



Evaluation of precipitation reanalysis products for regional hydrological modelling in the Yellow River Basin

Cong Jiang¹ · Eric J. R. Parteli² · Qian Xia^{3,4} · Yaping Shao¹

Received: 26 August 2023 / Accepted: 20 November 2023 / Published online: 13 December 2023
© The Author(s) 2023

Abstract

This study evaluates six precipitation reanalysis products for the Yellow River Basin using gridded rain gauge data, runoff data and the Atmospheric and Hydrological Modelling System (AHMS) simulations. The assessment begins with comparing the annual, seasonal, monthly and daily precipitation of the products with gridded rain gauge data. The AHMS is then run with each of the precipitation reanalysis products under two scenarios: one with calibrated rainfall-runoff and the other without. The simulated streamflow is then compared with the corresponding observations. It is found that non-gauge-corrected products tend to overestimate precipitation, especially for mountainous regions. Amongst the six products evaluated, the China Meteorological Forcing Dataset (CMFD) and WATCH Forcing Data methodology applied to ERA5 (WFDE5/CRU+GPCC) are identified as the most accurate products, supported by both statistical and hydrological comparisons. This consistency in statistical and hydrological comparisons suggests the potential applicability of the hydrological comparison method using the AHMS in ungagged catchments, even in the presence of significant anthropogenic impacts. Furthermore, the calibration of the hydrological model significantly impacts the model's response to precipitation, effectively compensating for deficiencies in rainfall data within certain limits. This study highlights accurate representation of extreme rainfall events in precipitation products has a significant impact on calibrated soil parameters and is particularly important in hydrological modelling. It enhances our understanding of the reliability of hydrological simulations and provides valuable insights for the assessment of precipitation reanalysis products in large arid and semiarid basins affected by human activities.

1 Introduction

Precipitation is usually measured using rain gauges, ground-based weather radars, satellite sensors, etc. For continental-scale estimates, precipitation reanalysis is generated by the assimilation of observations into weather prediction models or climate models, and precipitation reanalysis products have been widely used for hydrological modelling (Michaelides et al., 2009). Several such products have been developed

with different objectives, temporal ranges, data sources and resolutions, as summarized in Table 1. Precipitation is also one of the most important forcing variables for hydrological modelling. Thus, it is crucial to evaluate how hydrological models respond to different precipitation products to gain insight into model uncertainty and characterize the different biases in precipitation reanalysis. Although rainfall measurements are assimilated into the reanalysis, different data sources and assimilation systems lead to different accuracies, which are difficult to assess. Therefore, hydrological modellers are often confronted with the question of how their hydrological simulations depend on the choice of precipitation products and which product performs the best for a given region and time (Bitew et al. 2012).

Numerous studies have evaluated precipitation products to understand their respective advantages and limitations (Beck et al. 2019, 2020; Maggioni et al. 2016). Some studies (Dinku et al. 2008; Gao and Liu 2013; Hersbach et al. 2020; Hirpa et al. 2010) used statistical methods and rain gauge observations for the evaluation, while others (Bitew et al. 2012; Qi et al. 2016; Stisen and Sandholt 2010; Tang

✉ Cong Jiang
c.jiang@uni-koeln.de

¹ Institute for Geophysics and Meteorology, University of Cologne, Pohligstr. 3, D 50969 Cologne, Germany

² Faculty of Physics, University of Duisburg-Essen, Lotharstr. 1, D 47057 Duisburg, Germany

³ Hubei Water Resources and Hydropower Science and Technology Promotion Center, Hubei Water Resources Research Institute, 430070 Wuhan, China

⁴ School of Water Resources and Hydropower Engineering, Wuhan University, Bayi Road. 299, 430072 Wuhan, China

Table 1 Overview of seven (3 or 6-) hourly gridded precipitation products evaluated in this study. Abbreviations in the data source column are defined as follows: G: gauge; S: satellite; and R: reanalysis

Filename	Description	Data source	R data	G data	Spatial resolution	Temporal resolution	Temporal coverage	Reference
OBS	Gridded rain gauges data from CMA	G	None	CMA	1.0°	Hourly	1979–2003	Yang et al. (2010)
NCEP/NCAR	National Centers for Environmental Prediction and National Centers for Atmospheric Research	R	NCEP/NCAR	None	2.5°	6-Hourly	1979–2021	Kalnay et al. (1996)
ERA5	The fifth generation ECMWF reanalysis for the global climate and weather	R	ERA5	None	0.25°	Hourly	1959–present	Hersbach et al. (2020)
GLDAS 2.0	Global Land Data Assimilation System	G, S, R	NCEP/NCAR	GPCP and TRMM	0.5°	3-hourly	1979–2015	Sheffield et al. (2006)
WFDE5/CRU WFDE5/ (CRU+GPCC)	WATCH Forcing Data methodology applied to ERA5	G, R	ERA5	CRU TS CRU TS and GPCC	0.5°	Hourly	1979–2019	Cucchi et al. (2020)
CMFD	The China Meteorological Forcing Dataset	G, S, R	GLDAS	CMA	0.1°	3-hourly	1979–2018	He et al. (2020)

et al. 2016) employed hydrological modelling to compare the simulated streamflow or soil moisture with the corresponding observations. Although hydrological models offer the possibility of using observed water discharge to evaluate precipitation products, previous studies have focused primarily on wet headwater areas and small basins when using hydrological models (Bitew et al. 2012; Tang et al. 2016; Wang et al. 2023), because quantifying streamflow in large watersheds is difficult, especially in arid and semiarid areas (Pilgrim et al. 1988; Vorosmarty and Sahagian 2000).

This study aims to evaluate precipitation products in large-scale, arid and semiarid basins using a sub-continental scale hydrological model considering the significant impacts of human activities on the water cycle. Specifically, we evaluate the precipitation reanalysis products for the Yellow

River Basin using the offline Atmospheric and Hydrological Modelling System (AHMS) integrated with an irrigation module (AHMS-IRR hereafter, Jiang et al. (2022)). The Yellow River Basin offers a compelling case study for assessing the accuracy of these precipitation products, for its diverse range of climates and geomorphic features. We initially use the hydrological model to evaluate its response to various precipitation products and subsequently analyse the model's response. Our evaluation includes widely used near-surface precipitation reanalysis products, including NCEP/NCAR, ERA5, GLDAS 2.0, WFDE5/CRU, WFDE5/(CRU+GPCC) and CMFD. We directly compare these products with rain gauge observations and indirectly assess their reliability by comparing the model simulated streamflow with streamflow observations.

2 Study area and data

2.1 Study area

As shown in Fig. 1, the Yellow River Basin (795,000 km²) is the largest watershed in northern China, and the Yellow River is the second-longest river in the country (5464 km). The Yellow River flows through the Qinghai-Tibetan Plateau, Inner Mongolia and Loess Plateaus and the Huang-Huai-Hai (Yellow, Huai and Hai Rivers) Plain. Much of the Loess Plateau is an arid and semiarid region. The average annual temperature of the Yellow River Basin is -4° , and the annual precipitation amount is approximately 450 mm, which is heavily affected by the East Asian monsoon and unevenly distributed. According to the Yellow River Water Resources Bulletin, the middle and upper reaches up to the Huayuankou station cover an area of 73,036 km² and account for 91.82% of the total basin area. The annual runoff at the Huayuankou station is 56.7 billion m³. The Yellow River downstream from the Huayuankou station is an above-ground hanging river with a small catchment area covering 3% of the basin. Therefore, the present study evaluates precipitation data covering the upper reaches of the Huayuankou station in the Yellow River Basin.

2.2 Data

Table 1 presents seven gridded precipitation products, including one gridded rain gauge product and six precipitation reanalysis products. These products were classified into

three types: the first was based exclusively on gauge data, the second on reanalysis data that are not bias-corrected by gauge data and the third on the assimilation of reanalysis data and gauge data. Amongst these, we consider one first-type (OBS), two second-type (NCEP/NCAR and ERA5) and four third-type (WFDE5/CRU, WFDE5/(CRU+GPCC), GLDAS 2.0 and CMFD) products. These seven gridded precipitation products were further interpolated to the AHMS-IRR simulation grids with a resolution of 20 km in this study. They were used as forcing data for the model runs and for comparative analysis. Monthly, seasonal and annual precipitation data were generated by averaging the gridded daily precipitation data.

2.2.1 Non-gauge-corrected reanalysis precipitation products

NCEP/NCAR (National Centers for Environmental Prediction and the National Center for Atmospheric Research) Reanalysis is a global dataset of atmospheric variables to support the needs of the climate research communities (Kalnay et al. 1996). A frozen state-of-the-art global data assimilation system and a complete database were used in the NCEP/NCAR Reanalysis project.

ERA5 is the fifth-generation reanalysis product created by the European Center for Medium-Range Weather Forecasts (ECMWF) for global weather and climate from 1979 to near real-time (Hersbach et al. 2020). This reanalysis is carried out with the 4D-Var data assimilation and model forecasts in CY41R2 of the Integrated Forecast System (IFS). ERA5 hourly data on single levels are used in the present work.

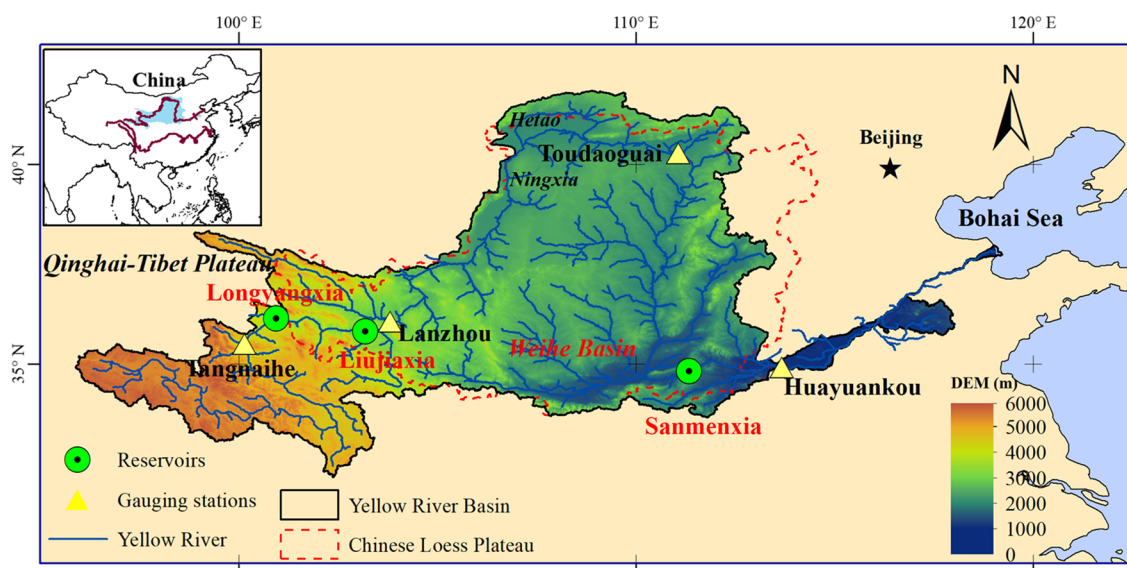


Fig. 1 Location and topography of the Yellow River Basin. The map includes the Yellow River network, the Loess Plateau and the four main hydrological stations referred to in the main text, i.e. Tangnaihe (TNH), Lanzhou (LZ), Toudaoguai (TDG) and Huayuankou (HYK)

ERA5 has the following main advantages compared to other reanalysis products: higher spatial and temporal resolutions, advanced modelling and data assimilation systems (4D-Var), and a more considerable amount of integrated historical observations and parameters.

2.2.2 Observed precipitation products

An observed hourly gridded precipitation product across China was developed by Yang et al. (2010) based on the observed daily precipitation data provided by the China Meteorological Administration (CMA). Hereby, the daily rain gauge data from 833 weather stations across China, spanning the period from 1951 to 2006, was gridded to 1° grid resolution using the method described in Milly and Dunne (2002) and Xia (2008), which considered the topographic effects on rainfall. This spatial resolution was conditioned by the distribution of weather stations in western China, shown in Fig. 17 for the Yellow River Basin. Furthermore, the daily gridded precipitation was downscaled to hourly by using a random statistical function that is the same over a calendar month (Waichler and Wigmosta 2003). The selection of this gridded precipitation observation product was made based on its comprehensive coverage of data from the majority of meteorological stations during the time period considered and the corresponding availability of hourly temporal resolution. These features are essential for conducting large-scale hydrological model simulations and facilitating climate model diagnostics.

2.2.3 Gauge-corrected reanalysis precipitation products

WATCH Forcing Data methodology applied to ERA5 (WFDE5) is derived from ERA5 with the methodology used to derive the widely employed water, energy and climate change (WATCH) forcing data. It is often used as a meteorological forcing dataset for land surface and hydrological models (Cucchi et al. 2020). The referenced reanalysis consists of a bias-corrected reconstruction of near-surface meteorological variables based on ERA5 (Hersbach et al. 2020).

Precipitation fields in the WFDE5 have been adjusted using monthly-scale bias corrections based on Climatic Research Unit (CRU) and Global Precipitation Climatology Centre (GPCC) data. Two different precipitation products (WATCH/CRU and WATCH/(CRU+GPCC)) come from WFDE5, one corrected using the CRU TS product and the other corrected using both the CRU TS and the GPCC observational products.

The China Meteorological Forcing Dataset (CMFD) is a high spatiotemporal resolution gridded near-surface

meteorological dataset that was developed for studies on land surface processes (He et al. 2020). This dataset combines remote sensing and reanalysis products with in situ observations from weather stations. Precipitation fields in the CMFD are produced based on the assimilation of 753 weather stations from the CMA and gridded background data (base reanalysis) GLDAS NOAH10SUBP 3H for the research period of this work, i.e. from 1979 to 1989.

