in the Ds, Dw, Df climate region, Mekong basin in Aw, Cw, ET climate region, Mississippi basin in Df, Cf, Cs climate region, Murray-Darling basin in BW, BS, Cf climate region, Nile basin in Aw, BW, BS climate region, Rhine basin in Cf, Df climate region and Yangtze basin in Cw, Cf, ET climate region. Moreover, the temperature difference was significant due to the latitude differences in basins. The average annual temperature of Amazon, Mekong, and Nile basins was above 20℃, while the average annual temperature of the Lena basin was below 0℃. The average annual precipitation ranges from more than 2000 mm in the Amazon basin to less than 500mm in Lena and Murray-Darling basins. Krysanova et al. (2017) shows that the runoff coefficient of Amazon and Rhine basins is above 0.7, while that of Murray-Darling and Nile basins is less than 0.12. The largest basin is the Amazon basin with an area of 6.915 million km2, and the smallest basin is the Rhine basin with an area of 173 thousand km2. Different meteorological conditions lead to altering runoff conditions. The diversity of climatic and hydrological characteristics of the eight selected typical basins ensures that they represent various conditions for the generation of global runoff.

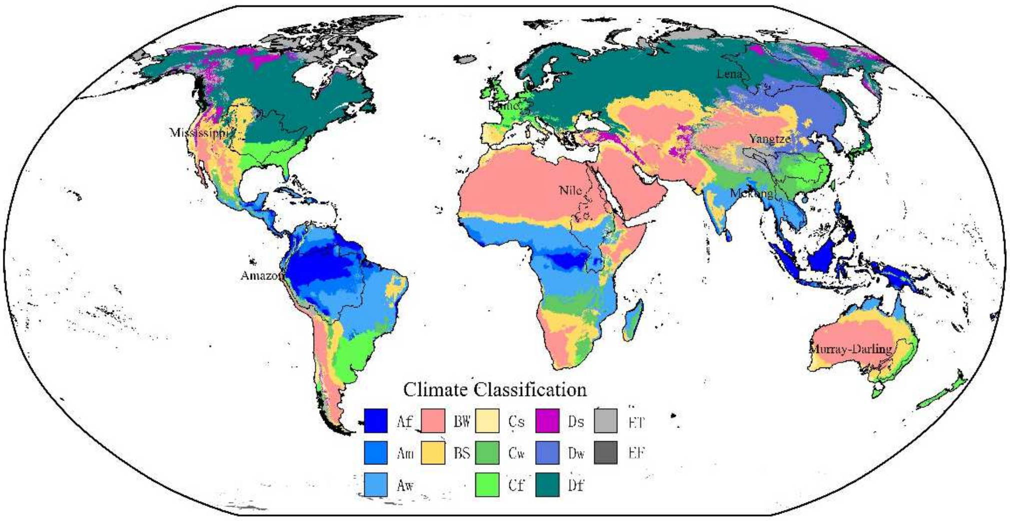

Fig. 1 Location map of the eight basins. According to Beck et al. (2018), the world can be divided into 13 climate zones: Af (Tropical, rainforest), Am (Tropical, monsoon), Aw (Tropical, savannah), BW (Arid, desert), BS (Arid, steppe), Cs (Temperate, dry summer), Cw (Temperate, dry winter), Cf (Temperate, no dry season), Ds (Cold, dry summer), Dw (Cold, dry winter), Df (Cold, no dry season), ET (Polar, tundra), and EF (Polar, frost).

2.2 Data

For each of CMIP model and the reference data set described below, this paper primary focus on monthly runoff from 1981 to 2012.

2.2.1 Model data

Monthly runoff output of CMIP6 historical runs were used in this study. Historical runoff simulations from 47 CMIP6 and 34 CMIP5 models have been released through the Earth System Grid Federation (ESGF) nodes (see https://esgf-node.llnl.gov/search/). The selected CMIP5 models have both historical and RCP8.5 experiments. Combining the historical experiment from 1980 to 2005 with the RCP8.5 experiment data from 2006 to 2012. For each phase of CMIP, the average value (A) and diagnostic standard deviation of the model ensemble members are estimated from all available models. Then, for each model, A±2 standard deviation interval is constructed around the set mean, and if the observed value ±20% contains the interval, the model is retained (Massonnet et al. 2012). The 14 CMIP6 and 5 CMIP5 models with large global deviations were removed. On a global scale, there are 33 CMIP6 and 29 CMIP5 models meet this requirement. Detailed information about these CMIP6(CMIP5) models can be viewed in Table A1(A2). The 33 CMIP6 models and 29 CMIP5 models are integrated according to the equal weight method (Massoud et al., 2019), which are labeled as "CMIP6_MMEs" and "CMIP5_MMEs", respectively. Compared with a single model, multi-model ensemble can better eliminate the uncertainty of the climate system (Abramowitz et al., 2019; Lehner et al., 2020).

2.2.2 Reference data set

Three reference data sets were used. The first is Linear Optimal Runoff Aggregate (LORA). It is a monthly global gridded synthesis runoff product (Hobeichi et al., 2019). It is a global gridded synthesis runoff product, that covers the period 1980-2012 on a 0.5° grid. The LORA data set has been extensively used in global and continental runoff assessment (Evans et al., 2020; Levizzani and Cattani, 2019). The second is Global Runoff Reconstruction (GRUN), It is an observation-based gridded global reconstruction of monthly runoff timeseries (Ghiggi et al., 2019), provided at 0.5° x 0.5° spatial resolution from 1902 to 2014. The third is ERA5-Land climate reanalysis data sets from European Centre for Medium-Range Weather Forecasts (ECMWF) and provided by EU-funded Copernicus Climate Change Service (C3S, 2019). This paper uses the monthly time series of ERA5-Land data set from 1981 to present on a 0.1° grid. Later in the text, EAR5-Land will be omitted as EAR5 for better readability.

In this paper, the overlapping time periods (1981-2012) of three reference data sets are selected to evaluate CMIP model. LORA is used as the primary reference data set, that as the reference baseline was compared with GRUN, ERA5-Land data sets and all models.

3 Methodology

3.1 Mann–Kendall Test

The Mann–Kendall (M–K) non-parametric statistical test (Mann, 1945; Kendall, 1975), has been widely used in meteorology and hydrological variables (Sharma and Ojha, 2019; Wang et al., 2020). The Mann–Kendall significance test Z and test statistic S is calculated using the following formula:

$$S=\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\operatorname{sgn}\left(x_j-x_i\right)$$

$$sgn\left(x_j\text{-}x_i\right)\begin{cases}+1 & \text{if } x_j>x_i\\ 0 & \text{if } x_j>x_i\\ -1 & \text{if } x_j>x_i\end{cases}$$

$$Var(S)=[n(n-1)(2n+5)-\sum_{i=1}^{m}t_i(t_i-1)(2t_i+5)\ ]/18$$

$$Z=\begin{cases}S-1/\sqrt{Var(S)}\ if\ S>0\\ \quad 0 \qquad\quad if\ S=0\\ S+1/\sqrt{Var(S)}\ \ if\ S<0\end{cases}$$

A positive value of S and Z indicates an ‘upward trend’; likewise, a negative value of S and Z indicates ‘downward trend’. P-value can be calculated from the test statistic Z.

3.2 Percent bias

To evaluate the runoff results of CMIP models in terms of temporal and spatial variation, this study mainly adopted percent bias (PBIAS) to evaluate the capability of model runoff simulation. PBIAS was described as follows:

$$\mathrm{PBIAS}=\left[\sum_{i=1}^{n}\mathrm{Sim}_i-\sum_{i=1}^{n}\mathrm{Rec}_i\right]\times 100\% \Big/ \sum_{i=1}^{n}\mathrm{Rec}_i$$

where Sim_i and Rec_i are the runoff of the model and LORA reference data set, respectively. The closer PBIAS is to 0, the better the simulation results of the model. The rating of PBIAS statistics refers to (Moriasi et al., 2007) (Table 1).

Table 1. Reported performance ratings for PBIAS

Value	Performance Rating
PBIAS≤10%	Very good
10%< PBIAS ≤15%	Good
15%< PBIAS ≤25%	Satisfactory
PBIAS>25%	Unsatisfactory

3.3 Taylor diagram

In this study, the Taylor diagram was used to perform uncertainty analysis in the simulated runoff of the CMIP model. Taylor diagram (Taylor, 2001) shows the graphical representation of the statistical relationship between simulations and reference data set in terms of correlation coefficient (CC), standard deviation (SD), and root mean square error (RMSE). It is widely used in the comparative study of geophysics and climate communities (Wang et al., 2020; Xu et al., 2016). The formula of CC, SD, and RMSE are as follows:

$$CC_{XY}=\mathrm{cov}(X,Y)/\sigma_X\sigma_Y$$

where

$$\mathrm{cov}(X,Y)=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})$$

where $\overline{X}$ and $\overline{Y}$ is the mean of variables X and Y, σ_X and σ_Y is the standard deviation of X and Y. RMS difference between X and Y is

$$\mathrm{RMSE}(X,Y)=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\left(X_i-\overline{X}\right)-\left(Y_i-\overline{Y}\right)\right]^2}$$

This paper uses the complete with the full Taylor diagram, which has two quadrants representing positive correlation and negative correlation, respectively. Because the different basin runoff may have widely varying numerical values, the results are normalized by LORA reference data set. The closer the position of the simulation values to that of the LORA reference value (at point REF), the better the model performance.

3.4 Comprehensive rating metrics

The comprehensive rating metrics (MR) is employed to effectively rank models (Jiang et al., 2015). The equation is as follows:

$$\mathrm{MR}=1-\frac{1}{\mathrm{nm}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{rank_i}$$

Where m is the number of models and n is the number of metrics. $\mathrm{rank_i}$ represents the ranking of the target model for index i. According to the sum of CMIP6, CMIP5 and two reference data sets, it is divided into 66 ranks per region. The model's rank is assigned based on the MR defined before. Each model is ranked from 1 (best) to 66 (worst) for BIAS, CC, and RMSE.

In addition, summarizing all the rankings should be useful in evaluating the CMIP models (Kim et al., 2020). The total ranking (TR) metrics was defined

$$\mathrm{TR}=(\mathrm{GR}+\mathrm{BR})/2$$

where GR and BR indicate the global and basin ranking, respectively as

$$\mathrm{BR_i(GR)}=(\mathrm{MR_{BIAS}}+\mathrm{MR_{CC}}+\mathrm{MR_{RMSE}})/3$$

$$\mathrm{BR}=\frac{1}{8}\sum_{\mathrm{i}=1}^{8}\mathrm{BR_i}$$

$\mathrm{BR_i}$ is the ranking of each of the eight basins.

