water

Article

Accuracy Evaluation of Multiple Runoff Products: A Case Study of the Middle Reaches of the Yellow River

Handi Cui [1] and Chang Huang [2,3,*]

[1] Shaanxi Key Laboratory of Earth Surface System and Environmental Carrying Capacity, College of Urban and Environmental Sciences, Northwest University, Xi'an 710127, China; cuihandi@stumail.nwu.edu.cn

[2] Engineering Technology Research Center of Resources Environment and Geo-Graphic Information System of Anhui Province, School of Geography and Tourism, Anhui Normal University, Wuhu 241002, China

[3] Key Laboratory of Earth Surface Processes and Regional Response in the Yangtze-Huaihe River Basin, School of Geography and Tourism, Anhui Normal University, Wuhu 241002, China

* Correspondence: chuang@ahnu.edu.cn

Abstract: Recent advances in hydrological modling have led to the generation of numerous global or regional runoff datasets, which have been widely used in hydrological analysis. However, it is not yet clear how their accuracy and reliabilities are. In this study, using observed gauge streamflow data at four stations (Hequ, Fugu, Wubu, and Longmen) in the middle reaches of the Yellow River as reference, we compare and evaluate the accuracy of three runoff gridded dataset products (GloFAS, GRFR v1.0, and WGHM) at four temporal scales: daily, monthly, annual, and wet/dry seasons. The results indicate the following: (1) As the temporal scale increases, the simulated streamflow accuracy of the three datasets gradually improves. The GloFAS dataset performs the best at daily scale, while the WGHM dataset outperforms the other two at monthly and annual scales. (2) The three datasets all tend to overestimate the total streamflow at the main stations. (3) Comparing the two hydrological scenarios of wet and dry seasons, all three datasets exhibit better performance during the wet season. (4) The capture of peak streamflow is influenced by dataset type, temporal scale, and station characteristics. In general, the three datasets perform better at stations with higher base streamflow, such as Longmen and Wubu stations. Additionally, this study discusses the possible reasons for their different performances, which can be mainly attributed to three aspects: the quality of meteorological input datasets, missing or simplified simulation processes, and incorrect model structure and parameterization. Future research will consider revising the datasets to obtain more accurate data sources and further enhance the accuracy of watershed streamflow simulations.

Keywords: streamflow simulation; Yellow River Basin; runoff datasets; performance evaluation

Received: 10 December 2024
Revised: 20 January 2025
Accepted: 4 February 2025
Published: 6 February 2025

Citation: Cui, H.; Huang, C. Accuracy Evaluation of Multiple Runoff Products: A Case Study of the Middle Reaches of the Yellow River. *Water* **2025**, *17*, 461. https://doi.org/10.3390/w17030461

1. Introduction

In the face of the serious challenges posed by global climate change and the impacts of human activities, the effective management and protection of water resources has become a core issue of concern for the scientific community, policymakers, and all sectors of society [1,2]. River runoff dynamics, a key element in the natural water cycle, directly support sustainable water resource development and utilization. Moreover, they play a pivotal role in material and carbon cycles within watersheds and are essential for monitoring and early warning systems for natural disasters such as floods and droughts [3–5]. However, the inherent complexity and spatial variability of runoff have long posed significant challenges for monitoring and data acquisition.

Traditional runoff observation is based on field measurements at hydrological stations, which can provide accurate runoff data, but its high construction and operation costs and limited spatial coverage cannot meet the needs of integrated management of large-scale river basins [6,7]. Hydrological modeling has emerged as an essential tool for studying and predicting the impacts of climate change on the global water cycle. With advancements in computer technology, hydrological sciences, and the rapid growth of data, particularly remote sensing data, global-scale hydrological modeling has flourished [8,9], producing numerous global runoff datasets with varying spatial and temporal resolutions. These runoff data products offer high temporal continuity and wide spatial coverage, which provide solid data support for water resource management decisions, hydrological forecasting, and disaster early warning response. However, in the hydrological modeling process, different models are built on a complex combination of different underlying processes and assumptions. When coupled with various climate forcings, this can result in significant differences in runoff outputs across models [10]. It has been claimed that the simulated runoff produced by these global models exhibits differences in magnitude, variability, and direction of change [11–13]; such variability and uncertainty pose a significant challenge to hydrologic analysis, scientific decision-making, and early warning systems. Therefore, conducting comparative accuracy analyses of multiple runoff datasets and exploring the root causes of the errors and the influencing factors to select the datasets that are most suitable for a specific study area are of significant theoretical and practical value.