The goal of the Global Land Data Assimilation System (GLDAS) is to integrate satellite and ground-based observational data products using advanced land surface modelling and data assimilation techniques to generate optimal fields of land surface states and fluxes (Rodell et al. 2004). GLDAS 2.0 is one of the two components of the GLDAS Version 2 (GLDAS-2) dataset, which was produced by GLDAS. Moreover, GLDAS 2.0 was derived from the global meteorological forcing dataset from Princeton University (Sheffield et al. 2006). Specifically, precipitation in GLDAS 2.0 was constructed by combining global observation-based products from the Global Precipitation Climatology Project (GPCP) daily product, the Tropical Rainfall Measuring Mission (TRMM, in operation from 1997 to 2015) 3-hourly real-time dataset and the National Centers for Environmental Prediction–National Center for Atmospheric Research (NCEP/NCAR) reanalysis.

2.2.4 Streamflow dataset

To calibrate and validate streamflow, observed daily streamflow ($\text{m}^3 \text{s}^{-1}$) for the period 1979–1987 was acquired from the National Science and Technology Infrastructure of China (<http://gre.geodata.cn>), and data came from the four main gauging stations in the basin, i.e. Tangnaihe (TNH), Lanzhou (LZ), Toudaoguai (TDG) and Huayuankou (HYK).

2.2.5 Evapotranspiration dataset

The Global Land Evaporation Amsterdam Model (GLEAM) v3.5 dataset (Martens et al. 2017) was applied to validate the model prediction of evapotranspiration. The choice of using the GLEAM data for evaluating simulated evapotranspiration is based on its robust foundation. GLEAM employs a water balance equation and assimilates multiple observations, including surface net radiation, near-surface air temperature and surface soil moisture, which collectively provide a comprehensive picture of the land surface dynamics (Martens et al. 2016, 2017). GLEAM's evapotranspiration estimates offer insights into real-world evapotranspiration, including the effects of irrigation.

3 Methods

3.1 The offline mode of the coupled Atmospheric and Hydrological Modelling System integrated IRRigation module (AHMS-IRR)

The Atmospheric and Hydrological Modelling System (AHMS) is a fully coupled atmospheric and hydrological modelling system (Jiang et al. 2022, 2020; Xia 2019; Xia et al. 2022) that combines the Weather Research and Forecasting Model, WRF (Skamarock and Klemp 2008), with a physically-based distributed regional hydrological model, HMS (Yu et al. 2006), through the land surface model NoahMP-LSM (Chen and Dudhia 2001; Niu et al. 2011). The coupling method developed in WRF-Hydro (Gochis et al. 2020) for downscaling and upscaling the variables associated with the land surface and hydrological models was adopted. AHMS can either be run offline by using prescribed near-surface atmospheric forcing variables or coupled with the WRF model. A schematic illustration of the online and offline AHMS versions is shown in Fig. 2. The near-surface atmospheric forcing data required to run the offline AHMS include incoming shortwave and longwave radiation, near-surface specific humidity, air temperature and wind, precipitation and surface pressure (see Table S1 in the Supplement).

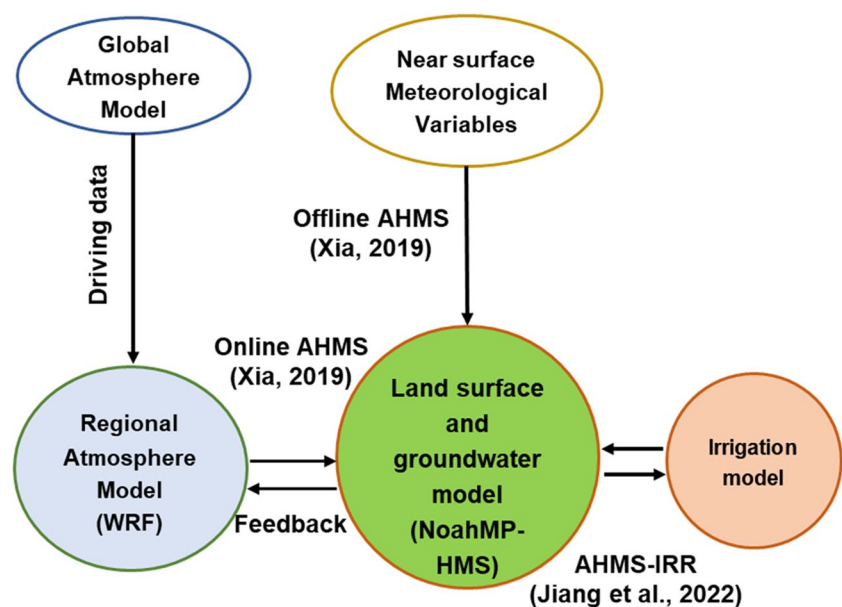
The AHMS-IRR used here was developed by Jiang et al. (2022) and incorporates a new irrigation module within the land surface, flow routing and groundwater models of the AHMS, allowing for the explicitly representation of irrigation processes in the water cycle. The irrigation amount was calculated based on the soil moisture deficits

method, water availability in rivers or lakes and irrigation fraction dataset (see Section S1 in the Supplement for more details). Figure 3 shows a schematic representation of the hydrological cycle in AHMS-IRR. In previous work, we have shown that simulations using the offline AHMS-IRR yield good agreement between predicted and observed streamflow in the Yellow River Basin for the time period considered (see Fig. S1 in the Supplemental Material of the present manuscript). Moreover, this agreement is not satisfactory if irrigation is neglected in our simulations (Jiang et al. 2022), as also shown in Fig. S1 of the Supplemental Material. By incorporating the irrigation into the simulation, a more realistic hydrological response was obtained near the outlet of the Yellow River Basin compared to previous studies (Cong et al. 2009). The model results thus demonstrate the ability of AHMS-IRR to reproduce long-term hydrological processes in the Yellow River Basin, provided that water taken for irrigation is included in the simulation.

3.2 Hydrological model setup

The AHMS-IRR model was configured for the Yellow River Basin with a temporal resolution of one hour and a spatial resolution of 20 km (Jiang et al. 2022). The key input static parameters, including vegetation type and soil texture data, are aggregated from the 30 arc-second data of the hybrid State Soil Geographic Database Food and Agriculture Organization (STATSGO/FAO) soil texture dataset and the USGS 24-category vegetation (land use) datasets. The performance of AHMS-IRR was rigorously assessed through a comprehensive evaluation, incorporating a combination of

Fig. 2 Schematic diagram of the AHMS-IRR which integrates the Atmospheric and Hydrological Modelling System (AHMS) with a new irrigation model (Jiang et al. 2022)



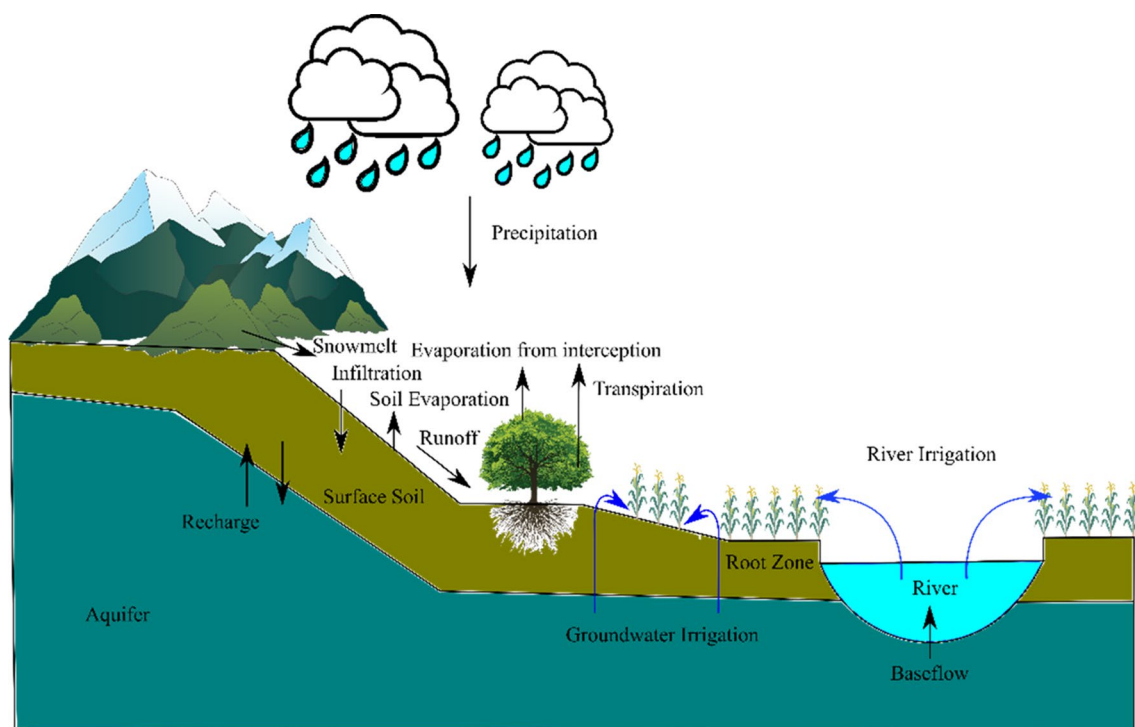


Fig. 3 Sketch of the hydrologic cycle simulated in AHMS-IRR, adopted from Jiang et al. (2022)

ground-based (in situ) and remote-sensing data sources. This evaluation encompassed various factors, including observed streamflow, terrestrial water storage data derived from the Gravity Recovery and Climate Experiment (GRACE) and assimilated evaporation data provided by the Global Land Evaporation Amsterdam Model (GLEAM). Based on sensitivity analysis conducted in Jiang et al. (2022), the most crucial parameters for calibrating the model against observations were identified as soil saturated hydraulic conductivity in the rainfall-runoff model, Manning's roughness coefficient in the flow routing model, irrigation-triggered soil moisture availability and leaf area index in the irrigation model. To effectively achieve our objective of evaluating various precipitation products using hydrological modelling in this study, we focused our calibration efforts solely on the rainfall-runoff model parameter known as soil-saturated hydraulic conductivity. We incorporated calibrated values of Manning's roughness coefficient, irrigation-induced soil water availability and leaf area index from the study of Jiang et al. (2022).

Furthermore, following the approach of Bitew et al. (2012) and Tang et al. (2016), we designed two scenarios to distinguish between uncertainties related to rainfall products and those related to models. For Scenario 1, the rain-runoff model parameters were calibrated using gauged precipitation and remained constant for the other precipitation products. For Scenario 2, the rain-runoff model parameters were recalibrated for each precipitation product.

Table 2 lists the two hydrological settings used for model calibration or validation. In Scenario 1, the offline AHMS-IRR was calibrated for gauged precipitation from 1 Jan 1979 to 31 Dec 1983. The model was then validated using the same parameters (Para 1) for each of the six (non-) gauge-corrected precipitation reanalysis products described above from 1 Jan 1984 to 31 Dec 1987.

In Scenario 2, the offline AHMS-IRR was recalibrated for each precipitation product, and the set of parameters was assigned corresponding names (Para 1 to Para 7). For scenarios 1 and 2, the performance of six (non-) gauge-corrected precipitation reanalysis products was assessed by comparing simulated and observed streamflow from major hydrological stations during the validation period. Moreover, it should be noted that the other required near-surface atmospheric forcing variables (as shown in Table S1), except for precipitation, were the same for each setup and were derived from ERA5 in Scenarios 1 and 2.

Furthermore, Table S3 in Section S5 of the Supplement shows the calibrated parameters (soil saturated hydraulic conductivity) of the rainfall-runoff model in the four subbasins, including Headwater-Tangnaihe, Tangnaihe-Lanzhou, Lanzhou-Toudaoguai and Toudaoguai-Huayuankou. Specifically, we calibrated the rainfall-runoff model against the mean annual observed runoff in the calibration period (from 1979 to 1983). The parameters in the rainfall-runoff model were calibrated here to keep the BIAS (see next section)

Table 2 Two scenarios were designed to evaluate different precipitation products using the offline AHMS-IRR

Scenario	Setup	Calibration or validation	Period	Precipitation product	Parameter set
I	1	Calibration	1 Jan 1979–31 Dec 1983 (same as below)	OBS	Para 1
	1–7	Validation	1 Jan 1984–31 Dec 1987 (same as below)	All products	
II	1	Calibration	-	OBS	Para 1
		Validation	-		
	8	Calibration	-	CMFD	Para 2
		Validation	-		
	9	Calibration	-	GLDAS 2.0	Para 3
		Validation	-		
	10	Calibration	-	WFDE5/CRU	Para 4
		Validation	-		
	11	Calibration	-	WFDE5/(CRU+GPCC)	Para 5
		Validation	-		
	12	Calibration	-	NCEP/NCAR	Para 6
		Validation	-		
	13	Calibration	-	ERA5	Para 7
		Validation	-		

of runoff for each reanalysis product as small as possible. Moreover, Table S4 in Section S5 of the Supplement presents the simulated and observed mean annual area runoff from 1979 to 1987, including the calibration and validation periods in the Yellow River Basin. Our choice for the simulation period from 1979 to 1987, with the first five years dedicated to calibration and the last four years for validation, was determined in view of the following constraints:

- (1) Meteorological Data Availability: Many near-surface meteorological products begin in 1979, making it a practical starting point for our simulation and analysis.
- (2) Human Activities: The construction of the Longyangxia reservoirs (with a storage capacity of $27.6 \times 10^9 \text{ m}^3$) in the 1980s significantly altered the Yellow River’s hydrology. Since the AHMS-IRR does not currently account for reservoirs and dams, and given the absence of comprehensive reservoir management data, we intentionally focused on the pre-1990 period to avoid the influence of these anthropogenic influences.
- (3) Limited Streamflow Data: Availability of streamflow observation data along the Yellow River is constrained, further impacting our choice of the simulation period.