3.5 Multi-model ensemble evaluation

Total uncertainty was assessed using reliability, sharpness metrics and Continuous Rank Probability Score (CRSP) (Pokorny et al., 2021; Zhou et al., 2016). Reliability was defined as the percentage of overlap of the LORA reference data set (annual) and the multi-model simulated ensemble bounds (annual) for the full period (1981–2012). Sharpness refers to the concentration of the models' outputs distributions. The average width ($\overline{W}$) of the confidence interval is used to measure sharpness performance:

$$\overline{W}=\frac{1}{T}\sum_{t=1}^{T}(q_{\overline{\alpha},t}-q_{\underline{\alpha},t})$$

in which $q_{\overline{\alpha},t}$ and $q_{\underline{\alpha},t}$ are the upper and lower bounds of the confidence interval, respectively. The more concentrated the confidence interval distributions, the sharper the simulation, and the sharper the better.

The CRPS (Hersbach, 2000) is a measure of the integrated squared difference between the cumulative distribution function of the forecasts and the corresponding cumulative distribution function of the reference value:

$$\mathrm{CRPS}=\frac{1}{\mathrm{T}}\sum_{\mathrm{i}=1}^{\mathrm{T}}\int_{-\infty}^{\infty}(\mathrm{F_t(x_t)}-\mathrm{H(x_t-y)})^2\mathrm{dx}$$

where $\mathrm{F_t(y)}$ is the cumulative distribution function (CDF) of an ensemble forecast at time t for variable $\mathrm{x_t}$, y is the LORA reference value, and H is the Heaviside step function which equals 0 if $\mathrm{x_t}\leq\mathrm{y}$ and equals 1 otherwise. The CRPS variates between 0 and$+\infty$; smaller value indicates better performances.

4. Results

4.1 Annual runoff variation

The annual runoff variation of CMIP6, CMIP5, and three reference data sets were shown in Figure. 2. In order to clearly compare the runoff difference between eight basins, this paper uses the average annual runoff in the world or in the basins, instead of the total runoff. Figure 2a shows the global average annual runoff change from 1981 to 2012. The runoff simulation results of CMIP6 models were higher than those of CMIP5 models. The 50% (25–75%) confidence intervals of runoff simulation results of

CMIP6 and CMIP5 only partially overlap. CMIP6_MMEs is about 0.1mm/day (accounting for 13% of CMIP5_MMEs) higher than CMIP5_MMEs. Although some CMIP models may capture the variation with the fluctuations of runoff, the lack of inter-annual variability consistent to all CMIP model results in a MMEs with smooth or even the absence of peaks. Moreover, because the wave phase of some CMIP models often deviates from the reference data, the amplitude is smaller than the reference data, especially in the vicinity of peaks and valleys, which is not ideal for the extreme value simulation of runoff. CMIP6_MMEs only showed the valleys corresponding to the reference data sets in 1983 and 1992, and other extreme points had a phase difference with the reference data sets.

Compared with the global scale, the interannual variation of runoff is more significant in the basin. The differences of climate and hydrological conditions among the eight basins have caused great differences in runoff simulation between the two generations of CMIP models in different watersheds. Among them, the simulation results of CMIP5 and CMIP6 models were highly consistent in Lena, Mississippi, Nile, and Rhine basins. The overlapping area of 25-75% confidence intervals of CMIP6 and CMIP5 models exceeds 70% of CMIP6 area. In the other four basins (Amazon, Mekong, Murray-Darling, and Yangtze basin), the runoff simulation results of CMIP6 model were much higher than those of CMIP5 model. On the basin scale, the fluctuation of annual runoff is more prominent than that of the whole world. In four basins (Amazon, Lena, Mississippi, and Rhine basins) with large runoff fluctuation, CMIP6 cannot capture the years of drought and flood. The amplitude of CMIP6_MMEs is less than 5% the amplitude of LORA reference data in Amazon and Lena basins. In other basins (Mekong Murray-Darling, Nile, and Yangtze basins) where runoff fluctuation is relatively gentle. CMIP6 can capture and reproduce the fluctuation of runoff from 1990 to 2012.

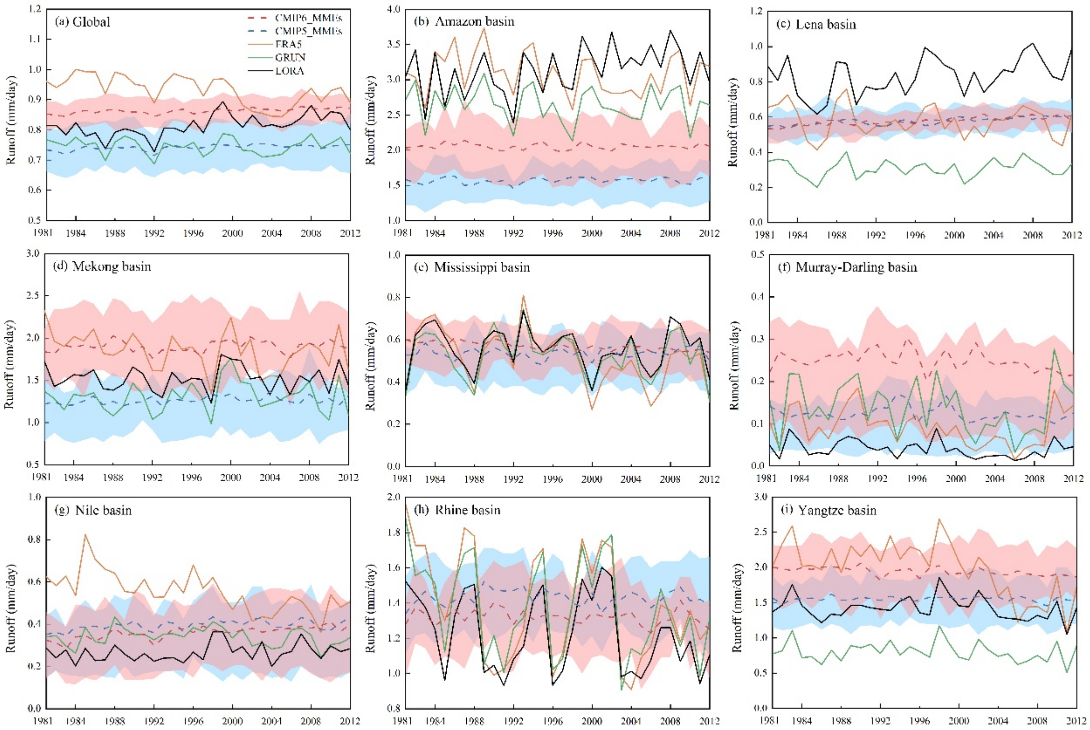

Fig. 2 Temporal change in annual runoff (1981–2012) derived from LORA (solid black curve), GRUN (solid green curve), ERA5 (solid orange curve) reference data set, 33 CMIP6 and 29 CMIP5 model simulations. (a) Global, (b) Amazon basin, (c) Lena basin, (d) Mekong basin, (e) Mississippi basin, (f) Murray-Darling basin, (g) Nile basin, (h) Rhine basin, (i) Yangtze basin. Red and blue dotted curve indicates CMIP6 and CMIP5 multi-model ensemble mean, respectively. The light red and blue shading respectively, denote the 50% confidence interval of the 33 CMIP6

models and 50% confidence intervals of the 29 CMIP5 model.

4.2 Trend from 1981 to 2012

The spatial distribution of global runoff trend changes from 1981 to 2012 is shown in Figure 3. The positive values denote increasing trends, whereas negative values denote decreasing trends. The trends that are significant at the 90% confidence level of the M-K test are stippled. CMIP6_MMEs and CMIP5_MMEs show a high degree of consistency in trend simulation in most parts of the world. CMIP6_MMEs is different from CMIP5_MMEs in a regional trend of runoff simulation results. CMIP5_MMEs only had an increasing trend in the equatorial region of South Asia and no obvious change trend in other regions. CMIP6_MMEs can simulate the increasing trend of runoff in the southern Himalayas and Indonesia, and the decreasing trend in the Yangtze basin. These changes were reflected in three reference data sets. However, CMIP6_MMEs had an increasing trend in Central Africa, reference data sets were basically stable or slightly decreasing. The analytical results for the M-K test are displayed in detail (Table 2). The test quantifies the overall trend on a global and basin scale in annual values of the average runoff. On the global, the trend of runoff simulation results of CMIP6_MMEs and CMIP5_MMEs show an increasing trend. The Z values of CMIP6_MMEs and CMIP5_MMEs were 4.330 and 3.649, respectively, with high reliability ($p<0.01$). In eight basins, CMIP6_MMEs passed the significance test ($p<0.05$) in Lena, Mississippi, Murray-Darling, Nile, and Yangtze five basins, while CMIP5_MMEs only passed in Lena and Nile basins.

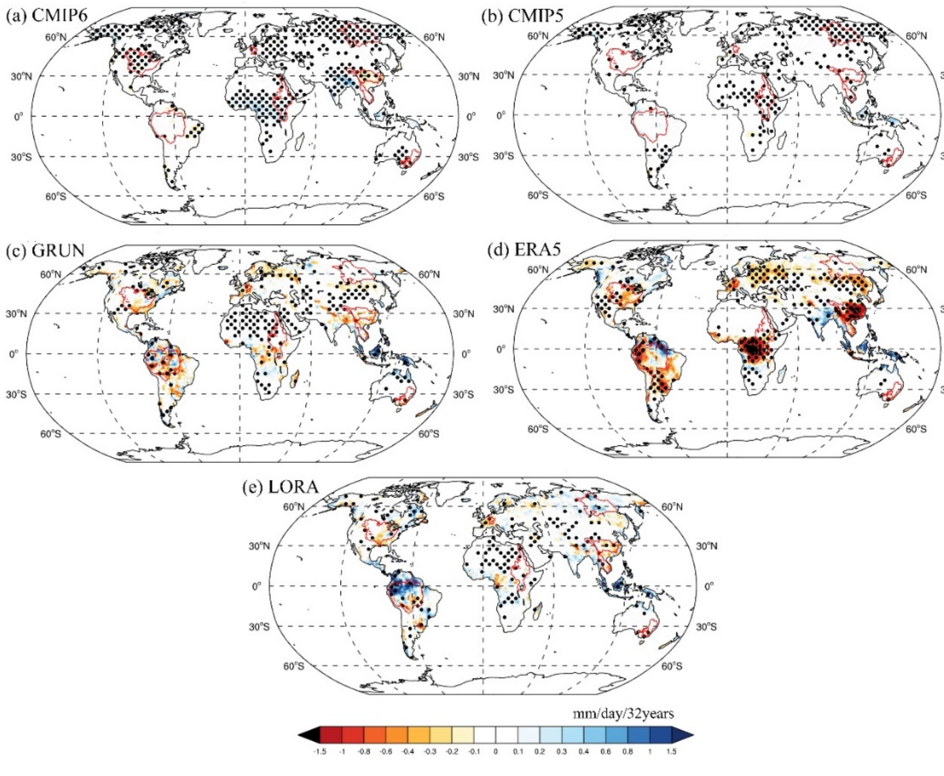

Fig. 3 Spatial distribution of runoff trends over the global land averaged from 1981 to 2012 for (a) CMIP5_MMEs;(b) CMIP6_MMEs;(c) GRUN reference data set;(d) ERA5 reference data set;(e) LORA reference data set, black dots indicate statistically significant ($p< 0.05$).