Recent studies have made advancements in evaluating runoff dataset performance [14]. For example, Hou et al. [10] comprehensively evaluated the simulated streamflow of 21 global models in terms of annual mean magnitude, interannual variability, annual trends, and intra-annual cycle. The result reveals significant uncertainty and caution in interpretation among the simulated streamflows of these global models. Beck et al. [15] assessed daily simulated runoff from six global hydrological models (GHMs) and four land surface models (LSMs), emphasizing the need for improved forcing data and parameterization schemes to achieve more accurate streamflow simulations. Similarly, Sikder et al. [16] compared the performance of multiple land surface models (LSMs) for simulating streamflow in a transboundary river basin and identified the best-performing model. To address the lack of runoff data in sparsely populated regions such as the Tibetan Plateau (TP), Bai et al. [17] evaluated the performance of four land surface models (CLM, Noah, VIC, and Mosaic) under the GLDAS project, focusing on monthly runoff simulation, seasonal cycles, annual trends, and component partitioning. These studies have significantly enhanced the understanding of runoff datasets' reliability and provided theoretical and methodological foundations for model improvement and optimization.

However, most of these studies focus on global or foreign regional scales, leaving a gap in the comparative analysis of runoff datasets at the regional scale within China. Our study focuses on the middle reaches of the Yellow River Basin, using observed streamflow from four key hydrological stations—Hequ, Fugu, Wubao, and Longmen (2006–2015)—as benchmarks. It systematically compares the performance of three mainstream gridded-runoff datasets (the GRFR v1.0 dataset [18], the GloFAS dataset [19], and the WGHM dataset [20]) across daily, monthly, annual, and seasonal periods. By analyzing the error characteristics of each dataset and their deviations from the observed data, the study aims to explore the possible causes of these accuracy differences so as to provide a solid scientific basis and practical reference guide for the efficient management of water resources and scientific research in the middle reaches of the Yellow River and other similar rivers worldwide.

2. Study Area

The Yellow River is the second-longest river in China. It originates from the Bayan Har Mountains on the Qinghai–Tibet Plateau, flowing through nine provinces and autonomous regions, including Qinghai, Sichuan, Gansu, Ningxia, Inner Mongolia, Shaanxi, Shanxi, Henan, and Shandong, ultimately emptying into the Bohai Sea in Kenli County, Shandong Province, with a total length of 5464 km. Its vast basin covers an area of 795,000 square kilometers, of which approximately 42,000 square kilometers is endorheic (Figure 1). The Yellow River basin boasts rich natural geographical features, spanning the Qinghai–Tibet Plateau, the Loess Plateau, the Inner Mongolia Plateau, and the North China Plain, and encompassing arid, semi-arid, and semi-humid regions, thereby nurturing unique landscapes and ecological environments.

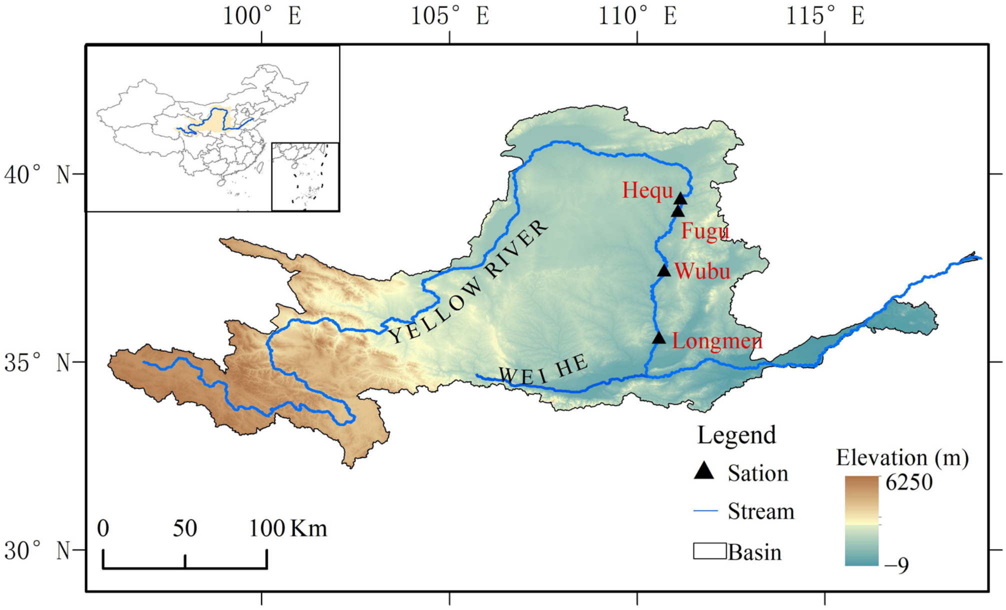

Figure 1. Study area.