3.3 Evaluation metrics

The agreement between the model-predicted and observed values was quantified by employing the following

performance metrics: the Nash–Sutcliffe model efficiency coefficient (NSE) (Nash and Sutcliffe 1970), the root-mean-square error (RMSE), the Pearson correlation coefficient (PCC), the mean absolute error (MAE), the mean absolute percentage error (MAPE), the relative bias in percent (BIAS) and the modified Kling-Gupta efficiency (mKGE) (Gupta et al. 2009; Kling et al. 2012). These performance metrics are defined as follows, and the value ranges and perfect agreement corresponding values are summarized in Table S2 in the Supplement.

$$NSE = 1.0 - \frac{\sum_{i=1}^N (P_s^i - P_o^i)^2}{\sum_{i=1}^N (P_o^i - \bar{P}_o)^2} \tag{1}$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (P_s^i - P_o^i)^2} \tag{2}$$

$$PCC = \frac{\sum_{i=1}^N (P_o^i - \bar{P}_o)(P_s^i - \bar{P}_s)}{\sqrt{\sum_{i=1}^N (P_s^i - \bar{P}_s)^2 (P_o^i - \bar{P}_o)^2}} \tag{3}$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |P_o^i - P_s^i| \tag{4}$$

$$MAPE = \frac{1}{N} \sum_{i=1}^N \left| \frac{P_o^i - P_s^i}{P_o^i} \right| \times 100\% \tag{5}$$

$$BIAS = \frac{\sum_{i=1}^N (P_s^i - P_o^i)}{\sum_{i=1}^N P_o^i} \times 100\% \tag{6}$$

$$mKGE = 1 - \sqrt{(PCC - 1)^2 + (\beta - 1)^2 + (\gamma - 1)^2}$$

$$\beta = \text{bias ratio} = \frac{\mu_s}{\mu_o}$$

$$\gamma = \text{variability ratio} = \frac{\sigma_s/\mu_s}{\sigma_o/\mu_o} \tag{7}$$

where P_s and P_o are the simulated and observed values of the P variable, respectively, and $\overline{P_o}$ is the mean value of P_o . Moreover, μ_s and μ_o represent the averages of the predicted and observed values, respectively, while σ_s and σ_o correspond to the standard deviations of the predicted and observed values, respectively.

3.4 Standard deviation of elevation

The standard deviation of elevation is used to depict the complexity of the terrain and is defined as follows:

$$SDE = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}} \tag{8}$$

where SDE is the standard deviation of elevation of the larger grid, which is at a 20 km resolution in this study, x_i is the elevation of the i th sub-grid at a 1 km resolution in the cell, N is the number of sub-grids at a 1 km resolution in the larger cell and μ is the average elevation of the cell. Here, a

moving window of 20×20 was used to calculate the standard deviation of elevations. Values close to zero indicate no variation (i.e. flat areas), while high standard deviations indicate areas with complex and steep terrain.

4 Results

4.1 Performance evaluation on the monthly, seasonal and annual scales using gauge observations

In this section, we evaluated the performance of six reanalysis precipitation products, including NCEP/NCAR, GLDAS 2.0, ERA5, WFDE5/CRU, WFDE5/(CRU+GPCC) and CMFD, across annual, seasonal and monthly timeframes by comparing them with gridded gauge precipitation (OBS). For this comparative analysis, we obtained monthly, seasonal and annual precipitation data by resampling the gridded daily precipitation data at a 20 km resolution using the mean method.

Figure 4 displays the average yearly precipitation of each product along with the observed annual precipitation. As shown, the non-gauge-corrected reanalysis precipitation products (NCEP/NCAR and ERA5) significantly overestimated precipitation by 50% and 45%, respectively. Furthermore, the CMFD results were almost identical to the observed precipitation on an annual scale.

Moreover, Table 3 lists the annual precipitation statistics of seven products from 1979 to 1989. As shown, NCER/NCAR yielded the highest average annual precipitation (664 m), had the widest range in the annual mean precipitation (514–772 mm) and was associated with the largest

Fig. 4 Annual precipitation in the Yellow River Basin from 1979 to 1989: NCEP/NCAR (brown dashed line), OBS (orange solid line), GLDAS 2.0 (orange dashed line), ERA5 (green dashed line), WFDE5/CRU (blue dashed line), WFDE5/(CRU+GPCC) (pink dashed line) and CMFD (green dotted line)

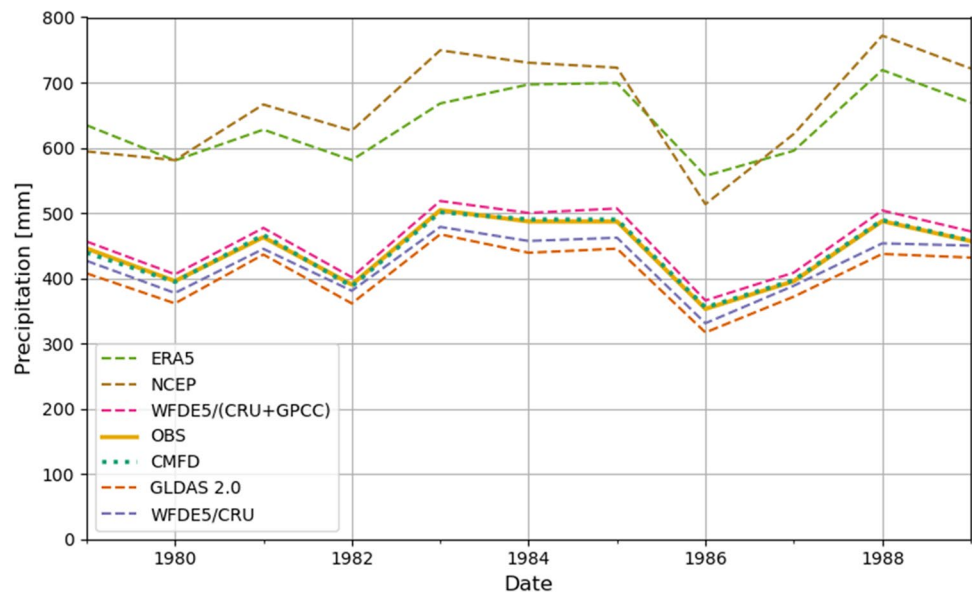


Table 3 Annual precipitation statistics for the Yellow River Basin from 1979 to 1989 (unit: mm), associated with the different products considered in this study

Index	Minimum	Median	Mean	Maximum	Standard deviation
OBS	353	457	442	504	50.6
NCEP/NCAR	514	666	664	772	82.3
ERA5	557	635	639	719	55.4
GLDAS 2.0	317	432	407	467	47.0
WFDE5/CRU	331	445	423	479	46.3
WFDE5/(CRU+GPCC)	366	472	456	518	52.2
CMFD	355	457	443	501	51.0

The maximum and minimum standard deviation are highlighted in the context

deviation (82.3 mm) over the Yellow River Basin. To compare, WFDE5/CRU had the narrowest range in the annual mean precipitation (331–479 mm) as well as the smallest associated deviation (46.3 mm). As shown in Fig. 5, the annual precipitation density from the CMFD yielded the best agreement with the gauge data (OBS), followed by the WFDE5/(CRU+GPCC). Moreover, WFDE5/CRU and GLDAS 2.0 yielded an overall low precipitation level with a peak of approximately 450 mm. Furthermore, the non-gauge-corrected NCEP/NCAR and ERA5 products significantly overestimated the precipitation levels in the study period.

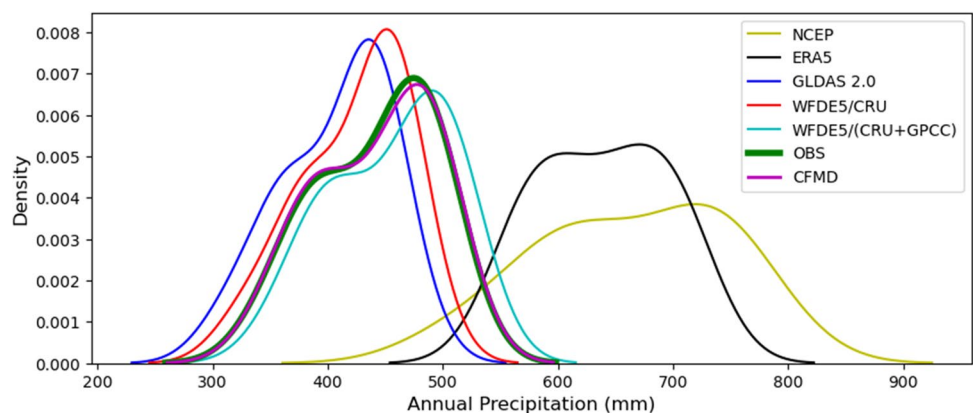
Figure 6 displays the spatial distributions of the precipitation from all products and the Pearson correlation coefficients (PCCs) for the six reanalysis products in the Yellow River Basin (upland of Huayuankou) from 1979 to 1989. This figure shows that the average annual precipitation gradually increased from northwest to southeast. The areas with the highest annual precipitation were located southern Yellow River Basin, including the upper reaches of this basin, the Weihe Basin and the Sanmenxia to Huayuankou area (see Fig. 1). In contrast, the lowest annual precipitation occurred at Ningxia and the Hetao Plateau in the northwest of the basin. According to the PCC and RMSE, precipitation data from WFDE5/(CRU+GPCC) (PCC=0.99 and RMSE=0.08) and CMFD (PCC=0.98 and RMSE=0.09) were the most consistent with OBS. GLDAS 2.0 and WFDE5/CRU slightly underestimated precipitation in

the upper reaches of the river, while the non-gauge-corrected precipitation products (NCEP/NCAR and ERA5) significantly overestimated precipitation, especially in the mountainous area of the basin. For further details and interpretations, please refer to the discussion in Sect. 5a.

To evaluate extreme precipitation, we calculated the 99th percentile of the daily precipitation for each precipitation product. The results are presented in Fig. 7. Based on the RMSE and PCC values calculated over the research area, the spatial distribution of extreme precipitation (i.e. 99th percentile precipitation) predicted from WFDE5/(CRU+GPCC) (PCC=0.97 and RMSE=1.81) and CMFD (PCC=0.96 and RMSE=2.13) exhibited the best agreement with the observations (OBS). Remarkably, although NCEP/NCAR and ERA5 exhibited similar annual precipitation, their spatial patterns of extreme precipitation showed significant differences, as indicated by the RMSE values of 8.09 and 4.13 and the PCC values of 0.21 and 0.93 for NCEP/NCAR and ERA5, respectively. As illustrated in Fig. 7, the statistical comparison suggests that all gauge-corrected reanalysis precipitation products underestimate extreme rainfall to various degrees, whereas non-gauge-corrected reanalysis precipitation products (ERA5) tend to overestimate extreme rainfall.

Figure 8 shows the monthly precipitation of the products in the Yellow River Basin from 1979 to 1989 and the average annual cycle of monthly precipitation for this

Fig. 5 Density curve (kernel density estimates KDE, which is the estimation of probability density functions of random variables based on kernels as weights) of annual precipitation of the products over the Yellow River Basin from 1979 to 1989



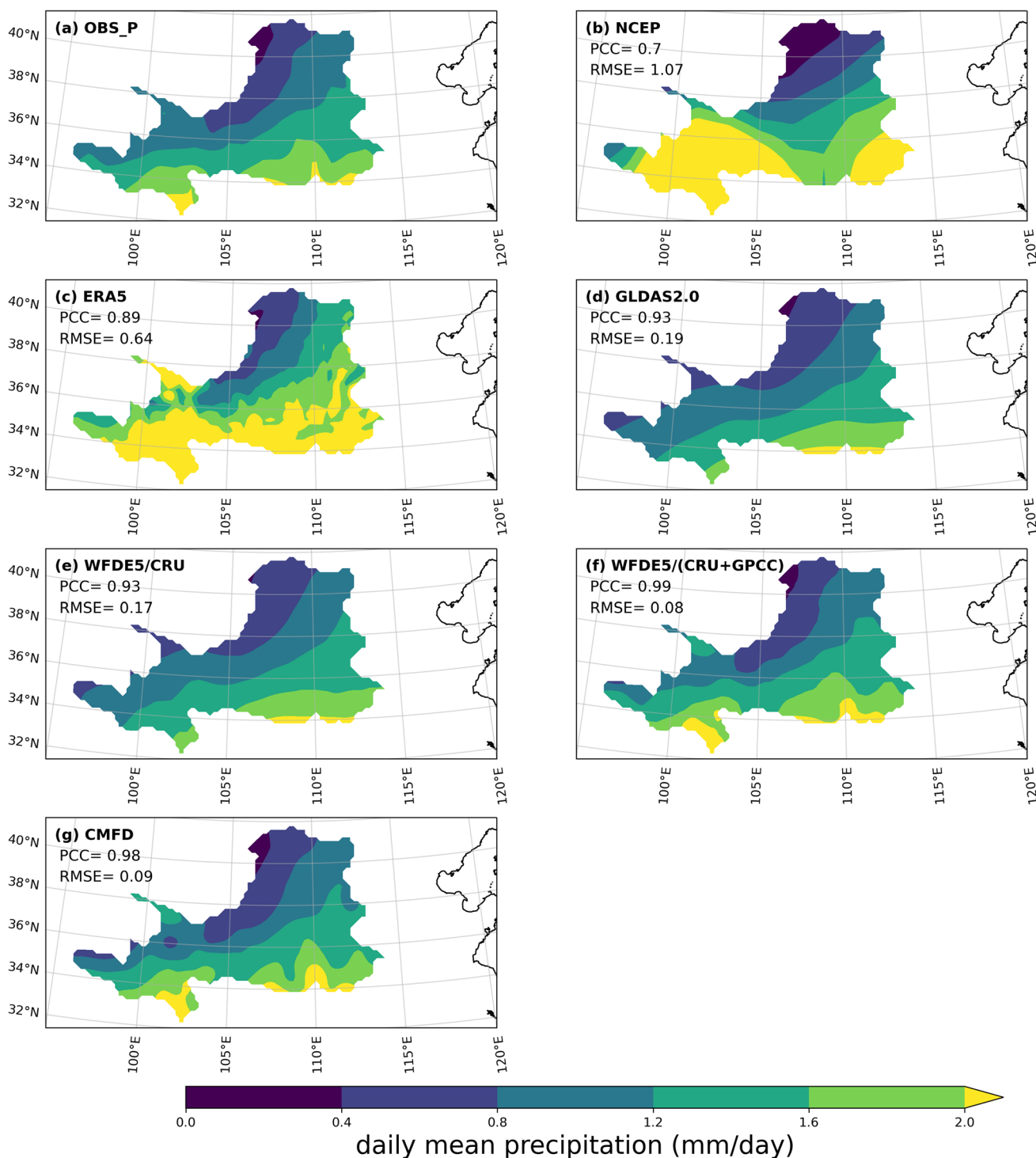


Fig. 6 Spatial distribution of the daily mean precipitation of the seven precipitation products considered in the present study, i.e. (a) OBS, (b) NCEP/NCAR, (c) ERA5, (d) GLDAS 2.0, (e) WFDE5/CRU, f WFDE5/(CRU+GPCC) and (g) CMFD, from 1979 to 1989 over the

Yellow River Basin (upland of Huayuankou), along with the root-mean-square error (RMSE, mm) and the Pearson correlation coefficient (PCC) between the reanalysis product and OBS for each grid in the basin

period. As shown in Fig. 8, NCEP/NCAR and ERA5 overestimated precipitation in all months, while all other products did not differ much in terms of monthly precipitation. Table 4 lists the monthly precipitation statistics of the six

reanalysis precipitation products. These monthly statistics showed that CMFD yielded the highest correlation coefficient. Similarly, CMFD attained the lowest MAE, MAPE and RMSE.