Table 2. Changes in the annual average values of runoff according to the Mann–Kendall (Z) test from 1981 to 2012

	CMIP6_MMEs	CMIP5_MMEs	LORA	GRUN	ERA5

	Z	p	Z	p	Z	p	Z	p	Z	p
global	4.330	0.000**	3.649	0.000**	3.227	0.001**	-0.924	0.355	-3.584	0.000**
Amazon	-0.308	0.758	1.249	0.212	1.735	0.083	-0.892	0.372	-1.151	0.250
Lena	5.465	0.000**	4.816	0.000**	2.059	0.039*	-0.016	0.987	-0.373	0.709
Mekong	1.054	0.292	1.800	0.072	-0.049	0.961	0.308	0.758	-1.930	0.054
Mississippi	-4.719	0.000**	0.795	0.427	-0.859	0.390	-0.730	0.466	-2.838	0.005**
Murray-Darling	-2.449	0.014*	-1.541	0.123	-1.573	0.116	-1.314	0.189	-1.346	0.178
Nile	4.849	0.000**	2.708	0.007**	1.346	0.178	-0.665	0.506	-4.200	0.000**
Rhine	-1.346	0.178	-0.114	0.910	-1.022	0.307	-1.443	0.149	-1.735	0.083
Yangtze	-3.714	0.000**	-0.146	0.884	-1.378	0.168	-0.827	0.408	-3.487	0.000**

** indicates p value < 0.01 and * indicates p value < 0.05

4.3 PBIAS of runoff

PBIAS measures the average tendency of CMIP models to be larger (positive PBIAS) or smaller (negative PBIAS) than their reference data set. Fig. 4a-b shows the PBIAS spatial distribution of the average annual runoff from LORA data set for CMIP6_MMEs and CMIP5_MMEs. Note that the PBIAS≤10% (green), 10%< PBIAS ≤15% (orange), 15%< PBIAS ≤25% (yellow), PBIAS>25% (gray) indicate performance very good, good, satisfactory and unsatisfactory, respectively. The positive (dark) and negative (light) PBIAS indicate overestimation and underestimation, respectively. The fraction (in %) of land area with positive and negative PBIAS is provided in the bottom corner. Fig. 4f shows the spatial distribution of multi-year average runoff from LORA data set.

Figures 1a and 1b show that the simulated runoff tends to be higher than LORA. According to Figure 1a and 1b, 58% (56%) of the land area shows a positive bias in CMIP6(5)_MMEs. The performance of PBIAS is satisfactory in northern Asia and Europe, eastern North America, southeast China and central Africa. It is known from Figure 4f in these areas that the average runoff is between 0.5 and 2.4 mm/day. When the runoff is in other ranges (below 0.5 mm/day or over 2.4 mm/day), the PBIAS of CMIP6_MMEs is unsatisfactory (PBIAS ≥ 25%), which means that CMIP6 has poor ability to capture extreme runoff. For example, the areas with low runoff: northern and southern Africa, Australia, western Argentina, western United States and northern China. And the areas with large runoff: Amazon and Indonesia. The performance of PBIAS in these areas is unsatisfactory.

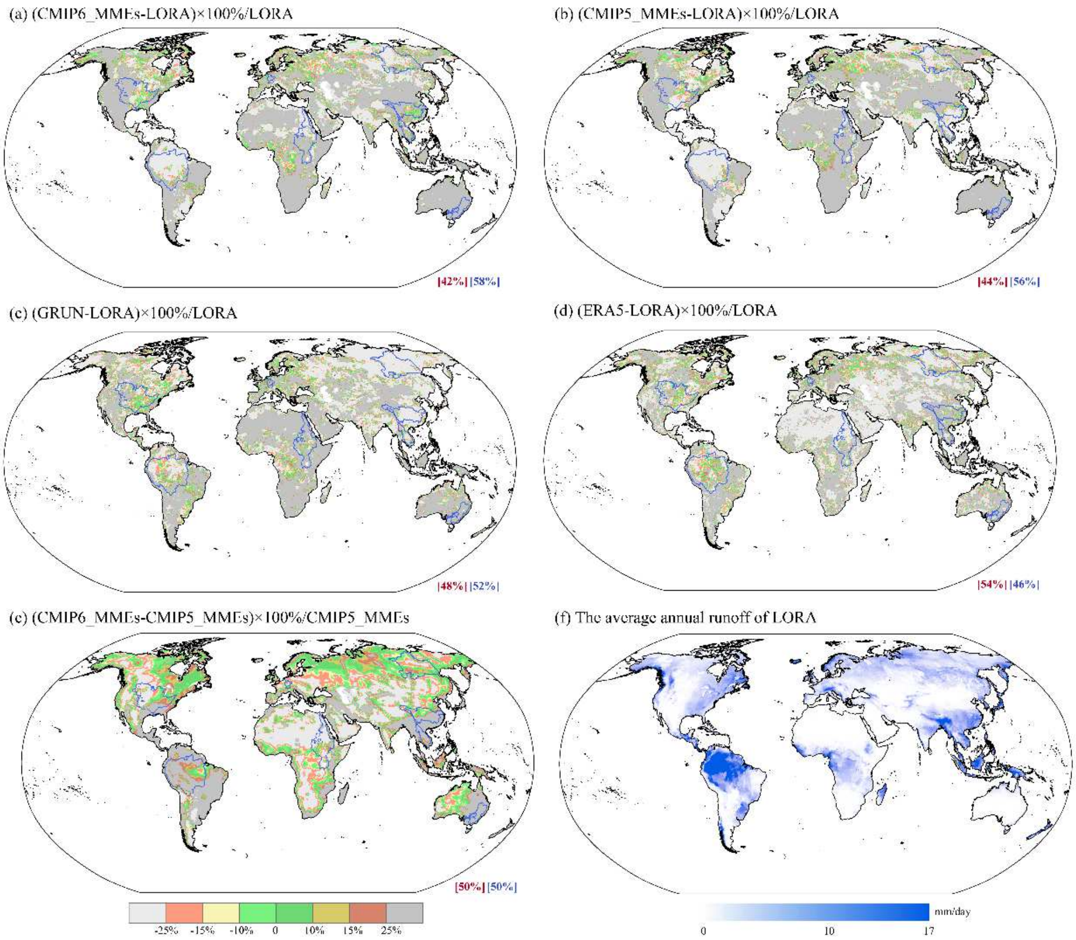

Fig. 4 The PBIAS for (a) CMIP6_ MMEs, (b) CMIP5_ MMEs, (c) GRUN, and (d) ERA5 relative to LORA; (e) PBIAS of CMIP6_ MMEs relative to CMIP5_ MMEs in 1981-2012 average annual runoff; (f) Global annual averages of the runoff in LORA data set from 1981-2012. For a-d, the percentage of land area showing negative (red) and positive (blue) PBIAS is denoted by the values in the bottom-right corner.

Due to the obvious seasonal variation of runoff, this study not only analyzed the annual PBIAS of runoff, but also analyzed the seasonal PBIAS. This paper breaks the analysis into four 3-month seasons: December–February (DJF), March–May (MAM), June–August (JJA), and September–November (SON) to calculate the PBIAS of the CMIP models and LORA reference data set.

Fig. 5 illustrates the PBIAS of runoff during the period 1980-2012 for global. CMIP6 models were less different and better performance than CMIP5 on a global scale. The PBIAS of CMIP6_MMEs and CMIP5_MMEs were good performance, which were 5.6% and -7.8%, respectively. The 25th and 75th percentile of CMIP6 (5) were 0.7% (-17.8%) and 10.4% (1.7%), respectively. PBIAS had a notable improvement in CMIP6 compared to CMIP5, as the MMEs was closer to 0, the whiskers were shorter and the interquartile model ranges was smaller. The runoff simulation results of CMIP6 and CMIP5 in the northern hemisphere were better than those in the southern hemisphere. In the northern hemisphere, the PBIAS of CMIP6_MMEs in DJF, MAM, JJA, and SON and annual were -8.8%, 14%, -6%, 0.6% and -0.4%, respectively. It was better than the PBIAS of CMIP5_MMEs were -15%, 18%, -28%, -15% and -12%, respectively. In the southern hemisphere, The PBIAS of CMIP6(5)_MMEs in DJF, MAM, JJA, and SON and annual were 46.8%(31.3%), 21.8%(0.2%), -25.8%(-49.2%), 29.3%(2.5%) and 20.8%(-0.1%), respectively. Overall, the PBIAS of CMIP5 models were better than CMIP6 models in the southern hemisphere. However, CMIP6 whiskers were shorter, and the quartile range was smaller than CMIP5 in the southern hemisphere. The same was true in the northern hemisphere and on a global scale.

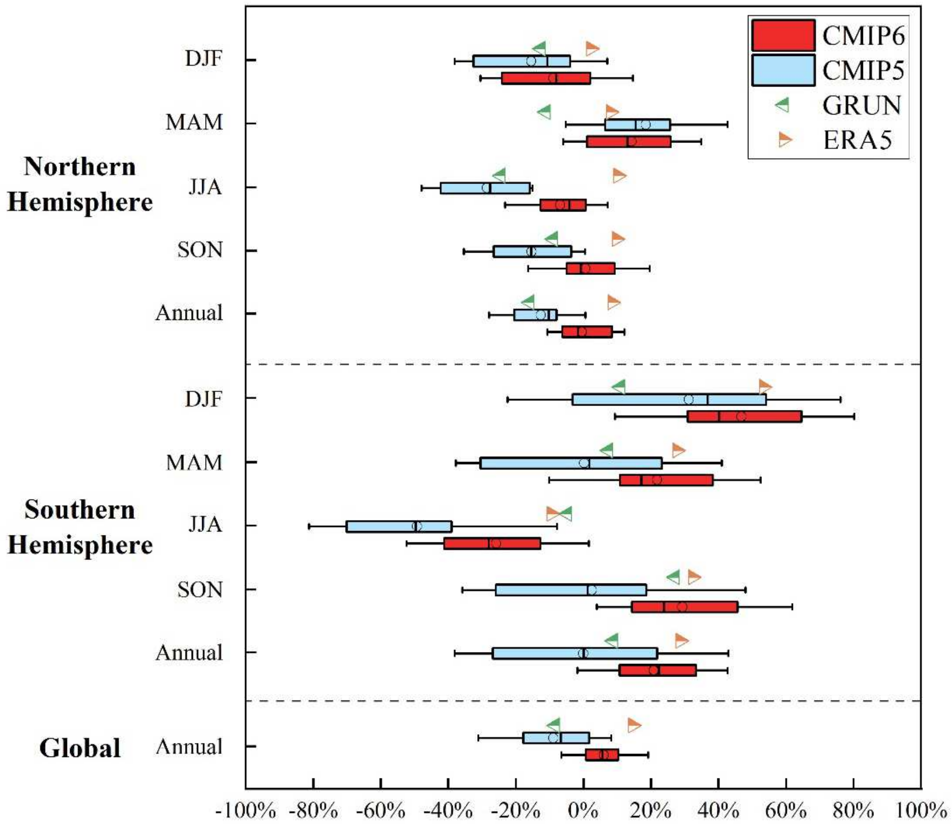

Fig. 5 Box-and-whisker plots for runoff PBIAS calculated from 33 CMIP6 (red) and 29 CMIP5 (blue) models. Upper panel is the northern hemisphere, the middle panel is the southern hemisphere, and the bottom panel is global. The box marks the median and interquartile range, the line marks the 5%–95% range，circle represents the average of multiple models. The reference data sets are indicated by different colored arrows of GRUN (green) and ERA5 (Orange).

On the basin scale, the PBIAS of CMIP cannot all achieve satisfactory performance (Fig. 6). The annual PBIAS of CMIP6(5)_MMEs in Lena, Mississippi, Nile, and Rhine basin were -30.6% (-30.6%), 2.1% (-3.9%), 35.1 % (48.3%), 7.9% (16.9%), respectively. This result means that the annual PBIAS of CMIP6 models were better than CMIP5 in these basins. The PBIAS of CMIP6 and CMIP5 models were the best in Mississippi basin. The PBIAS of CMIP6(5)_MMEs in DJF, MAM, JJA and SON were -8.3% (-2.4%), 11.1% (1.3%), 9.5% (-9.0%) and -9.0% (-12.6%), respectively. This result means that the PBIAS in seasons were not as optimistic as the annual. The same situation also occurs in other basins. In Amazon、Mekong、Murray-Darling, and Yangtze basin, the annual PBIAS of CMIP6(5)_MMEs were -34.5% (-49.9%), 24.9% (-16.9%), 517.4% (208.1%) and 36.9% (9.4%), respectively. The performance of CMIP6 is not better than CMIP5, even worse than CMIP5. Figure 6a,6b shows that runoff was obviously underestimated in Amazon and Lena basin. In the Amazon basin, the PBIAS of CMIP6(5)_MMEs were -17.6% (-27.9%), -38.3% (-51.8%) ,-61.1% (-80.5%), -28.1% (-44.1%) and -34.2% (-52.3%) in DJF, MAM, JJA, SON and annual, respectively. In Lena basin, they were -86.5% (-94.0%), -38.0% (-27.1%), -25.5% (-46.3%), -12.0% (-25.1%) and -31.3% (-32.6%), respectively. It should be noted that the runoff of the Murray-Darling and Nile basin was low, the result of PBIAS was often larger, the scale of the X axis is adjusted here (Fig 6e, 6h). In these two basins, the simulation results of CMIP model tend to be higher, which is more prominent in winter. The 25th and 75th percentile PBIAS of CMIP6(5) in DJF in Murray-Darling basin were 150.5% (13.7%) and 922.0% (656.2%), respectively.

To summarize, compared with the global scale, CMIP model has greater differences, longer beard and wider quartile range in basin scale. The PBIAS of CMIP6 models has been improved in winter (except Murray-Darling basin).

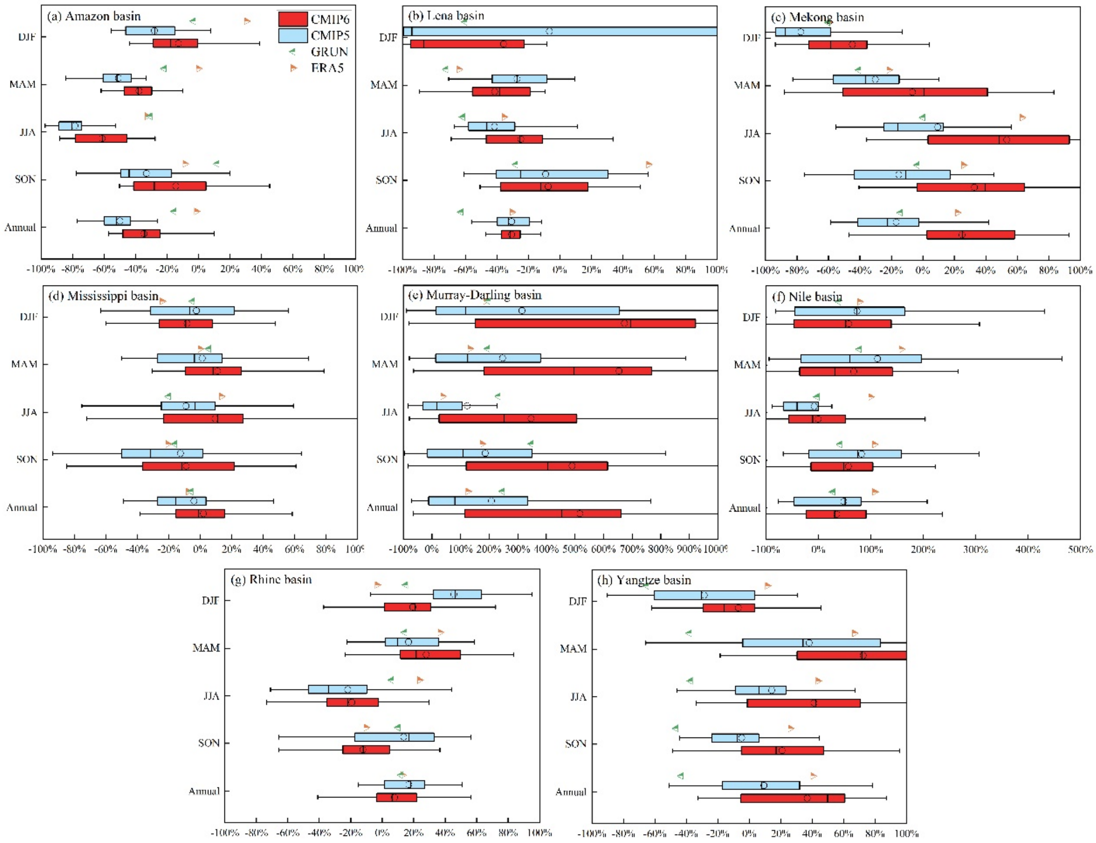

Fig. 6 Box-and-whisker plots for runoff percent bias from CMIP models in eight basins.

4.4 Taylor diagram analysis

PBIAS can well evaluate the differences in multi-year average state of runoff, but it has some limitations in evaluating temporal changes. Taylor diagram is used to represent the statistical variables of CC and RMS together, and the uncertainty caused by the temporal and spatial is analyzed.

Fig 7 the normalized Taylor diagrams of the average runoff from the historical simulations (1981-2012) of CMIP model and 2 reference data sets. Note that 33 CMIP6 models and 29 CMIP5 models in the paper are represented by red and blue dots in Figure 7. The simulation result is assumed close to the reference value, when there would be relatively high correlation, low RMS errors and minimum difference of standard deviation with respect to the reference value.

On a global scale (Fig 7a), most CMIP6 models had the CC between 0 and 0.4, the RMSE between 1 and 1.5, and the SD between 0.8 and 1.2. Compared to the CMIP models, the simulation results of MMEs were superior to other models, especially CC was much higher than any single model. The CC of CMIP6(5)_MMEs was 0.536 (0.590), which passed the significance test of 99% reliability (i.e., α = 0.01,CC=0.436). The SD and RMSE were the smallest, about 1.1 (1.3) and 1 (1.1) respectively.

However, in eight basins, CMIP6 models have the CC between -0.3 and 0.3. The best CC was 0.304 in Lena basin cannot pass the significance test of 95% reliability (i.e., α = 0.05, CC=0.339). The CC of CMIP5_MMEs passed the significance test of 95% reliability in Amazon and Nile basin, which were

0.364 and 0.411, respectively. The RMSE of CMIP model was mainly between 1 and 1.5, but in Mekong, Rhine and Yangtze basin was between 1.25 and 1.75. The SD of CMIP models in eight basins was between 0.7 and 1.3. Among them, the SD of most models in Amazon, Lena, and Mississippi basin was less than 1, which indicates that the CMIP models have lower variability in these basins.

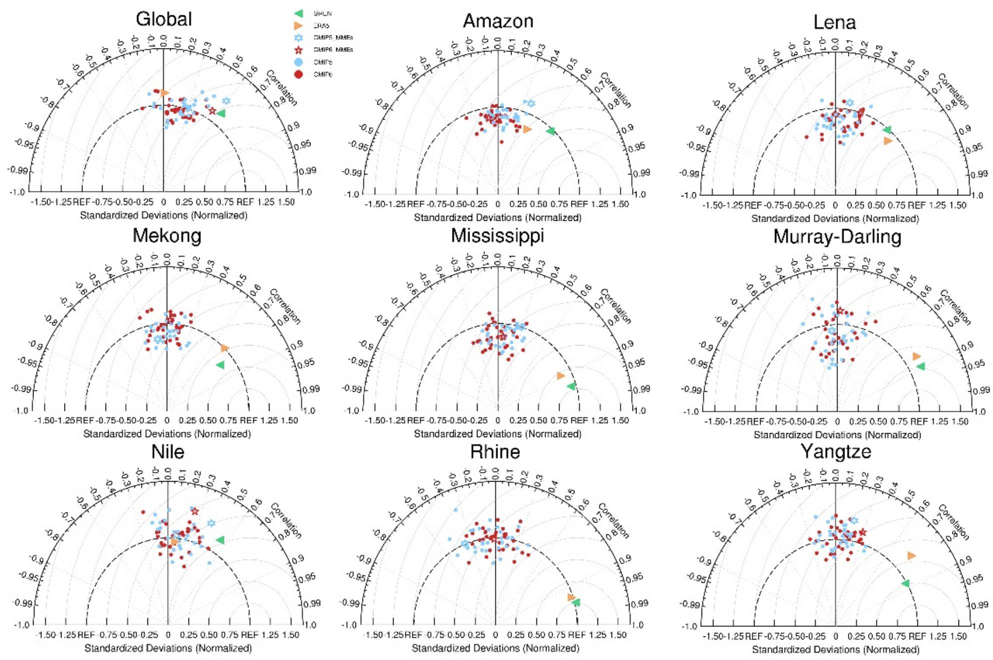

Fig. 7 Taylor diagram of the average runoff from the historical simulations (1981-2012) of 33 CMIP6 models (red dot), 29 CMIP5 models (blue dot) and 2 reference data sets (triangle) compared with the LORA data set. The azimuthal angle denotes the correlation coefficient between model and LORA reference results (gray solid line), the radial values are normalized spatial standard deviations of the runoff time series referenced or modeled (where referenced or modeled correspond to the "REF" or reference value of 1.0).

4.5 Ranking of climate model

In this section, CMIP models are ranked according to PBIAS, CC, and RMSE three metrics (Fig. 8). The global ranking (GR) and basin ranking (BR) were the comprehensive ranking of three metrics in the global and eight basins, respectively (Please refer to appendix 1 for the ranking of the three metrics in the basin). The total ranking (TR) was the average of BR and GR. The blue line shows a higher ranking in most metrics and the red line shows lower ranking.