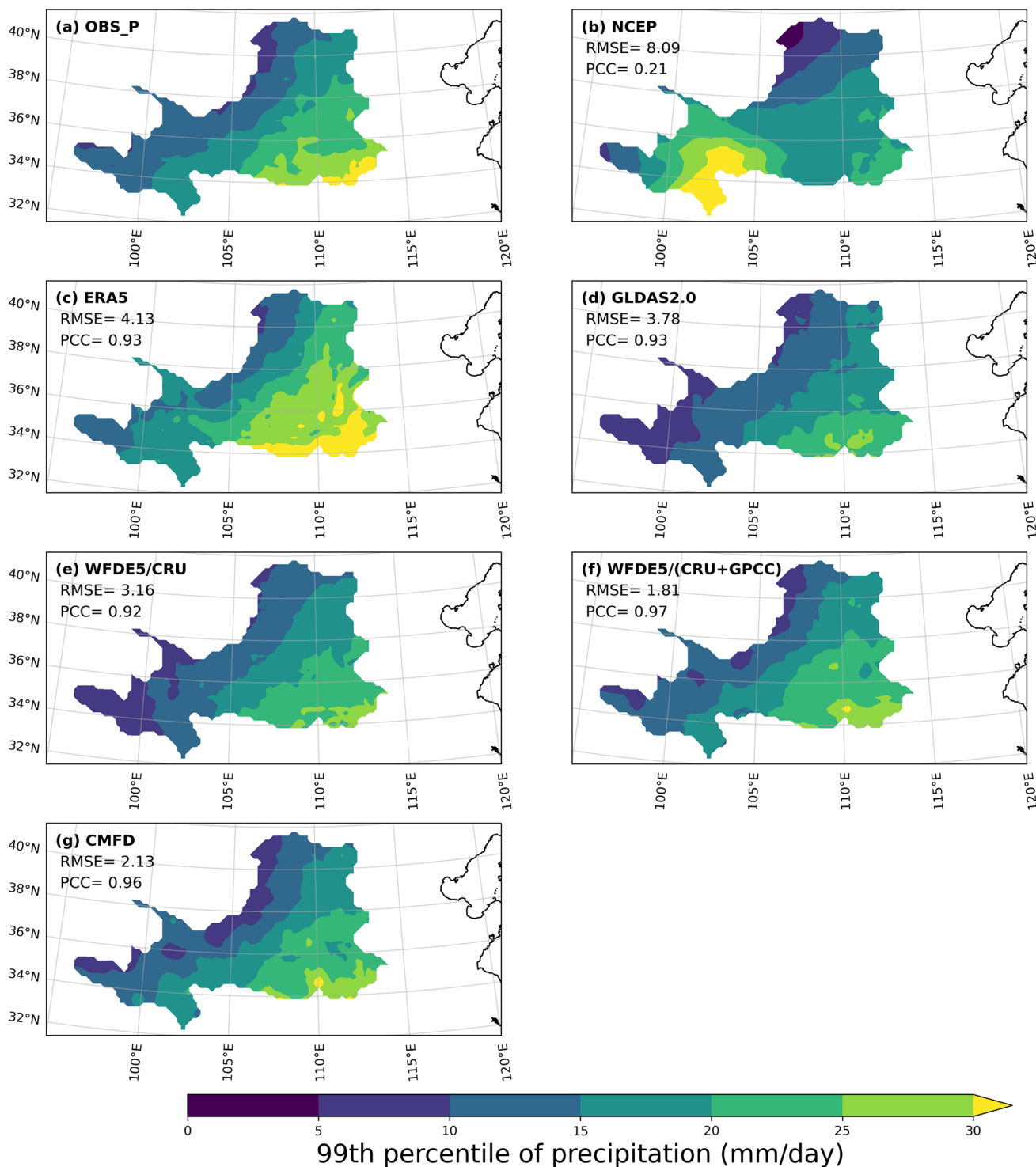


Fig. 7 Spatial distribution of the 99th percentile precipitation of the seven precipitation products considered in the present study, i.e. (a) OBS, (b) NCEP/NCAR, (c) ERA5, (d) GLDAS 2.0, (e) WFDE5/CRU, (f) WFDE5/(CRU+GPCC) and (g) CMFD, from 1979 to 1989

over the Yellow River Basin (upland of Huayankou), along with the root mean square error (RMSE, mm) and the Pearson correlation coefficient (PCC) between the reanalysis product and OBS for each grid in the basin

Figure 9 shows the average monthly evaluation metrics of the precipitation products in the Yellow River Basin from 1979 to 1989. As seen, Fig. 9 indicates that CMFD had the lowest

RMSE, MAE and MAPE, while NCEP/NCAR performed the worst. Compared to other months, all data showed the most considerable error in August. On the contrary, Fig. 9d shows

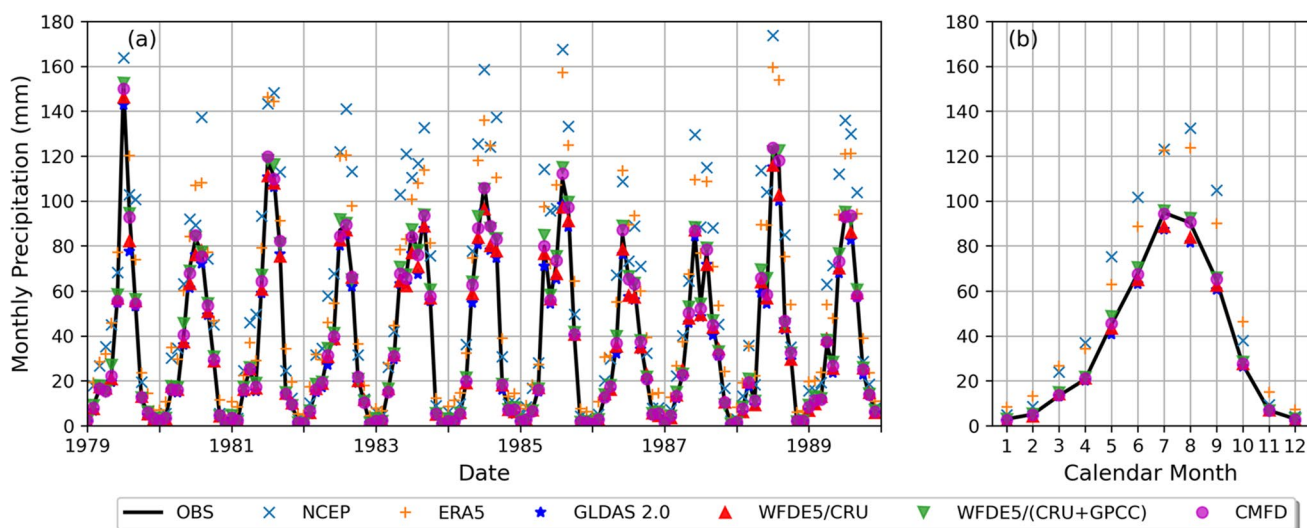


Fig. 8 (a) Monthly precipitation of the products over the Yellow River Basin from 1979 to 1989. (b) The annual cycle of monthly precipitation averaged over 1979–1989

Table 4 Statistics of the monthly precipitation of the products over the Yellow River Basin from 1979 to 1989 (unit: mm)

Index	PCC	RMSE	MAE	MAPE
NCEP/NCAR	0.976	25.18	18.43	61.80
ERA5	0.993	19.11	16.39	104.44
GLDAS	0.998	4.53	3.02	13.64
WFDE5/CRU	0.997	0.99	2.18	10.40
WFDE5/(CRU+GPCC)	0.999	2.02	1.34	5.68
CMFD	1.000	0.88	0.61	4.20

The maximum and minimum standard deviation are highlighted in the context

that the percentage error in precipitation for these products was significantly higher in December, January and February.

Similarly, Fig. 10a, c, d shows that CMFD (NCEP/NCAR) produced the lowest (largest) RMSE, MAE and MAPE for seasonal precipitation, while all products led to an overestimation of this precipitation in the summer. Figure 10d shows that the percentage error in precipitation for these products increased significantly in the winter, especially for NCEP/NCAR and ERA5. This finding is consistent with Beck et al. (2017) and Jiang et al. (2021), as it occurs because the atmospheric model produces spurious drizzle and tends to overestimate low rainfall due to deficiencies in rainfall generation parameterization scheme. For further insights and interpretations, please refer to Sect. 5a.

4.2 Performance evaluation using the offline AHMS-IRR

In this section, we evaluated the performance of the six reanalysis precipitation data products by comparing the

offline AHMS-IRR simulated monthly streamflow with the corresponding observations. Figure 11 compares simulated monthly streamflow with gauge-calibrated parameters and observed monthly streamflow in both the calibration and the validation periods at all four main hydrological stations (Tangnaihe, Lanzhou, Toudaoguai and Huayuankou). As seen, the monthly streamflow was significantly overestimated when using NCEP/NCAR and ERA5 products and slightly underestimated when using gauge-corrected reanalysis precipitation products. The Nash–Sutcliffe model efficiency coefficient (NSE) and the Kling–Gupta efficiency (KGE) for the validation periods are displayed in Fig. 12, and additional data for both calibration and validation periods can be referenced in Table S5 in the Supplement. As shown in Fig. 12, the model performed much better upstream than in the middle stream area. The disparities in observed and simulated streamflow may be linked to inadequate representations of the precipitation patterns of these products. Notably, the model's performance appears to degrade as we move downstream, implying that errors accumulate along the river's course.

In particular, the highest values of NSE (0.82) and KGE (0.82) were observed at the Tangnaihai station in the validation period (from 1984 to 1987). Moreover, WFDE5/(CRU+GPCC) and CMFD outperformed the other products during the validation period, with the highest mean NSE and KGE values of the four stations, consistent with our statistical comparison discussed in the previous text.

Figure 13 compares simulated and observed monthly streamflow, where the simulated values were obtained from the hydrological model run with recalibrated parameters. Furthermore, Fig. 14 and Table S6 in the Supplement show the model performance (NSE and KGE) associated with the

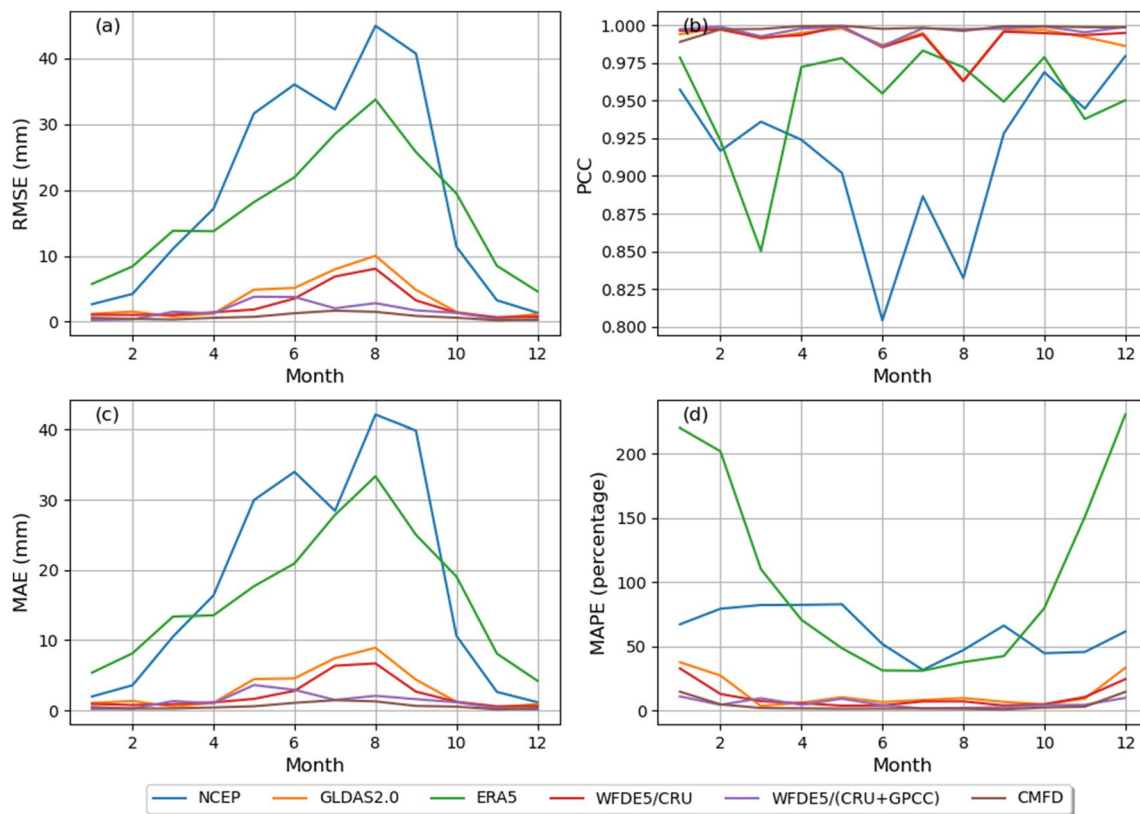


Fig. 9 Averaged monthly evaluation metrics of the precipitation products over the Yellow River Basin from 1979 to 1989: (a) RMSE (mm), (b) PCC, (c) MAE (mm) and (d) MAPE (%)

different precipitation products in Scenario 2. The recalibration of the rainfall-runoff model substantially improved the model performance for all precipitation products compared to the results of Scenario 1 (Fig. 12). The improvement in model performance was particularly notable for models using gauge-corrected reanalysis precipitation products. For the validation period considered in the present work, the gauge-corrected reanalysis precipitation products, including WFDE5/(CRU+GPCC) and CMFD, demonstrated slightly better performance than other gauge-corrected products. However, non-gauge-corrected reanalysis precipitation products such as NCEP/NCAR and ERA5 exhibited poorer performance than gauge-corrected products. Overall, the performance of all gauge-corrected reanalysis precipitation products was similar.