In CMIP6, the FGOALS-f3-L, CNRM-CM6-1, and TaiESM1 were ranked in the top three in TR (Fig. 8). In CMIP5, the CCSM4, CESM1-BGC and BNU-ESM models were ranked in the top three in TR. The top models do not have good rankings for all global and basins. For example, GR and BR in TaiESM1 are ranked 3,25 respectively. By analyzing the global model ranking of each diagnostic metric, it was found that there was nine of the top ten in BIAS are in CMIP6 model. The RMES performance is satisfactory in TaiESM1, INM-CM5-0 and CAS-ESM2-0 models. However, CC ranked higher in the MPI-ESM-MR and CCSM4 from CMIP5, which also participate in CMIP6 (Gettelman et al., 2019; Mauritsen et al., 2019). The GR, BR, and TR of CMIP6_MMEs (CMIP5_MMEs) are 2 (30), 9 (8) and 2 (11), respectively. In general, the runoff simulation results of MME were excellent and consistently shows better performance than most single models. Strong evidence of Fig. 8 was found that CMIP6 has obvious improvement in BIAS and RMSE. The blue line appears more frequently in CMIP6 than in CMIP5, indicating that the models of CMIP6 show good performance regardless of the metrics. Thus, the model performance in CMIP6 is superior overall to that in CMIP5.

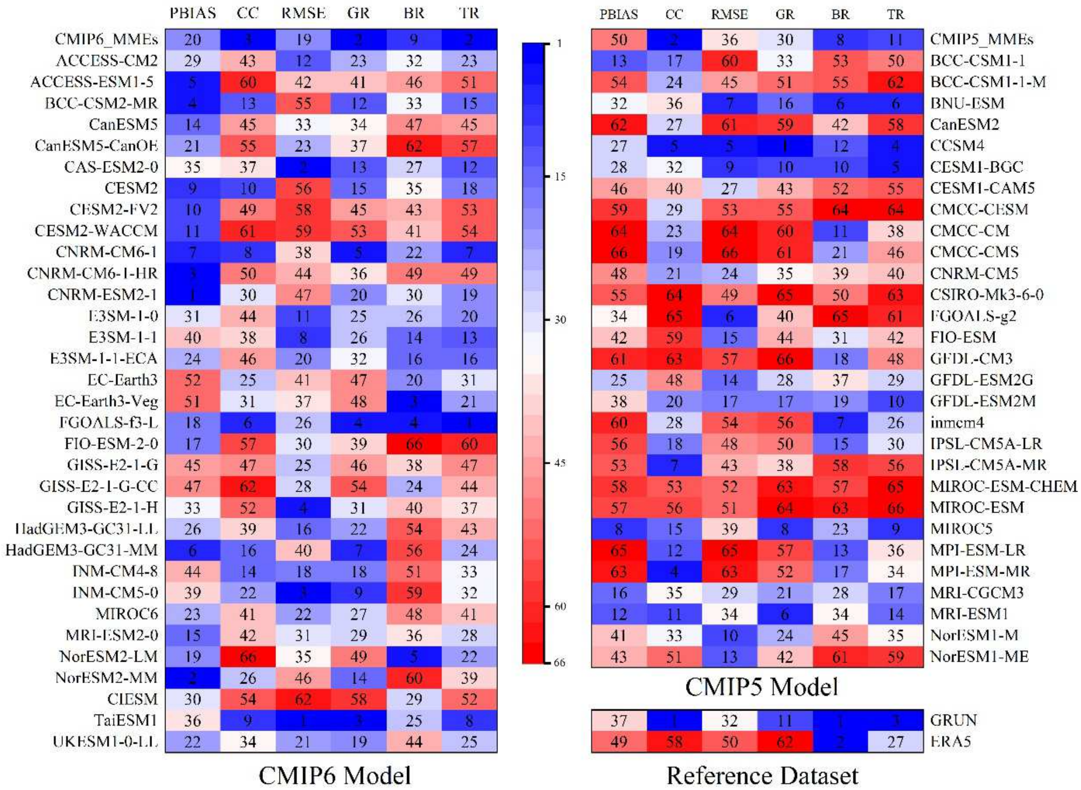

Fig.8 The portrait diagram for the rankings of PBIAS, CC and RMSE. between runoff for CMIP6 (left), CMIP5 (Top right) and reference data set (bottom right). The global comprehensive rating metrics (GR) was the comprehensive ranking of three indicators on a global scale, the basins comprehensive rating metrics (BR) was the comprehensive ranking of three indicators in eight basins, TR was the average of BR and GR. Color denotes the model's rank for each index.

5. Discussion

This paper results show that CMIP6_MMEs has a good ability to capture runoff during the period 1981-2012, particularly on a global scale. Importantly, the simulated trend change range of CMIP6 is more obvious than CMIP5 (Fig .2), which can better simulate the trend change of runoff in Yangtze Basin and Qinghai-Tibet Plateau. However, compared with the reference data set, the trend change is still small. Due to the sharp reduction of Arctic glaciers and sea ice, the runoff in the high latitudes of the northern hemisphere has increased significantly (Jahfer et al., 2017; Lutz et al., 2014), and this trend CMIP6 has also been well captured.

In this article, the seasonal runoff simulation results of CMIP model are obviously worse than the annual, especially in JJA. The results show that the runoff simulation results of CMIP6 model on a global scale are better than those at the basin scale. Previous studies have shown that CMIP models have greater uncertainty on regional scale than global scale (Fiedler et al., 2020; Waliser et al., 2020; Watterson, 2015). Most CMIP models obviously underestimate the annual average runoff in the Amazon basin because of underestimation of precipitation in the Amazon basin (Coppola et al., 2021; Zhou et al., 2012). Beck et al. (2017) pointed out that CMIP5 models underestimate of simulated runoff occurred in snow-dominated areas (Lena basin), this situation has not been improved in CMIP6. The runoff capture capacity of CMIP6 is poor in Murray-Darling and Nile basins. Poor vegetation coverage and soil hydrophobicity may lead to serious higher runoff results of CMIP models in arid and semi-arid areas (Deb et al., 2019; Kling et al., 2015). It may also be related to the hydrological structure defects of CMIP models in arid and semi-arid areas (Schewe et al., 2014; Zhang et al., 2016). Gudmundsson and Seneviratne (2015) showed that global hydrological models (GHMs) struggle in reproducing the seasonality of runoff. The CMIP model's selection were determined according to standard deviation interval of the global reference runoff, which also results in better simulation results on a global scale. Therefore, the reference data of the corresponding basin can also be used to screen out the CMIP model more suitable for simulating the basin.

When calculating the comprehensive rating index (CMR), the trend index (Z) is not added. The trend of runoff time series simulated by CMIP calculated by MK method in the world and eight basins only has a few models passed the statistical significance of student's T standard. Some articles also pointed out that the interannual variation of runoff lacks obvious trend (Gelfan et al., 2020). Usually, models with higher horizontal spatial resolution tend to produce better simulations (Sarmadi et al., 2019; Travis et al., 2016), but they are not shown in runoff simulation. In this paper, CNRM-CM6-1-HR, E3SM-1-0, E3SM-1-1, E3SM-1-1-ECA, EC-Earth3, EC-Earth3-Veg, HadGEM3-GC31-MM and CMCC-CM (CMIP5) were high-resolution models (Table A1 and A2), but they have no high ranking on a global or basin scale (Fig. 8).

5.1 Uncertainty of the CMIP

It is known that CMIP data sets are uncertain due to many reasons, such as convective parameterization, tunable parameters, model resolution. In this paper, the uncertainty of CMIP model is analyzed from two aspects: the uncertainty between model and model and the uncertainty between model and reference data set.

For the uncertainty between model and reference data set, this paper used objective functions PBIAS, CC and RMS were taken into consideration.

The PBIAS of CMIP6 has been significantly improved on a global scale (Fig. 5). However, PBIAS still cannot reach the satisfactory performance (PBIAS $\leq$25%) on some basins (Fig.6). Figure 9 is obtained by calculating the area ratio of PBIAS in performance rating from Figure 4. Results show that the ratio of PBIAS with very good and satisfactory performance in CMIP6_MMEs is higher than that in CMIP5_MMEs (except Murray-Darling basin). In CMIP6_MMEs, the area (in %) of PBIAS for very good, good, and satisfactory performance was 11.62%, 17.26%, and 27.68%, respectively in the world. The PBIAS performance of CMIP5_MMEs (-16.9% and 9.36%) is better than that of CMIP6_MMEs (24.94% and 36.9%) in Mekong and Yangtze basins, respectively (Fig. 6). However, the satisfactory area ratio of CMIP6_MMEs (32.47% and 44.74%) is higher than CMIP5_MMEs (22.51% and 21.93%) in these two basins. This shows that although CMIP6 is captured accurately in some areas (for example: lower Yangtze basin), but the greater uncertainty caused by overestimation in some areas. Compared with CMIP5, CC and RMSE of CMIP 6 do not improve Compared with CMIP5, CC and RMSE of CMIP6 have not improved in the world and eight basins. The uncertainty in temporal and spatial has not been reduced.

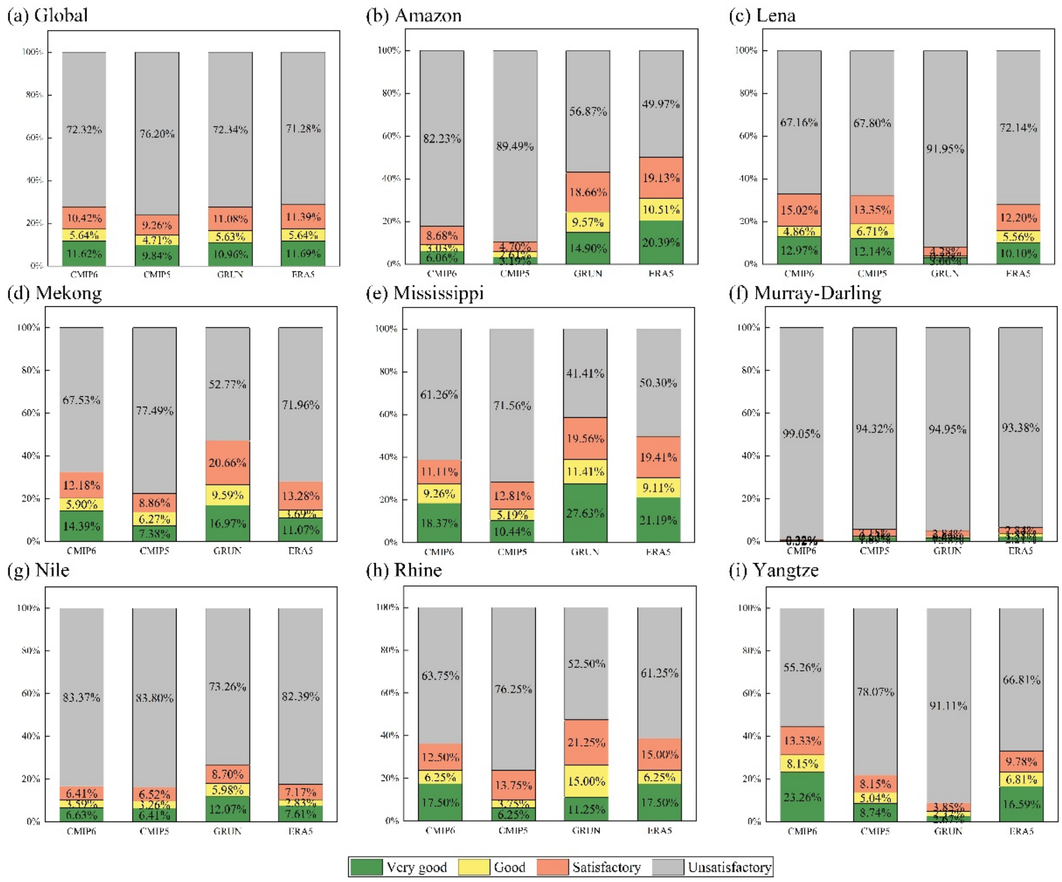

Fig.9 The area (in %) of PBIAS from CMIP6_MMEs, CMIP5_MMEs, GRUN, and ERA5 in performance rating.

For the uncertainty between model and model, this paper used reliability (the coverage of LORA reference data set), sharpness (CMIP simulation interval width) and CRPS. The smaller the CRPS, the lower the uncertainty. The results of these functions are presented in Fig.10. Compared with CMIP5 models, CMIP6 models have been significantly improved on a global scale. The reliability of 10% confidence interval of CMIP6 model is 19% and the interval width is 0.014 mm/day, which has greatly improved. The CRPS of CMIP6 models is 0.034 mm/day, which is better than 0.046 of CMIP5.Among the eight basins, the CPRS of CMIP6 is best (0.065) in Mississippi Basin. The CPRS of CMIP6 and CMIP5 in Murray-Darling and Nile basins are also less than 0.1, which is mainly caused by the low annual average runoff. The CPRS is worst performance in Amazon basin. It can be seen from Fig.10 that not only the confidence interval width is large, but also the reliability is low in Amazon basin.

Compared with CMIP5, CMIP6 model has less uncertainty in Amazon, Mississippi, and Rhine basins and the whole world. This is a particularly reassuring result.

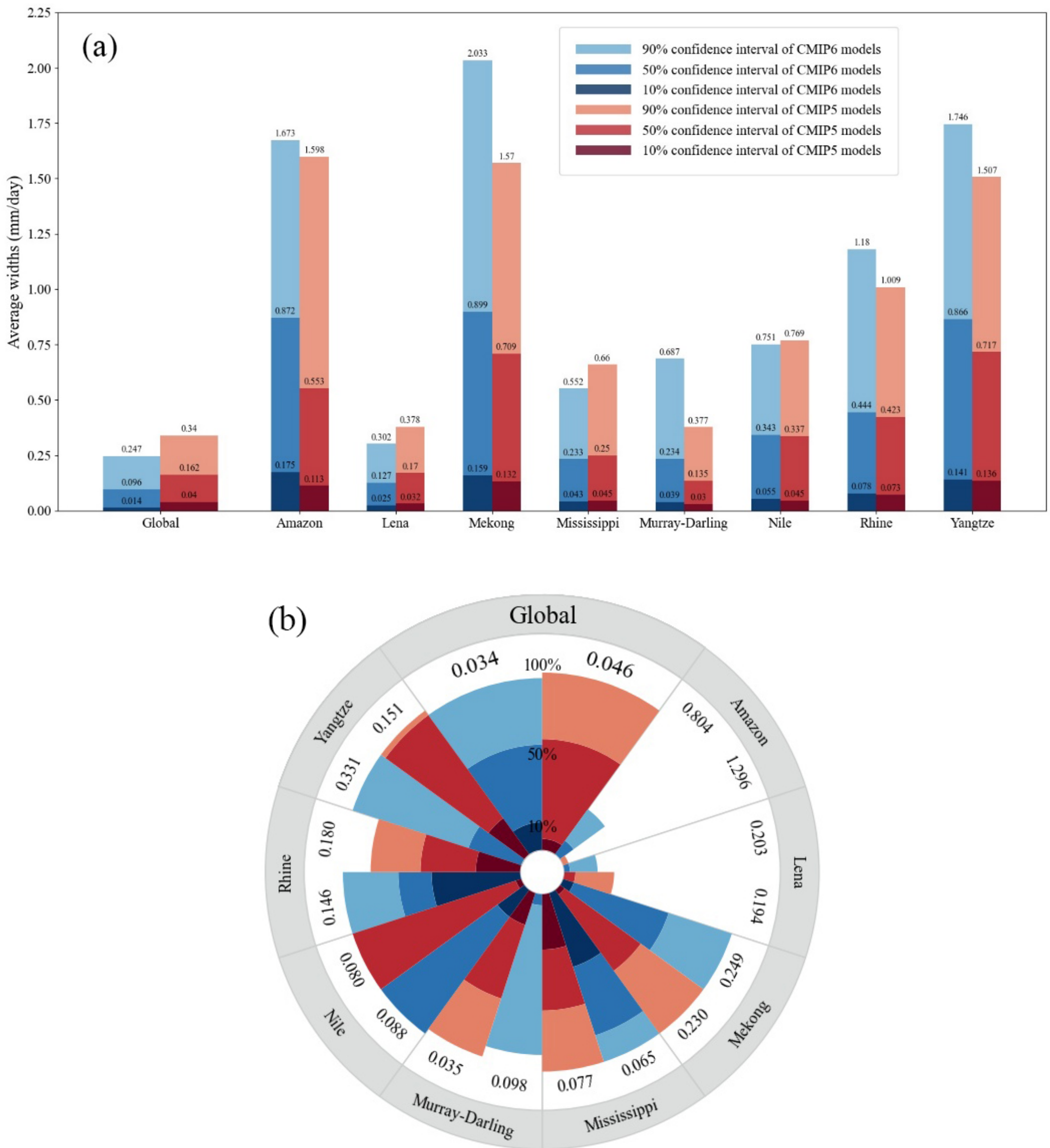

Fig.10 (a) The average width of the confidence interval of CMIP6 models and CMIP5 models during 1980-2012. (b) The reliability of confidence interval of CMIP model, in which the outer number represents CRPS.

5.2 Uncertainty of the reference data set

In the model evaluation, the uncertainty of the reference data set is often ignored by some scientists. Some scientists assume that the uncertainty of the model is dominant and ignore the uncertainty of the reference data set (Knutti et al., 2017). Or other thinks that the reference data set is true and accurate (Lloyd, 2012). Ignoring data set uncertainty can thus lead to false or distorted conclusions (Zumwald et al., 2020). In the past studies of evaluation variables, only one or two different data sets were usually used (Flato et al., 2013).

In this paper, there are great differences among the three reference data sets in evaluating runoff. For the trend, the biggest difference from the spatial distribution of the three reference data sets is in the Amazon basin of South America (Fig. 2). The trend of LORA was increased significantly in the northern and decreased slightly in the southern Amazon basin. Similar trends have been reported by Espinoza Villar et al. (2009) in their work on regional discharge evolutions in the Amazon basin. In other regions, the trend of LORA and GRUN is roughly consistent, but EAR5 has a downward trend in central Africa and southeast China. For PBIAS, the PBIAS of GRUN and ERA5 are -8.3% and 14.6%, respectively, on the global scale (Fig.5). In the northern hemisphere, the PBIAS of GRUN in each season is stable at about -14%, and that of ERA5 is about 8%. Compared with the northern hemisphere, the PBIAS of GRUN and EAR5 were increased by about 20% in each season (except ERA5 in JJA) from the southern hemisphere. This shows that the runoff results of LORA in the southern hemisphere may be lower than the measured values. The PBIAS of ERA5 has obvious fluctuation in different seasons, which was similar to that of CMIP6_MMEs. The PBIAS of ERA5 (CMIP6_MMEs) in DJF, MAM, JJA, SON and annual were 53.3% (46.8%), 27.8% (21.8%), -9.5% (-25.8%), 32.3% (29.3%) and 28.6% (20.8%) respectively. Among the eight basins, the annual PBIAS of GRUN and ERA5 were less than |15%| (performance good) in Amazon, Rhine and Mississippi basins. Only the PBIAS of GUNR in Rhine performed good in each season. This shows that the reference data has great uncertainty on PBIAS. Excluding the influence of PBIAS, the three reference data sets are highly consistent in terms of interannual variation in 8 basins and the world (Fig. 2). Except for a few years, the occurrence time and the increase and decrease of the drought and flood years are the same. Fig.7 shows a strong CC of GRUN and LORA on a global scale and eight basins, and the lowest CC value was 0.545 (i.e., $\alpha = 0.01$, CC=0.436) in Nile basin. The CC of ERA5 and LORA (except global and Nile basin) passed the significance test of 99% reliability (i.e., $\alpha = 0.01$, CC=0.436). There were high CC of GRUN and ERA5 in Rhine basin, which were 0.967 and 0.947, respectively. In terms of rankings, GRUN and ERA5 reference data sets ranked 1 and 2 respectively on the basin scale but ranked 11 and 62 respectively on a global scale. This reflects that the simulation effect of some CMIP models on the global scale can be comparable to the reference data sets, but they still need to be strengthened the capture ability at the basins.

To sum up, LORA data set has better reproduced the historical trend of runoff change and the average climate state from 1980 to 2012, which has a good correlation with GRUN and ERA5 data sets. Therefore, LORA data set is selected as the primary reference data set in this paper.

6. Conclusion

This study evaluated the capability of simulated runoff from MMEs of CMIP6 and CMIP5 models. Model trend and biases on global scale and basin scale were compared between CMIP6 and CMIP5 and with three reference data sets (LORA, GRUN, and ERA5). Besides the MMEs, this paper has shown the differences and uncertainties of individual models as well as those of the reference data sets. The main findings of the study are:

The results of this study suggest that CMIP6 models can well capture the characteristics of annual and seasonal runoff on global, especially CMIP6_MMEs. The simulation results of some CMIP6 models were better than the reference data set.

In the eight basins, the simulation results of CMIP6 were not as good as those on a global scale. Mississippi and Rhine basins were the best ones, while Murray-Darling and Nile basins were not ideal. This is highly consistent with CMIP5.

In the three reference data sets selected in the article, we cannot conclude which data set is the best. We encourage using an ensemble of observations from different sources and centers to estimate runoff and better assess their associated uncertainties.

In total, CMIP6 has improved the simulation performance of runoff compared with CMIP5. However, GCMs still have great potential of further improvement in arid regions. Although the deviation still exists, it is gradually decreasing. It shows that with the development of the climate model, it is increasingly suitable to analyze the changes on a large scale.

Acknowledgment:

We acknowledge Climate Change Research Centre, University of New South Wales for the data from LORA, ETH Zurich for the data from GRUN, European Centre for Medium-Range Weather Forecasts (ECMWF) for the reanalysis data set from ERA5-land. We also gratefully acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups for producing and making available their model output, the Earth System Grid Federation (ESGF) for archiving the data and providing access, and the multiple funding agencies who support CMIP6 and ESGF. We sincerely appreciate the anonymous reviewers' helpful comments and the editor's efforts in improving this manuscript.