Furthermore, we found the discrepancies in the pattern of extreme rainfall events drive variations in calibrated soil parameters with different precipitation products. By comparing Table S2 and Fig. 7, we observed a clear correlation between extreme rainfall events and the calibrated soil parameters. Specifically, in areas where extreme rainfall was underestimated by the precipitation product, the calibration process led to a reduction in soil saturated hydraulic conductivity. This reduction compensated for the underestimated

extreme rainfall rates, aligning the model's surface runoff with observed data. This dependency illustrates that the accuracy of rainfall products to depict extreme rainfall events is critical for hydrological modelling.

Figure 15 displays the performance of the offline AHMS (RMSE and NSE) in predicting evapotranspiration by comparing it with the GLEAM results (Martens et al. 2017). The analysis is associated with the different precipitation products used in Scenarios 1 and 2. The results suggest that simulations driven by the non-gauge-corrected precipitation products significantly overestimated evapotranspiration. Moreover, the recalibration of the rainfall-runoff model led to an increase in runoff and an improvement in the evapotranspiration simulation for the gauge-corrected precipitation products, whereas it did not produce similar effects for the non-gauge-corrected precipitation products. Therefore, both Figs. 14 and 15 reveal that when the precipitation bias exceeded a specific threshold (e.g. ERA5 and NCEP/NCAR with $\text{BIAS} \geq 45\%$), the calibration of the hydrological model was no longer valid in terms of improving the model's performance.

In Scenario II, as shown in Fig. 15 and Table S7, the performance of simulated evapotranspiration exceeds the OBS data for selected precipitation products (CMFD,

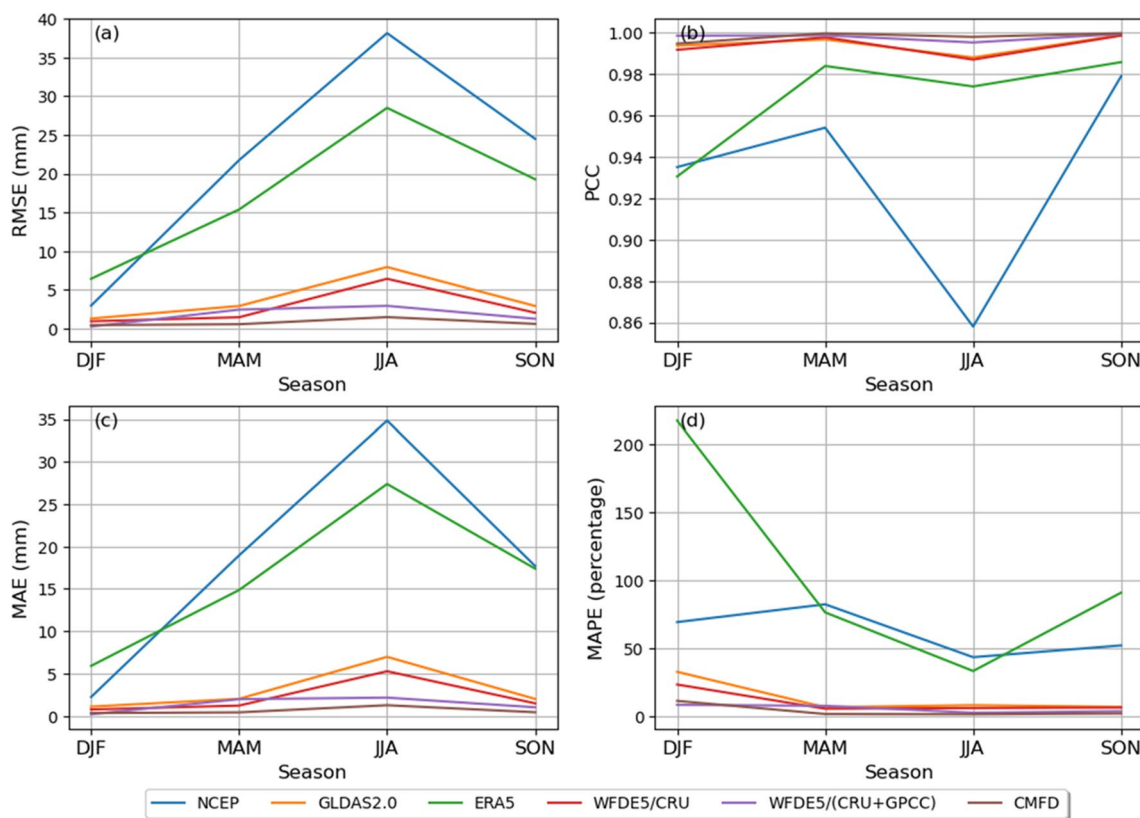


Fig. 10 Seasonal evaluation metrics of the precipitation products over the Yellow River Basin from 1979 to 1989: DJF (winter, December to February), MAM (spring, March to May), JJA (summer, June to

August), SON (autumn, September to November), (a) RMSE (mm), (b) PCC, (c) MAE (mm) and (d) MAPE (%)

GLDAS 2.0, WFDE5/CRU, WFDE5/(CRU+GPCC)). The reasons for this divergence between the AHMS-IRR simulation and the GLEAM data become apparent as we delve deeper into the analysis.

As shown in Fig. S2, the AHMS-IRR simulation driven by observed precipitation tends to slightly overestimate evapotranspiration during the summer, especially in response to extreme rainfall events. It is worth noting that, as indicated in Fig. S3, this overestimation is, in part, offset by the underestimation of precipitation during extreme rainfall events (see Fig. 7), resulting in a closer agreement with GLEAM data.

Our current model calibration focused primarily on soil parameters, specifically against runoff. We acknowledge that incorporating additional parameters and targets, such as vegetation characteristics and evapotranspiration, may further enhance model performance. Future model calibration efforts will consider these factors to achieve a more comprehensive representation of land surface processes.

5 Discussion

5.1 Why do the non-gauge-corrected reanalysis products overestimate the precipitation in the Yellow River Basin?

Many previous studies concluded that the reanalysis product overestimates the observed precipitation in areas of complex terrain, which is consistent with the present study (see Fig. 6). Amjad et al. (2020) and Izadi et al. (2021) found that ERA5 consistently overestimated the observed precipitation over the mountainous areas of Turkey and Iran, while Jiang et al. (2021) reported that ERA5 tended to overestimate light precipitation events, especially in mountainous areas.

There are two possible reasons for the overestimation of precipitation in areas with complex terrain. On the one hand, this overestimation may be due to the scarcity of weather stations in areas of complex terrain, where the limited observations do not capture precipitation patterns in sufficient detail (Jiao et al. 2021). On the other hand, the

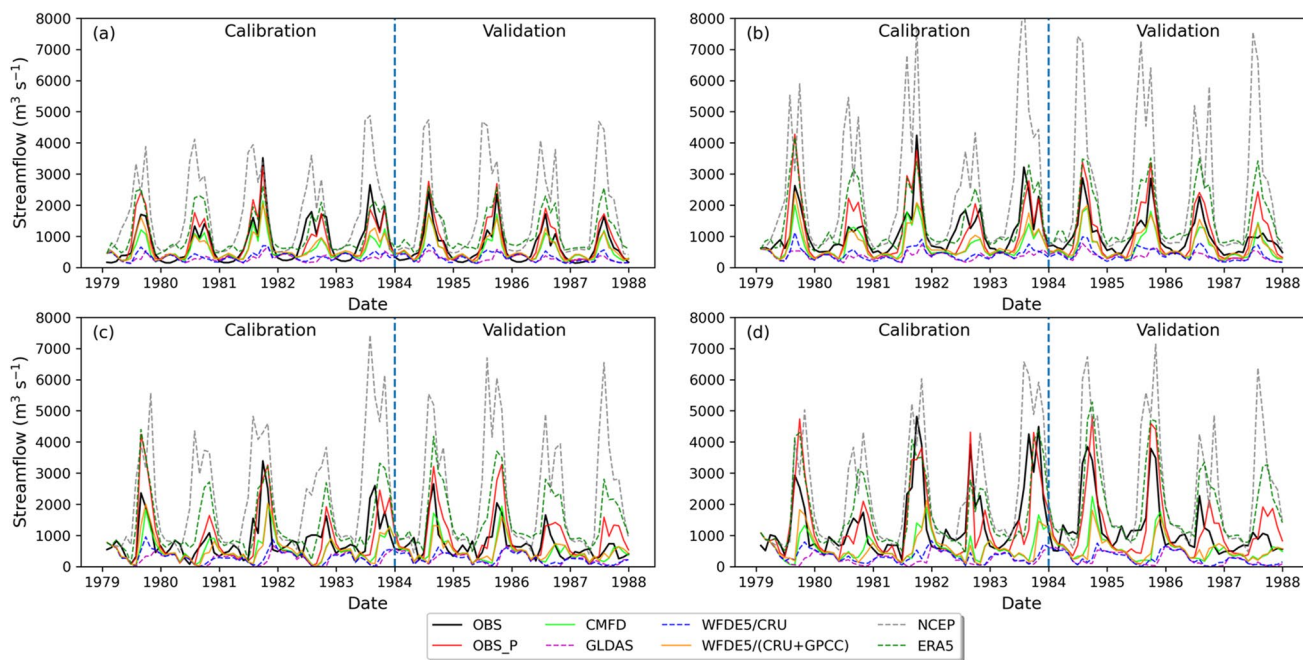


Fig. 11 Comparison of monthly streamflow simulated by offline AHMS-IRR with gauge-calibrated parameters (Scenario 1). The comparison includes seven different precipitation products against observed monthly streamflow at all four main hydrological stations, i.e. Tangnaihe (a), Lanzhou (b), Toudaoguai (c) and Huayuankou (d).

The calibration period ranges from January 1, 1979, to December 31, 1983, while the validation period spans from January 1, 1984, to December 31, 1987. Note that OBS represents the observed streamflow, while OBS_P represents the streamflow simulated using the gauged precipitation

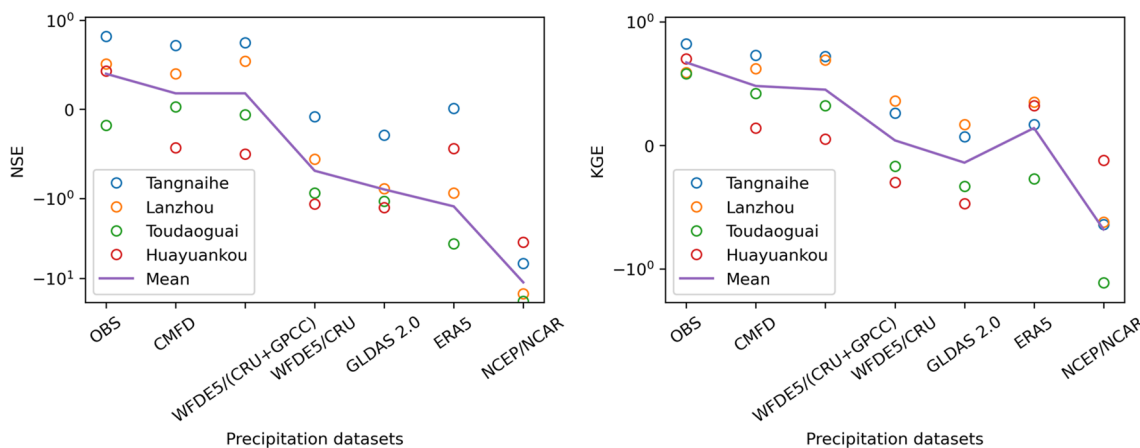


Fig. 12 Performance of streamflow simulations forced by seven different precipitation products using the static rainfall-runoff parameters during the validation period (Scenario 1)

precipitation simulations in complex terrain areas are not accurate because the resolution of the precipitation simulation is relatively coarse in areas of complex terrain. Specifically, as discussed in previous work (Amjad et al. 2020; Izadi et al. 2021), the mechanisms of humidity injection and turbulence through land surface-atmosphere interactions in complex terrain areas are poorly understood. Furthermore,

the lack of snow removal processes in the ERA5 simulation could lead to excessive snowfall and precipitation in the Tibetan Plateau region, according to Orsolini et al. (2019). These findings are consistent with the fact that the percentage errors of non-gauge-corrected precipitation products (ERA5 and NCEP/NCAR) were much larger in winter than in summer (see Figs. 9, 10 and Sect. 4.1).