Data availability

The datasets used in the present study are freely available: (i) The CMIP6 and CMIP5 datasets from https://esgf-node.llnl.gov/search/ (Eyring et al. 2016). (ii) The LORA dataset is freely available for download on https://geonetwork.nci.org.au/geonetwork/srv/eng/catalog.search#/metadata/f9617_9854_8096_5291 (Hobeichi et al., 2019). (iii) The GRUN dataset is available from the ETHZ Research Collection at https://doi.org/10.3929/ethz-b-000324386 (Ghiggi et al., 2019). (iv) The ERA5-Land reanalysis datasets from https://www.ecmwf.int/en/era5-land (C3S, 2019).

Code availability

Not applicable.

References

Abramowitz, G. *et al.*, ESD Reviews: Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing, *Earth System Dynamics* **10**(2019), pp. 91-105.

Adnan, M., Nabi, G., Saleem Poomee, M., Ashraf, A., Snowmelt runoff prediction under changing climate in the Himalayan cryosphere: A case of Gilgit River Basin, *Geoscience Frontiers* **8**(2017), pp. 941-949.

Alkama, R., Marchand, L., Ribes, A., Decharme, B., Detection of global runoff changes: results from observations and CMIP5 experiments, *Hydrology and Earth System Sciences* **17**(2013), pp. 2967-2979.

Beck, H.E. *et al.*, MSWEP: 3-hourly 0.25° global gridded precipitation (1979–2015) by merging gauge, satellite, and reanalysis data, *Hydrology and Earth System Sciences* **21**(2017), pp. 589-615.

Beck, H.E. *et al.*, Present and future Koppen-Geiger climate classification maps at 1-km resolution, *Sci Data* **5**(2018), p. 180214.

C3S ERA5-land Reanalysis. Copernicus Climate Change Service date of access December, 2019, https://cds.climate.copernicus.eu/cdsapp#!/home (2019).

Chen, H., Xu, C.-Y., Guo, S., Comparison and evaluation of multiple GCMs, statistical downscaling and hydrological models in the study of climate change impacts on runoff, *Journal of Hydrology* **434-435**(2012), pp. 36-45.

Coppola, E. *et al.*, Climate hazard indices projections based on CORDEX-CORE, CMIP5 and CMIP6 ensemble, *Climate Dynamics*(2021).

Dankers, R., Kundzewicz, Z.W., Grappling with uncertainties in physical climate impact projections of water resources, *Climatic Change* **163**(2020), pp. 1379-1397.

Deb, P., Kiem, A.S., Willgoose, G., A linked surface water-groundwater modelling approach to more realistically simulate rainfall-runoff non-stationarity in semi-arid regions, *Journal of Hydrology* **575**(2019), pp. 273-291.

Dobrovolski, S.G., Yushkov, V.P., Istomina, M.N., Statistical Modeling of the Global River Runoff Using GCMs: Comparison with the Observational Data and Reanalysis Results, *Water Resources* **46**(2019), pp. S17-S24.

Espinoza Villar, J.C. *et al.*, Contrasting regional discharge evolutions in the Amazon basin (1974–2004), *Journal of Hydrology* **375**(2009), pp. 297-311.

Evans, J., Abramowitz, G., Hobeichi, S., Conserving Land–Atmosphere Synthesis Suite (CLASS), *Journal of Climate* **33**(2020), pp. 1821-1844.

Eyring, V. *et al.*, Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, *Geoscientific Model Development* **9**(2016a), pp. 1937-1958.

Eyring, V. *et al.*, Towards improved and more routine Earth system model evaluation in CMIP, *Earth System Dynamics* **7**(2016b), pp. 813-830.

Fiedler, S. *et al.*, Simulated Tropical Precipitation Assessed across Three Major Phases of the Coupled Model Intercomparison Project (CMIP), *Monthly Weather Review* **148**(2020), pp. 3653-3680.

Flato, G. et al. (2013). Evaluation of climate models. In: Climate change 2013: The physical science basis. Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change. In Climate change 2013 (Vol. 5, pp. 741–866). Cambridge, MA: Cambridge University Press.

Gain, A.K., Apel, H., Renaud, F.G., Giupponi, C., Thresholds of hydrologic flow regime of a river and investigation of climate change impact—the case of the Lower Brahmaputra river Basin, *Climatic Change* **120**(2013), pp. 463-475.

Gao, H.K. *et al.*, Assessing glacier retreat and its impact on water resources in a headwater of Yangtze River based on CMIP6 projections, *Science of the Total Environment* **765**(2021).

Gelfan, A. *et al.*, Does a successful comprehensive evaluation increase confidence in a hydrological model intended for climate impact assessment?, *Climatic Change* **163**(2020), pp. 1165-1185.

Gettelman, A. *et al.*, High Climate Sensitivity in the Community Earth System Model Version 2 (CESM2), *Geophysical Research Letters* **46**(2019), pp. 8329-8337.

Ghiggi, G., Humphrey, V., Seneviratne, S.I., Gudmundsson, L., GRUN: an observation-based global gridded runoff dataset from 1902 to 2014, *Earth System Science Data* **11**(2019), pp. 1655-1674.

Giuntoli, I., Vidal, J.P., Prudhomme, C., Hannah, D.M., Future hydrological extremes: the uncertainty from multiple global climate and global hydrological models, *Earth System Dynamics* **6**(2015), pp. 267-285.

Gosling, S.N., Arnell, N.W., A global assessment of the impact of climate change on water scarcity, *Climatic Change* **134**(2013), pp. 371-385.

Gudmundsson, L., Seneviratne, S.I., Towards observation-based gridded runoff estimates for Europe, *Hydrology and Earth System Sciences* **19**(2015), pp. 2859-2879.

Hersbach, H., Decomposition of the continuous ranked probability score for ensemble prediction systems, *Weather and Forecasting* **15**(2000), pp. 559-570.

Hobeichi, S., Abramowitz, G., Evans, J., Beck, H.E., Linear Optimal Runoff Aggregate (LORA): a global gridded synthesis runoff product, *Hydrology and Earth System Sciences* **23**(2019), pp. 851-870.

Jahfer, S., Vinayachandran, P.N., Nanjundiah, R.S., Long-term impact of Amazon river runoff on northern hemispheric climate, *Sci Rep* **7**(2017), p. 10989.

Jiang, Z., Li, W., Xu, J., Li, L., Extreme Precipitation Indices over China in CMIP5 Models. Part I: Model Evaluation, *Journal of Climate* **28**(2015), pp. 8603-8619.

Kendall, M.G. Rank Correlation Methods, 4th ed.; Charless Grin: London, UK, 1975; ISBN 0195208374.

Kim, M.K. *et al.*, Performance Evaluation of CMIP5 and CMIP6 Models on Heatwaves in Korea and Associated Teleconnection Patterns, *Journal of Geophysical Research: Atmospheres* **125**(2020).

Kling, H., Stanzel, P., Fuchs, M., Nachtnebel, H.-P., Performance of the COSERO precipitation–runoff model under non-stationary conditions in basins with different climates, *Hydrological Sciences Journal* **60**(2015), pp. 1374-1393.

Knutti, R., Sedláček, J., Robustness and uncertainties in the new CMIP5 climate model projections, *Nature Climate Change* **3**(2012), pp. 369-373.

Knutti, R. *et al.*, A climate model projection weighting scheme accounting for performance and interdependence, *Geophysical Research Letters*(2017).

Kooperman, G.J. *et al.*, Plant Physiological Responses to Rising CO2 Modify Simulated Daily Runoff Intensity With Implications for Global-Scale Flood Risk Assessment, *Geophysical Research Letters* **45**(2018).

Krysanova, V. *et al.*, Intercomparison of regional-scale hydrological models and climate change impacts projected for 12 large river basins worldwide—a synthesis, *Environmental Research Letters* **12**(2017).

Kumar, S. *et al.*, Terrestrial contribution to the heterogeneity in hydrological changes under global warming, *Water Resources Research* **52**(2016), pp. 3127-3142.

Lehner, F. *et al.*, Partitioning climate projection uncertainty with multiple large ensembles and CMIP5/6, *Earth System Dynamics* **11**(2020), pp. 491-508.

Levizzani, V., Cattani, E., Satellite Remote Sensing of Precipitation and the Terrestrial Water Cycle in a Changing Climate, *Remote Sensing* **11**(2019).

Lloyd, E.A., The role of 'complex' empiricism in the debates about satellite data and climate models, *Studies in History and Philosophy of Science* **43**(2012), pp. 390-401.

Lutz, A.F., Immerzeel, W.W., Shrestha, A.B., Bierkens, M.F.P., Consistent increase in High Asia's runoff due to increasing glacier melt and precipitation, *Nature Climate Change* **4**(2014), pp. 587-592.

Mann, H.B., NONPARAMETRIC TESTS AGAINST TREND, *Econometrica* **13**(1945), pp. 245-259.

Massonnet, F. *et al.*, Constraining projections of summer Arctic sea ice, *The Cryosphere* **6**(2012), pp. 1383-1394.

Massoud, E.C., Espinoza, V., Guan, B., Waliser, D.E., Global Climate Model Ensemble Approaches for Future Projections of Atmospheric Rivers, *Earth's Future* **7**(2019), pp. 1136-1151.

Mauritsen, T. *et al.*, Developments in the MPI-M Earth System Model version 1.2 (MPI-ESM1.2) and Its Response to Increasing CO2, *J Adv Model Earth Syst* **11**(2019), pp. 998-1038.

Milly, P.C., Dunne, K.A., Vecchia, A.V., Global pattern of trends in streamflow and water availability in a changing climate, *Nature* **438**(2005), pp. 347-350.

Mockler, E.M., Chun, K.P., Sapriza-Azuri, G., Bruen, M., Wheater, H.S., Assessing the relative importance of parameter and forcing uncertainty and their interactions in conceptual hydrological model simulations, *Advances in Water Resources* **97**(2016), pp. 299-313.

Moriasi, D.N. *et al.*, Model evaluation guidelines for systematic quantification of accuracy in watershed simulations, *Transactions of the Asabe* **50**(2007), pp. 885-900.

Padrón, R.S. *et al.*, Observed changes in dry-season water availability attributed to human-induced climate change, *Nature Geoscience* **13**(2020), pp. 477-481.

Pokorny, S. *et al.*, Cumulative Effects of Uncertainty on Simulated Streamflow in a Hydrologic Modeling Environment, *Elementa-Science of the Anthropocene* **9**(2021).

Sarmadi, F., Huang, Y., Thompson, G., Siems, S.T., Manton, M.J., Simulations of orographic precipitation in the Snowy Mountains of Southeastern Australia, *Atmospheric Research* **219**(2019), pp. 183-199.

Schewe, J. *et al.*, Multimodel assessment of water scarcity under climate change, *Proc Natl Acad Sci U S A* **111**(2014), pp. 3245-3250.

Seibert, J., Beven, K.J., Gauging the ungauged basin: how many discharge measurements are needed?, *Hydrology and Earth System Sciences* **13**(2009), pp. 883-892.

Sharma, Ojha, Changes of Annual Precipitation and Probability Distributions for Different Climate Types of the World, *Water* **11**(2019).

Simpkins, G., Progress in climate modelling, *Nature Climate Change* **7**(2017), pp. 684-685.

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

Teklesadik, A.D. *et al.*, Inter-model comparison of hydrological impacts of climate change on the Upper Blue Nile basin using ensemble of hydrological models and global climate models, *Climatic Change* **141**(2017), pp. 517-532.

Travis, K.R. *et al.*, Why do Models Overestimate Surface Ozone in the Southeastern United States?, *Atmos Chem Phys* **16**(2016), pp. 13561-13577.

Vaze, J. *et al.*, Climate non-stationarity – Validity of calibrated rainfall–runoff models for use in climate change studies, *Journal of Hydrology* **394**(2010), pp. 447-457.

Waliser, D. *et al.*, Observations for Model Intercomparison Project (Obs4MIPs): status for CMIP6, *Geoscientific Model Development* **13**(2020), pp. 2945-2958.

Wang, G., Dommenget, D., Frauen, C., An evaluation of the CMIP3 and CMIP5 simulations in their skill of simulating the spatial structure of SST variability, *Climate Dynamics* **44**(2014), pp. 95-114.

Wang, Z., Zhan, C., Ning, L., Guo, H., Evaluation of global terrestrial evapotranspiration in CMIP6 models, *Theoretical and Applied Climatology* **143**(2020), pp. 521-531.

Watterson, I.G., Improved Simulation of Regional Climate by Global Models with Higher Resolution: Skill Scores Correlated with Grid Length*, *Journal of Climate* **28**(2015), pp. 5985-6000.

Wen, X. *et al.*, Two-phase extreme learning machines integrated with the complete ensemble empirical mode decomposition with adaptive noise algorithm for multi-scale runoff prediction problems, *Journal of Hydrology* **570**(2019), pp. 167-184.

Wen, X. *et al.*, Future changes in Yuan River ecohydrology: Individual and cumulative impacts of climates change and cascade hydropower development on runoff and aquatic habitat quality, *Sci Total Environ* **633**(2018), pp. 1403-1417.

Xu, Z., Hou, Z., Han, Y., Guo, W., A diagram for evaluating multiple aspects of model performance in simulating vector fields, *Geoscientific Model Development* **9**(2016), pp. 4365-4380.

Yin, J.B. *et al.*, Does the Hook Structure Constrain Future Flood Intensification Under Anthropogenic Climate Warming?, *Water Resources Research* **57**(2021).

Zhang, Y., Shao, Q., Zhang, S., Zhai, X., She, D., Multi-metric calibration of hydrological model to capture overall flow regimes, *Journal of Hydrology* **539**(2016), pp. 525-538.

Zhou, R.R., Li, Y., Lu, D., Liu, H.X., Zhou, H.C., An optimization based sampling approach for multiple metrics uncertainty analysis using generalized likelihood uncertainty estimation, *Journal of Hydrology* **540**(2016), pp. 274-286.

Zhou, X. *et al.*, Benchmarking global land surface models against the observed mean annual runoff from 150 large basins, *Journal of Hydrology* **470-471**(2012), pp. 269-279.

Zumwald, M. *et al.*, Understanding and assessing uncertainty of observational climate datasets for model evaluation using ensembles, *WIREs Climate Change* **11**(2020).

Statements & Declarations

Funding:

This research was funded by the National Key R&D Program of China [grant number 2017YFA0603702]; the National Natural Science Foundation of China [grant number 41701023].

Ethics declarations

Ethics approval

This research did not involve human subjects. Meteorological datasets used in this study can all be obtained from publicly accessible archives.

Consent to participate

This research did not involve human subjects.

Consent for publication

This research did not involve personal information for which consent was to be sought.

Conflict of interest

The authors declare no competing interests.

Author Contributions

All authors contributed to the study conception and design. Hai Guo: data curation, formal analysis, visualization, software, writing—original draft preparation. Zhonghe Li: visualization. Like Ning: conceptualization, methodology, writing—reviewing and editing. Chesheng Zhan: supervision. All authors read and approved the final manuscript.

Appendix A:

Table A1 Model names, institution, and resolution for CMIP6 models used in the paper

No.	Model ID/acronym	Resolution	institution	country
1	ACCESS-CM2	192 x 144	CSIRO-ARCCSS	Australia
2	ACCESS-ESM1-5	192 x 145	CSIRO	Australia
3	BCC-CSM2-MR	320 x 160	BCC	China
4	CanESM5	128 x 64	CCCma	Canada
5	CanESM5-CanOE	128 x 64	CCCma	Canada
6	CAS-ESM2-0	256 x 128	CAS	China
7	CESM2	288 x 192	NCAR	USA
8	CESM2-FV2	144 x 96	NCAR	USA
9	CESM2-WACCM	288 x 192	NCAR	USA
10	CIESM	288 x 192	THU	China

11	CNRM-CM6-1	256 x 128	CNRM-CERFACS	France
12	CNRM-CM6-1-HR	720 x 360	CNRM-CERFACS	France
13	CNRM-ESM2-1	256 x 128	CNRM-CERFACS	France
14	E3SM-1-0	360 x 180	E3SM-Project	USA
15	E3SM-1-1	360 x 180	E3SM-Project	USA
16	E3SM-1-1-ECA	360 x 180	E3SM-Project	USA
17	EC-Earth3	512 x 256	EC-Earth-Consortium	Many Countries in Europe
18	EC-Earth3-Veg	512 x 256	EC-Earth-Consortium	Many Countries in Europe
19	FGOALS-f3-L	288 x 192	CAS	China
20	FIO-ESM-2-0	288 x 192	CAS	China
21	GISS-E2-1-G	144 x 90	NASA-GISS	USA
22	GISS-E2-1-G-CC	144 x 90	NASA-GISS	USA
23	GISS-E2-1-H	144 x 90	NASA-GISS	USA
24	HadGEM3-GC31-LL	192 x 144	MOHC, NERC	UK
25	HadGEM3-GC31-MM	432 x 324	MOHC, NERC	UK
26	INM-CM4-8	180 x 120	INM	Russia
27	INM-CM5-0	180 x 120	INM	Russia
28	MIROC6	256 x 128	MIROC	Japan
29	MRI-ESM2-0	320 x 160	MRI	Japan
30	NorESM2-LM	144 x 96	NCC	Norway
31	NorESM2-MM	288 x 192	NCC	Norway
32	TaiESM1	288 x 192	AS-RCEC	Taiwan
33	UKESM1-0-LL	192 x 144	MOHC, NERC, NIMS-KMA, NIWA	UK, Korea, New Zealand

Table A2 CMIP5 models used in the paper, details are the same as table A1

No.	Model ID/acronym	Resolution	institution	country
1	BCC-CSM1-1	320 x 160	BCC	China
2	BCC-CSM1-1-M	128 x 64	BCC	China
3	BNU-ESM	128 x 64	GCESS	China
4	CanESM2	128 x 64	CCCma	Canada
5	CCSM4	288 x 192	NCAR	USA
6	CESM1-BGC	288 x 192	NSF-DOE-NCAR	USA
7	CESM1-CAM5	288 x 192	NSF-DOE-NCAR	USA
8	CMCC-CESM	96 x 48	CMCC	Italy
9	CMCC-CM	480 x 240	CMCC	Italy
10	CMCC-CMS	192 x 96	CMCC	Italy
11	CNRM-CM5	256 x 128	CNRM-CERFACS	France
12	CSIRO-Mk3-6-0	192 x 96	CSIRO-QCCCE	Australia
13	FGOALS-g2	128 x 60	LASG-CESS	China
14	FIO-ESM	128 x 64	FIO	China
15	GFDL-CM3	144 x 90	NOAA GFDL	USA
16	GFDL-ESM2G	144 x 90	NOAA GFDL	USA
17	GFDL-ESM2M	144 x 90	NOAA GFDL	USA
18	INM-CM4	180 x 120	INM	Russia
19	IPSL-CM5A-LR	96 x 96	IPSL	France

20	IPSL-CM5A-MR	144 x 143	IPSL	France
21	MIROC5	256 x 128	MIROC	Japan
22	MIROC-ESM	128 x 64	MIROC	Japan
23	MIROC-ESM-CHEM	128 x 64	MIROC	Japan
24	MPI-ESM-LR	192 x 96	MPI-M	Germany
25	MPI-ESM-MR	192 x 96	MPI-M	Germany
26	MRI-CGCM3	320 x 160	MRI	Japan
27	MRI-ESM1	320 x 160	MRI	Japan
28	NorESM1-M	144 x 96	NCC	Norway
29	NorESM1-ME	144 x 96	NCC	Norway

Appendix B: Model Ranking in Basin

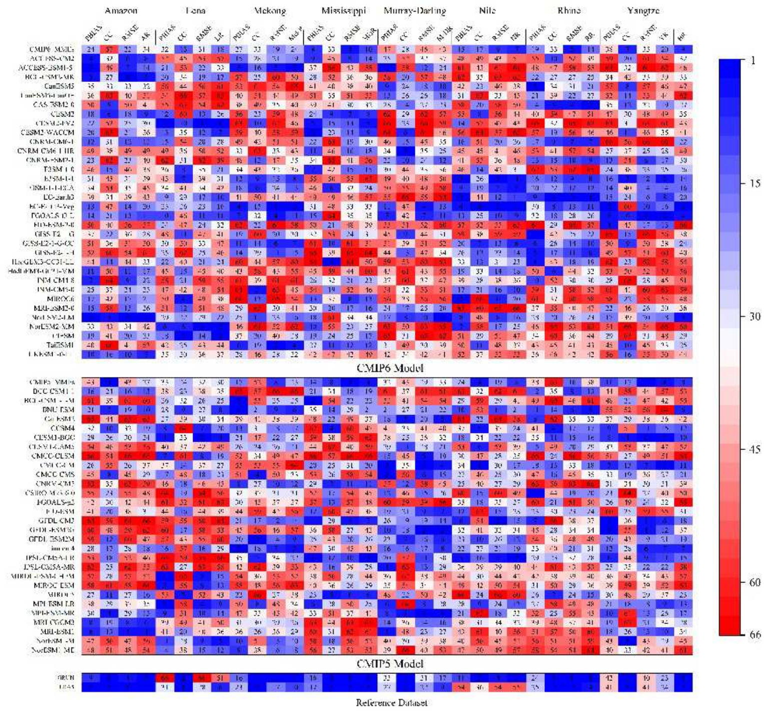

Figure B The portrait diagram for the rankings of PBIAS, CC, RMSE. Upper panel is the CMIP6 model, the middle panel is the CMIP5 model, and the bottom panel is reference data set. The AR, LR, MEKR, MISR, M-DR, NR, RR and YR are the comprehensive rating metrics of Amazon, Lena, Mekong, Mississippi, Murray-Darling, Nile, Rhine and Yangtze basin, respectively. The basins comprehensive rating metrics (BR) is the comprehensive ranking of three indicators in eight basins.