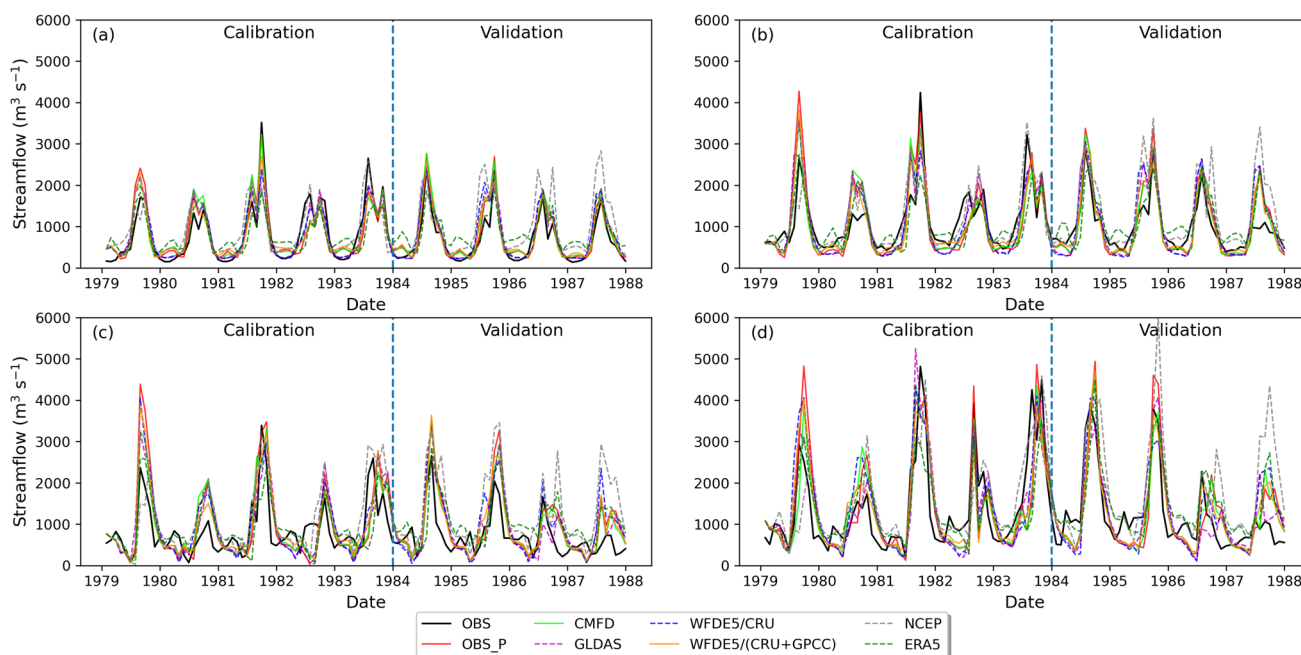


Fig. 13 Same as in Fig. 11 but with product-specific calibrated parameters for each of the seven precipitation products (Scenario 2)

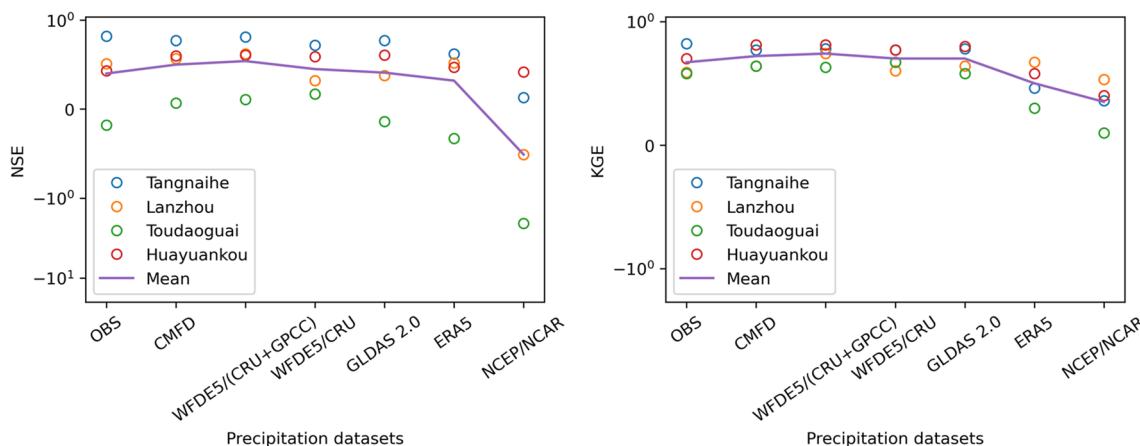


Fig. 14 Performance of offline AHMS simulations of streamflow forced by seven different precipitation products using the dynamic rainfall-runoff parameters during the validation period (Scenario 2)

To analyse the effect of the complex topography of the Yellow River Basin on the quality of the reanalysis precipitation data, Pearson correlation coefficients (PCCs) were calculated between precipitation observations and individual reanalysis precipitation using the standard deviation of elevation and mean annual precipitation over the Yellow River Basin. According to Eq. 8, the standard deviation of elevation was calculated at a 20 km resolution and was based on a digital estimation model at a 1 km resolution (USGS 1K). As shown in Fig. 16a, the Pearson correlation coefficient between the average annual observed precipitation and NCEP/NCAR reanalysis

precipitation decreased rapidly with the standard deviation of elevation, while ERA5 precipitation maintained a highly stable correlation with the observed data. This behaviour was most likely due to the higher resolution of ERA5 (2.5° of NCEP/NCAR and 0.25° of ERA5) and the advanced 4d-Var data assimilation system of ERA5. Surprisingly, as the standard deviation of elevation increased (>230 m), the Pearson correlation coefficient for ERA5 also slightly increased and fluctuated, which was also observed in the gauge-corrected reanalysis precipitation products (see Fig. 16b). As shown in Fig. 16c, this behaviour may be explained by the fact that grids with

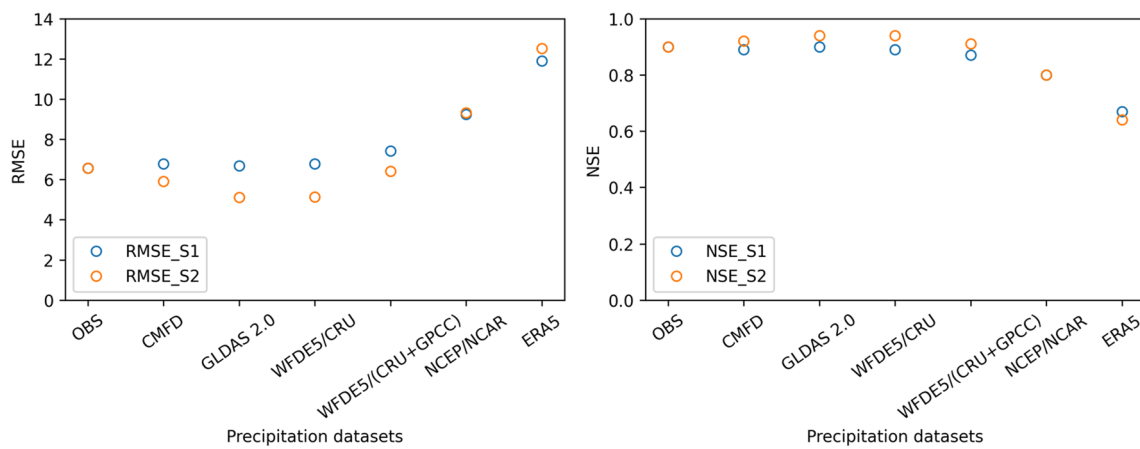


Fig. 15 Performance of offline AHMS-IRR simulations in terms of evapotranspiration forced by seven different precipitation products using the static (Scenario 1 in blue) and dynamic (Scenario 2 in yellow) rainfall-runoff parameters for the period 1980–1987

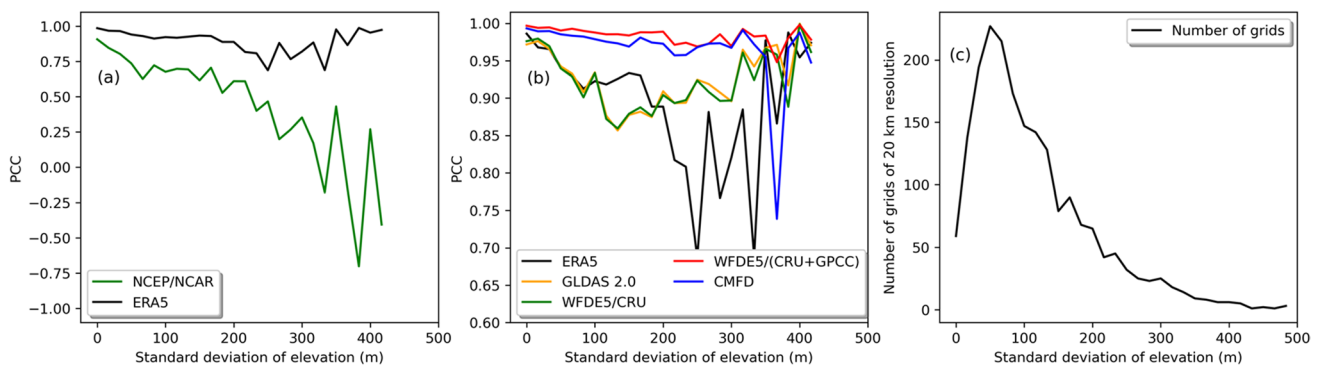


Fig. 16 Pearson correlation coefficient (PCC) between the average annual observed precipitation and six reanalysis precipitation products with a standard deviation of elevation calculated in each 20 km

grid over the Yellow River Basin (a) and (b). (c) The number of grids with a 20 km resolution and standard deviation of elevation over the Yellow River Basin

higher standard deviation elevations are scarce. Furthermore, Fig. 16b shows that the gauge-corrected reanalysis precipitation data maintained a stable correlation with the observed data as the standard deviation of elevation changed compared with the non-gauge-corrected reanalysis precipitation products. Specifically, WFDE5/(CRU+GPCC) and CMFD performed better than GLDAS 2.0 and WFDE5/CRU.

Considering the influence of complex terrain on precipitation simulations, the simulation grid for future climate modelling should be refined in areas of complex terrain to reach a pre-defined threshold of the standard deviation of elevation (Bacon et al. 2000; Dietachmayer 1992). This refinement process improves the simulation of precipitation in areas with complex terrain without wasting computing capacity (Skamarock and Klemp 2008).

5.2 Why do CMFD and WFDE5/(CRU+GPCC) agree the best with the precipitation observations?

The strong agreement between the CMFD and WFDE5/(CRU+GPCC) with precipitation observations can be attributed to several key factors discussed below.

As indicated by the comparison of annual and monthly precipitation (Figs. 4, 6, 7 and 9) and monthly streamflow (Figs. 12 and 14), CMFD performs exceptionally well and is nearly on par with the gridded rain gauge product and WFDE5/CRU+GPCC. However, its superior performance can be attributed to a combination of factors, including—in addition to its high spatial resolution (0.1°)—a substantial volume of rain gauge data, as well as the incorporation of remote sensing data. As indicated by He et al. (2020), CMFD integrates a more extensive product of weather station observations from the China Meteorological Administration

(CMA), which began sharing data through the China Meteorological Data Service Center (CMDC) in recent years.

However, when comparing the performance of CMFD and WFDE5/CRU+GPCC, it is clear that the high spatial resolution of CMFD (0.1°) does not yield a clear advantage over WFDE5/(CRU+GPCC) (0.5°). Therefore, we attribute the better performance of CMFD and WFDE5/(CRU+GPCC) to the incorporation of a significantly larger number of weather stations, as explained below.

The disparities in the performances of gauge-corrected reanalysis precipitation products are closely linked to the number of assimilated rain gauges. During the study period (1979~1989), only a few weather stations from CMA were shared globally through the Global Telecommunications System (GTS). Figure 17 shows the distribution of a total of 82 CMA weather stations in the Yellow River Basin, including 22 national benchmark climate stations shared worldwide through the GTS and the other 60 national ordinary stations not shared via the GTS. This is an important reason why the gauge-corrected reanalysis precipitation products (WFDE5/CRU and GLDAS 2.0) underestimated the annual precipitation compared with gauge observations and explained why CMFD and WFDE5/(CRU+GPCC) agreed the best with the precipitation observations.

Specifically, CMFD and WFDE5/(CRU+GPCC) employed a more significant number of stations than WFDE5/CRU and GLDAS 2.0. According to He et al.

(2020) and Yu et al. (2020), the CMFD uses approximately 753 stations across China, while WFDE5/(CRU+GPCC) essentially assimilates the same stations. In contrast, only approximately 300 stations across China, shared by the Global Telecommunications System (GTS), have been incorporated to derive the WFDE5/CRU and GLDAS 2.0 products. In this study, we found that WFDE5/CRU and GLDAS 2.0 slightly underestimated the observed annual precipitation, which was consistent with the observations of Shi et al. (2017). These authors also concluded that the CRU TS product underestimated precipitation over the three-river headwaters region of China.

5.3 What challenges do hydrological model uncertainties and anthropogenic influences pose in this study?

In this study, we employed an offline AHMS-IRR and measured water discharge to assess (non-) gauge-corrected precipitation reanalysis. Although the use of offline AHMS-IRR and measured water discharge provides us with an additional dimension for assessing precipitation, it is associated with additional uncertainties related to the parameters of the rainfall-runoff model and the flow routing model. Therefore, to reduce model uncertainty as much as possible, we recalibrated the rainfall-runoff model separately with each meteorological forcing data in the present study. In addition, we used two evaluation

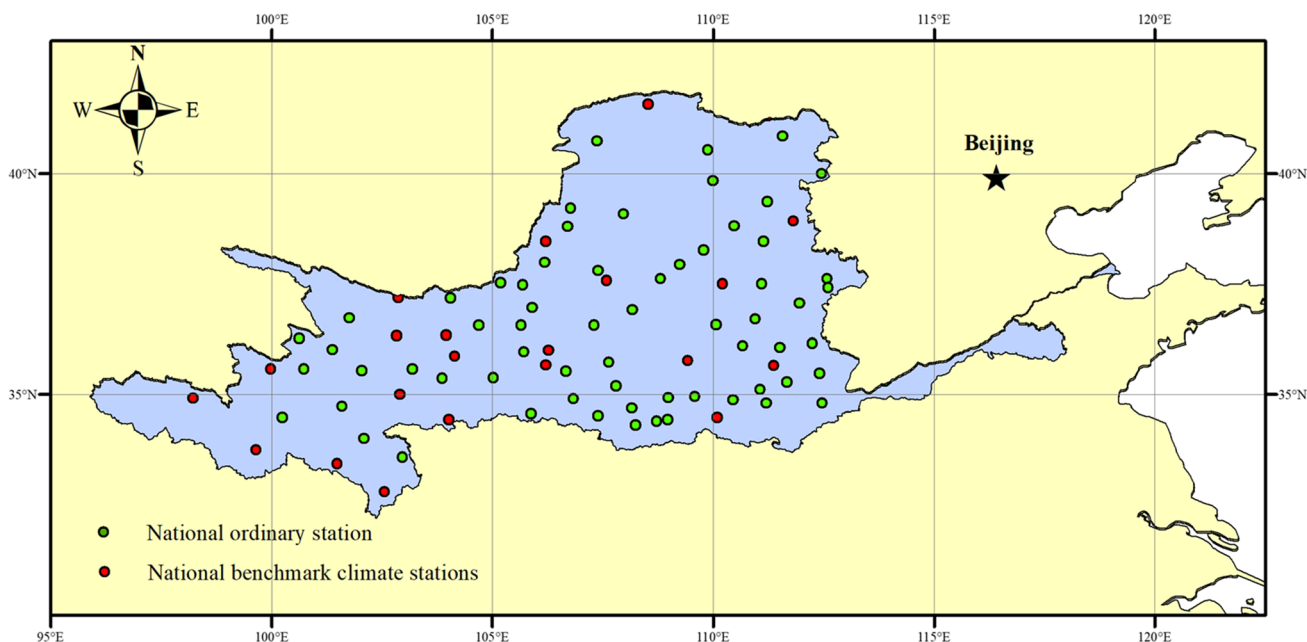


Fig. 17 Distribution of weather stations of the China Meteorological Administration in the Yellow River Basin (upstream of HYK station). Twenty-two national benchmark climate stations are shared world-

wide through the Global Telecommunication System (GTS) (red points), and the other 60 national ordinary stations (green points) are not shared through the GTS

metrics to evaluate model performance, i.e. the Nash-Sutcliffe model efficiency coefficient (NSE) and the modified Kling-Gupta efficiency (mKGE) (Gupta et al. 2009; Kling et al. 2012). Hereby, we focused on the agreements between the observed and simulated streamflow peak and correlation, bias and variability.

Furthermore, anthropogenic factors such as irrigation play an important role in hydrological processes and profoundly impact the water balance in arid and semiarid basins such as the Yellow River Basin. To quantify anthropogenic influences, we used the recently introduced and extended offline AHMS-IRR that applies to arid and semiarid regions by explicitly considering the effects of irrigation on hydrological processes (Jiang et al. 2022). The extended offline AHMS-IRR integrates a new irrigation module, which computes the dynamic irrigation demand, into the modified land surface, flow routing and groundwater models. This new extended offline AHMS-IRR explicitly represents anthropogenic processes, such as irrigation, which significantly impact the water balance in the arid and semiarid areas of the Yellow River Basin.

The pronounced discrepancy at Toudaoguai, also observed in companion papers (Jia et al. 2006; Yin et al. 2021), may arise from various factors, including the complex diversion and drainage process in the Hetao irrigation district of Inner Mongolia. Another potential contributing factor could be the reliance on a single set of empirical hydraulic geometry equations in the channel routing model of AHMS-IRR and the absence of bathymetry observation data, which may not comprehensively capture the intricate bathymetry of the Yellow River, especially in its middle and lower reaches. Furthermore, NSE is sensitive to peak flow values (Krause et al. 2005). As suggested by Gupta et al. (2009) and Kling et al. (2012), mKGE (Modified Kling-Gupta Efficiency) offers a more balanced metric, taking into account three independent criteria: correlation coefficient, bias ratio and variability ratio.

As seen in Figs. S3 and S4, Tables S8 and S9, we evaluated of the model's performance based on daily streamflow for the Tangnaihe, Lanzhou, Toudaiguai and Huayuankou stations. We indeed observe a slightly worse performance of the model results for the daily streamflow, compared to results associated with the monthly streamflow, particularly when using the NSE metric.

It is worth noting that our study area encompasses a large basin with extensive human interventions—including complex diversion and drainage processes in the irrigation district, damming and reservoir operations—which can strongly affect the behaviour of daily streamflow. Moreover, the profiles of riverbed, width and Manning roughness coefficient are poorly known, thus constraining the accuracy of flow dynamics simulation within the routing model and

limiting our ability to accurately represent daily streamflow dynamics. As mentioned earlier, our choice to employ the NSE metric is influenced by its sensitivity to peak flows, and monthly streamflow offers a more robust metric for assessing model performance across various precipitation products.

5.4 Why do the calibrated model parameters vary with the different precipitation products?

In our study, the calibration process is primarily based on matching the model-simulated surface runoff with observed data. Surface runoff is fundamentally determined by the balance between incoming water from rainfall and the soil's infiltration capacity. Detailed calculations for infiltration capacity and infiltration-excess runoff in this study can be found in Section S2 of the Supplement.

However, the following two main factors make the calibration process challenging:

- (1) Heterogeneity of rainfall and soil properties: Both rainfall and the surface soil properties exhibit spatiotemporal heterogeneity (Yu 1998). Rainfall patterns vary across time and space, and soil properties also exhibit variations across the watershed. This inherent heterogeneity complicates the calibration process, as it requires modifying the soil parameters to account for inaccuracies in the rainfall data.
- (2) Impact of extreme rainfall events: Surface runoff generation and peak river runoff are strongly influenced by extreme rainfall events. These events often play a crucial role in shaping the calibrated soil parameters. The accurate representation of rainfall extremes in precipitation products significantly impacts the magnitude of the calibrated soil parameters.

In our analysis, we examined the relationship between extreme rainfall events and the calibrated soil parameters. By comparing Table S2 and Fig. 7, we observed a clear correlation. Specifically, in areas where extreme rainfall was underestimated by the precipitation dataset, the calibration process led to a reduction in soil saturated hydraulic conductivity. This reduction compensated for the underestimated extreme rainfall rates, aligning the model's surface runoff with observed data. This dependency illustrates that the accuracy of rainfall products to depict extreme rainfall events is critical for hydrological modelling.

In short, we found that the discrepancies in the pattern of extreme rainfall events drive variations in the calibrated soil parameters with different precipitation products. The calibration process aims to ensure that the model accurately reproduces surface runoff, and variations in calibrated soil

properties are a means to achieve this goal, considering the challenges posed by heterogeneity in both rainfall and soil characteristics.

6 Summary and conclusions

In this study, we evaluated the performance of six widely used precipitation reanalysis products in the hydrology community for the Yellow River Basin, focusing on their accuracy against observed data and their influence on streamflow predictions using offline AHMS-IRR. The evaluated products included both non-gauge-corrected (NCEP/NCAR and ERA5) and gauge-corrected reanalysis products (GLDAS 2.0, WFDE5/CRU, WFDE5/(CRU+GPCC) and CMFD). We conducted a comprehensive evaluation based on annual, seasonal and monthly precipitation observation statistics, alongside a comparison of streamflow observations at four main hydrological stations along the Yellow River with corresponding predictions using a recently introduced regional hydrological model (Jiang et al. 2022). Each precipitation product served as the meteorological input for offline AHMS-IRR simulations, considering two model setups with static and dynamic parameters of the rainfall-runoff model. Our key findings can be summarized as follows:

- 1) CMFD and WFDE5/(CRU+GPCC) demonstrated the best agreement with observed precipitation and streamflow through statistical and hydrological comparisons in the Yellow River Basin. This superior performance can be attributed to CMFD and WFDE5/(CRU+GPCC) assimilate a significantly larger number of rain gauges compared to other precipitation products, as well as their high spatial resolutions.
- 2) Non-gauge-corrected reanalysis products (NCEP/NCAR and ERA5) significantly overestimated annual precipitation, especially in mountain areas. ERA5 performed better due to its higher resolution and the advanced data assimilation system.
- 3) The hydrological response in the Yellow River Basin is significantly influenced by the precipitation, and calibration of the rainfall-runoff model can compensate for data deficiencies. We found that the accurate representation of extreme rainfall events in the precipitation dataset plays a crucial role in shaping the calibrated soil parameters. The dependency illustrates that the accuracy of rainfall products in depicting extreme rainfall events is critical for hydrological modelling.

Hence, we recommend the utilization of CMFD or WFDE5/(CRU+GPCC) as near-surface meteorological variables in the hydrological modelling in the Yellow

River Basin. To enhance reanalysis products, it is crucial to integrate weather station observations sharing through the China Meteorological Data Service Center (CMDC) by the China Meteorological Administration (CMA). Furthermore, improving precipitation simulations in mountainous regions can be achieved through a deeper understanding and parameterization of land surface-atmosphere interactions, coupled with higher-resolution simulations. It is important to carefully calibrate the hydrological model to reduce dependence on the quality of precipitation input, which has significant impacts on the components of the water balance. Our evaluation contributes not only to improved hydrological modelling practices but also holds broader implications for Earth system and climate modelling across various spatiotemporal scales.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s00704-023-04758-w>.

Acknowledgements The authors thank Xin Yin (Institute for Geophysics and Meteorology, University of Cologne) for helpful discussion about the results. We would also like to thank the anonymous reviewer and the editor for their valuable comments and suggestions, which helped to improve this paper.

Author contribution CJ, EP and YS developed the concept; CJ wrote the main manuscript text; CJ and QX performed the analysis; and all authors reviewed the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. We thank the German Research Foundation (DFG) for funding through the Heisenberg Programme “Multiscale Simulation of Earth Surface Processes” (project number: 434377576) and the Collaborative Research Center SFB1211 “Earth—Evolution at the Dry Limit” (project number: 268236062). We furthermore thank the Regional Computing Center of the University of Cologne (RRZK) for providing computing time on the DFG-funded (Funding number: INST 216/512/1FUGG) High Performance Computing (HPC) system CHEOPS as well as support.

Data availability NCEP/NCAR is available through the NOAA Physical Sciences Laboratory (PSL) <https://psl.noaa.gov/data/gridded/data.ncep.reanalysis.html>. ERA5, WFDE5/CRU and WFDE5/(CRU+GPCC) are available through the Climate Data Store, <https://cds.climate.copernicus.eu/>. CMFD is available through the China National Tibetan Plateau/Third Pole Environment Data Center, <https://data.tpcd.ac.cn/en/data/8028b944-daaa-4511-8769-965612652c49/>. GLDAS 2.0 is available through the NASA Goddard Earth Sciences (GES) Data and Information Services Center (DISC), <https://disc.gsfc.nasa.gov/>. Daily water discharges of the Yellow River at four hydrological stations, including Tangnaihe, Lanzhou, Toudaoguai and Huayankou, are available on request from the National Geographic Resource Science SubCenter, the National Earth System Science Data Center and the National Science and Technology Infrastructure of China (<http://ge.geodata.cn>). Statistical irrigation data in the Yellow River Basin during the study period can be obtained from Jia et al. (2006) and downscaled to annual data (Jiang et al. 2022) according to the Yellow River Bulletin of Water Resources from 2000 to 2009, published by the Yellow River Conservancy Commission (YRCC) of the Ministry of Water Resources of China (<http://www.yrcc.gov.cn/other/hhgb/>). The source code for the AHMS-IRR is free to use and can be found in the

Github repository (<https://github.com/JiangCong1990/AHMS-IRRIG>). Model input and simulation data of AHMS-IRR in the Yellow River Basin can be found in this Github repository (https://github.com/JiangCong1990/Yellow_River_Basin).

Declarations

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Amjad M, Yilmaz MT, Yucel I, Yilmaz KK (2020) Performance evaluation of satellite- and model-based precipitation products over varying climate and complex topography. *J Hydrol* 584:124707. <https://doi.org/10.1016/j.jhydrol.2020.124707>
- Bacon DP, Ahmad NN, Boybeyi Z, Dunn TJ, Hall MS, Lee PCS, Sarma RA, Turner MD, Waight KT, Young SH, Zack JW (2000) A dynamically adapting weather and dispersion model: the operational multiscale environment model with grid adaptivity (OMEGA). *Mon Weather Rev* 128(7):2044–2076. [https://doi.org/10.1175/1520-0493\(2000\)128<2044:ADAWAD>2.0.CO;2](https://doi.org/10.1175/1520-0493(2000)128<2044:ADAWAD>2.0.CO;2)
- Beck HE, Pan M, Roy T, Weedon GP, Pappenberger F, Van Dijk AIJM, Huffman GJ, Adler RF, Wood EF (2019) Daily evaluation of 26 precipitation datasets using Stage-IV gauge-radar data for the CONUS. *Hydrol Earth Syst Sci* 23(1):207–224. <https://doi.org/10.5194/hess-23-207-2019>
- Beck HE, Vergopolan N, Pan M, Levizzani V, Van Dijk AIJM, Weedon GP, Brocca L, Pappenberger F, Huffman GJ, Wood EF (2017) Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling. *Hydrol Earth Syst Sci* 21(12):6201–6217. <https://doi.org/10.5194/hess-21-6201-2017>
- Beck HE, Vergopolan N, Pan M, Levizzani V, van Dijk AIJM, Weedon GP, Brocca L, Pappenberger F, Huffman GJ, Wood EF (2020) Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling. *Adv Glob Change Res* 69(12):625–653. https://doi.org/10.1007/978-3-030-35798-6_9
- Bitew MM, Gebremichael M, Ghebremichael LT, Bayissa YA (2012) Evaluation of high-resolution satellite rainfall products through streamflow simulation in a hydrological modeling of a small mountainous watershed in Ethiopia. *J Hydrometeorol* 13(1):338–350. <https://doi.org/10.1175/2011JHM1292.1>
- Chen F, Dudhia J (2001) Coupling and advanced land surface-hydrology model with the Penn State-NCAR MM5 modeling system. Part I: Model implementation and sensitivity. *Mon Weather Rev* 129(4):569–585. [https://doi.org/10.1175/1520-0493\(2001\)129<0569:CAALSH>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<0569:CAALSH>2.0.CO;2)
- Cong Z, Yang D, Gao B, Yang H, Hu H (2009) Hydrological trend analysis in the Yellow River basin using a distributed hydrological model. *Water Resour Res* 45(7). <https://doi.org/10.1029/2008WR006852>
- Cucchi M, Weedon GP, Amici A, Bellouin N, Lange S, Müller Schmied H, Hersbach H, Buontempo C (2020) WFDE5: Bias-adjusted ERA5 reanalysis data for impact studies. *Earth Syst Sci Data* 12(3):2097–2120. <https://doi.org/10.5194/essd-12-2097-2020>
- Dietachmayer GS (1992) Application of continuous dynamic grid adaptation techniques to meteorological modeling. Part II: efficiency. *Mon Weather Rev* 120(8):1707–1722. [https://doi.org/10.1175/1520-0493\(1992\)120<1707:AOCDDG>2.0.CO;2](https://doi.org/10.1175/1520-0493(1992)120<1707:AOCDDG>2.0.CO;2)
- Dinku T, Chidzambwa S, Ceccato P, Connor SJ, Ropelewski CF (2008) Validation of high-resolution satellite rainfall products over complex terrain. *Int J Remote Sens* 29(14):4097–4110. <https://doi.org/10.1080/01431160701772526>
- Gao YC, Liu MF (2013) Evaluation of high-resolution satellite precipitation products using rain gauge observations over the Tibetan Plateau. *Hydrol Earth Syst Sci* 17(2):837–849. <https://doi.org/10.5194/hess-17-837-2013>
- Gochis, DJ, Barlage M, Cabell R, Casali M, Dugger A, FitzGerald K, McAllister M, McCreight J, RafieeiNasab A, Read L, Sampson K, Yates D, Zhang Y (2020). The WRF-Hydro® modeling system technical description, (Version 5.1.1). NCAR Technical Note. 107 pages. Available online at: <https://ral.ucar.edu/sites/default/files/docs/water/wrf-hydro-v511-technical-description.pdf>
- Gupta HV, Kling H, Yilmaz KK, Martinez GF (2009) Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. *J Hydrol* 377(1–2). <https://doi.org/10.1016/j.jhydrol.2009.08.003>
- He J, Yang K, Tang W, Lu H, Qin J, Chen Y, Li X (2020) The first high-resolution meteorological forcing dataset for land process studies over China. *Sci Data* 7(1):1–11. <https://doi.org/10.1038/s41597-020-0369-y>
- Hersbach H, Bell B, Berrisford P, Hirahara S, Horányi A, Muñoz-Sabater J, Nicolas J, Peubey C, Radu R, Schepers D, Simmons A (2020) The ERA5 global reanalysis. *Q J R Meteorol Soc* 146(730):1999–2049
- Hirpa FA, Gebremichael M, Hopson T (2010) Evaluation of high-resolution satellite precipitation products over very complex terrain in Ethiopia. *J Appl Meteorol Climatol* 49(5):1044–1051. <https://doi.org/10.1175/2009JAMC2298.1>
- Izadi N, Karakani EG, Saadatabadi AR, Shamsipour A, Fattahi E, Habibi M (2021) Evaluation of era5 precipitation accuracy based on various time scales over Iran during 2000–2018. *Water (Switzerland)* 13(18):2538. <https://doi.org/10.3390/w13182538>
- Jia Y, Wang H, Zhou Z, Qiu Y, Luo X, Wang J, Yan D, Qin D (2006) Development of the WEP-L distributed hydrological model and dynamic assessment of water resources in the Yellow River basin. *J Hydrol* 331(3–4):606–629. <https://doi.org/10.1016/j.jhydrol.2006.06.006>
- Jiang C, Parteli EJR, Shao Y (2020) Application of a coupled atmospheric and hydrological modelling system (AHMS) to the Yellow River Basin, China. In: EGU General Assembly Conference Abstracts, vol 5197. <https://doi.org/10.5194/egusphere-egu2020-5197>
- Jiang C, Parteli EJR, Xia Q, Yin X, Shao Y (2022) A regional hydrological model for arid and semi-arid river basins with consideration of irrigation. *Environ Model Softw* 157:105531. <https://doi.org/10.1016/j.envsoft.2022.105531>
- Jiang Q, Li W, Fan Z, He X, Sun W, Chen S, Wen J, Gao J, Wang J (2021) Evaluation of the ERA5 reanalysis precipitation dataset over Chinese Mainland. *J Hydrol* 595:125660. <https://doi.org/10.1016/j.jhydrol.2020.125660>
- Jiao D, Xu N, Yang F, Xu K (2021) Evaluation of spatial-temporal variation performance of ERA5 precipitation data in China. *Sci Rep* 11(1):1–13. <https://doi.org/10.1038/s41598-021-97432-y>

- Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Woollen J, others (1996) The NCEP/NCAR 40-year reanalysis project. *Bull Am Meteorol Soc* 77(3):437–472
- Kling H, Fuchs M, Paulin M (2012) Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *J Hydrol* 424. <https://doi.org/10.1016/j.jhydrol.2012.01.011>
- Krause P, Boyle DP, Bäse F (2005) Comparison of different efficiency criteria for hydrological model assessment. *Adv Geosci* 5:89–97. <https://doi.org/10.5194/adgeo-5-89-2005>
- Maggioni V, Meyers PC, Robinson MD (2016) A review of merged high-resolution satellite precipitation product accuracy during the Tropical Rainfall Measuring Mission (TRMM) era. *J Hydrometeorol* 17(4):1101–1117. <https://doi.org/10.1175/JHM-D-15-0190.1>
- Martens B, Miralles D, Lievens H, Fernández-Prieto D, Verhoest NEC (2016) Improving terrestrial evaporation estimates over continental Australia through assimilation of SMOS soil moisture. *Int J Appl Earth Obs Geoinf* 48:146–162. <https://doi.org/10.1016/j.jag.2015.09.012>
- Martens B, Miralles DG, Lievens H, Van Der Schalie R, De Jeu RAM, Fernández-Prieto D, Beck HE, Dorigo WA, Verhoest NEC (2017) GLEAM v3: satellite-based land evaporation and root-zone soil moisture. *Geosci Model Dev* 10(5):1903–1925. <https://doi.org/10.5194/gmd-10-1903-2017>
- Milly PCD, Dunne KA (2002) Macroscale water fluxes: 1. Quantifying errors in the estimation of basin mean precipitation. *Water Resour Res* 38(10):23-1–23-14. <https://doi.org/10.1029/2001wr000759>
- Nash E, Sutcliffe V (1970) River flow forecasting through conceptual models part I. A discussion of principles. *J Hydrol* 10(3):282–290
- Niu GY, Yang ZL, Mitchell KE, Chen F, Ek MB, Barlage M, Kumar A, Manning K, Niyogi D, Rosero E, Tewari M, Xia Y (2011) The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *J Geophys Res-Atmos* 116(12). <https://doi.org/10.1029/2010JD015139>
- Pilgrim DH, Chapman TG, Doran DG (1988) Problems of rainfall-runoff modelling in arid and semiarid regions. *Hydrol Sci J* 33(4). <https://doi.org/10.1080/02626668809491261>
- Qi W, Zhang C, Fu G, Sweetapple C, Zhou H (2016) Evaluation of global fine-resolution precipitation products and their uncertainty quantification in ensemble discharge simulations. *Hydrol Earth Syst Sci* 20(2):903–920. <https://doi.org/10.5194/hess-20-903-2016>
- Rodell M, Houser PR, Jambor U, Gottschalck J, Mitchell K, Meng CJ, Arsenault K, Cosgrove B, Radakovich J, Bosilovich M, Entin JK, Walker JP, Lohmann D, Toll D (2004) The Global Land Data Assimilation System. *Bull Am Meteorol Soc* 85(3):381–394. <https://doi.org/10.1175/BAMS-85-3-381>
- Sheffield J, Goteti G, Wood EF (2006) Development of a 50-year high-resolution global dataset of meteorological forcings for land surface modeling. *J Clim* 19(13):3088–3111. <https://doi.org/10.1175/JCLI3790.1>
- Shi H, Li T, Wei J (2017) Evaluation of the gridded CRU TS precipitation dataset with the point raingauge records over the Three-River Headwaters Region. *J Hydrol* 548:322–332. <https://doi.org/10.1016/j.jhydrol.2017.03.017>
- Skamarock WC, Klemp JB (2008) A time-split nonhydrostatic atmospheric model for weather research and forecasting applications. *J Comput Phys* 227(7):3465–3485. <https://doi.org/10.1016/j.jcp.2007.01.037>
- Stisen S, Sandholt I (2010) Evaluation of remote-sensing-based rainfall products through predictive capability in hydrological runoff modelling. *Hydrol Process* 24(7):879–891. <https://doi.org/10.1002/hyp.7529>
- Tang G, Zeng Z, Long D, Guo X, Yong B, Zhang W, Hong Y (2016) Statistical and hydrological comparisons between TRMM and GPM Level-3 products over a midlatitude Basin: Is day-1 IMERG a good successor for TMPA 3B42V7? *J Hydrometeorol* 17(1):121–137. <https://doi.org/10.1175/JHM-D-15-0059.1>
- Vorosmarty, C. J., & Sahagian, D. (2000). Anthropogenic disturbance of the terrestrial water cycle. In *BioScience* (50, 9). [https://doi.org/10.1641/0006-3568\(2000\)050\[0753:ADOTTW\]2.0.CO;2](https://doi.org/10.1641/0006-3568(2000)050[0753:ADOTTW]2.0.CO;2)
- Waichler SR, Wigmosta MS (2003) Development of hourly meteorological values from daily data and significance to hydrological modeling at H. J. Andrews experimental forest. *J Hydrometeorol* 4(2). [https://doi.org/10.1175/1525-7541\(2003\)4<251:DOHMF>2.0.CO;2](https://doi.org/10.1175/1525-7541(2003)4<251:DOHMF>2.0.CO;2)
- Wang J, Zhuo L, Han D, Liu Y, Rico-Ramirez MA (2023) Hydrological model adaptability to rainfall inputs of varied quality. *Water Resour Res* 59(2). <https://doi.org/10.1029/2022WR032484>
- Xia Q (2019) Development and application of a coupled atmospheric and hydrological modelling system, pp 1–125
- Xia Q, Liu P, Fan Y, Cheng L, An R, Xie K, Zhou L (2022) Representing irrigation processes in the land surface-hydrological model and a case study in the Yangtze River Basin, China. *J Adv Model Earth Syst* 14(7). <https://doi.org/10.1029/2021MS002653>
- Xia Y (2008) Adjustment of global precipitation data for orographic effects using observed annual streamflow and the LaD model. *J Geophys Res-Atmos* 113(4). <https://doi.org/10.1029/2007JD008545>
- Yang C, Lin Z, Yu Z, Hao Z, Liu S (2010) Analysis and simulation of human activity impact on streamflow in the Huaihe river basin with a large-scale hydrologic model. *J Hydrometeorol* 11(3):810–821. <https://doi.org/10.1175/2009JHM1145.1>
- Yin Z, Otlé C, Ciaï P, Zhou F, Wang X, Jan P, Dumas P, Peng S, Li L, Zhou X, Bo Y, Xi Y, Piao S (2021) Irrigation, damming, and streamflow fluctuations of the Yellow River. *Hydrol Earth Syst Sci* 25(3):1133–1150. <https://doi.org/10.5194/hess-25-1133-2021>
- Yu B (1998) Theoretical justification of SCS method for runoff estimation. *J Irrig Drain Eng* 124(6). [https://doi.org/10.1061/\(asce\)0733-9437\(1998\)124:6\(306\)](https://doi.org/10.1061/(asce)0733-9437(1998)124:6(306))
- Yu Y, Schneider U, Yang S, Becker A, Ren Z (2020) Evaluating the GPCC full data daily analysis version 2018 through ETCCDI indices and comparison with station observations over mainland of China. *Theor Appl Climatol* 142(3–4):835–845. <https://doi.org/10.1007/s00704-020-03352-8>
- Yu Z, Pollard D, Cheng L (2006) On continental-scale hydrologic simulations with a coupled hydrologic model. *J Hydrol* 331(1–2):110–124. <https://doi.org/10.1016/j.jhydrol.2006.05.021>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.