The Yellow River basin can be divided into upstream, midstream, and downstream sections based on geographical characteristics. The upstream section, from the source to Hekou in Inner Mongolia, features steep terrain and a significant river gradient, abundant in hydropower resources and designated as a crucial area for ecological conservation. The midstream section extends from Hekou to Taohuayu in Henan, where it receives a vast influx of tributaries from the Loess Plateau, resulting in a sharp increase in sediment content and severe soil erosion. The downstream section, from Taohuayu to the estuary, is renowned for its broad and winding river channels and the unique "above-ground river" landform. The climatic conditions across the basin are highly diverse, with annual precipitation fluctuating between 200 and 650 mm, while annual water surface evaporation can reach as high as 1100 mm. This pronounced imbalance between precipitation and evaporation, coupled with the basin's erodible soil characteristics, exacerbates soil erosion problems in the midstream region. This issue not only degrades the local ecological environment and reduces land productivity but also has far-reaching impacts on downstream regions, intensifying flood risks and compromising the sustainable utilization of water resources.

Given the severity of soil erosion in the midstream region of the Yellow River basin, our study selects four representative hydrological stations—Hequ, Fugu, Wubu, and Longmen—in the midstream section as the research subjects. It is noted that there are no water conservancy facilities such as dams or reservoirs existing within this river section. Monitoring the streamflow here yields essential data for soil erosion assessments, thereby supporting ecological conservation efforts in the Yellow River basin and contributing to the sustainable management of water resources.

3. Materials and Methods

3.1. Introduction to the Runoff Datasets

In this study, three runoff datasets (Table 1), including GRFR V1.0, GloFAS, and WGHM datasets, are selected for evaluation.

Table 1. Introduction to the runoff dataset products.

Runoff Dataset	Production Method	Forcing	Temporal Resolution	Spatial Resolution	Coverage	River Routing
GRFR V1.0	Built using a global high-resolution and high-accuracy natural river runoff simulation system, with distributed hydrological model VIC and river routing model RAPID as the core, integrating multiple sources of data and modeling methods.	MSWEP 2.2 ERA5	Daily	0.05°	1980–2019	RAPID
GloFAS	This dataset is the result of extensive hydrological modeling that combines a grid-based hydrological routing model with a terrestrial surface model. During this process, the land surface model, known as H-TESSEL, calculates the water balance to generate surface and subsurface runoff, and the routing model, RAPID, is utilized to determine the flow within river channels.	ERA5	Daily	0.1°	1979–2018	LISLOOD
WGHM	A conceptual model based on the water balance method, which calculates the net runoff (precipitation minus evapotranspiration and changes in soil water storage) for each grid cell and routes the runoff through the river network to simulate river streamflow.	GSWP3 W5E5 v2.0	Monthly	0.5°	1901–2019	Fractional Routing scheme

Both GRFR V1.0 and GloFAS are global-scale grid-based runoff datasets. Global Reach-level 3-hourly River Flood Reanalysis (GRFR V1.0) [18] is a comprehensive dataset that provides detailed river discharge records for approximately 2.94 million river reaches over 40 years from 1980 to 2019. Its underlying modeling chain integrates the well-calibrated and bias-corrected land surface model VIC, operating at a 0.05° spatial resolution with a

3 h temporal resolution, and the RAPID routing model, which incorporates 2.94 million river and catchment vectors. The VIC model [21], developed collaboratively by institutions including the University of Washington and Princeton University, draws upon the water storage capacity curve of the Xin'anjiang model for runoff computations. It is famous for its robust simulation capabilities, accurately reproducing complex terrestrial surface processes such as water and energy balances, snowmelt, and frozen soil dynamics. The RAPID program [22], a river routing model specifically designed to simulate river flow transport and routing, employs a matrix-based Muskingum method to calculate water flow and volume at each river segment within a river network. Its efficient parallel computation capabilities make it particularly adept at handling large-scale river network simulations, making it well suited for early flood warning and real-time streamflow prediction. In the GRFR V1.0 project, no lakes or reservoirs or human regulations are considered and their influences on flow are ignored. The VIC model and RAPID program operate synergistically, using precipitation data from the MSWEP 2.2 global dataset (1979–present) with a 3 h temporal resolution and a 0.1° spatial resolution and other meteorological fields, including surface air temperature, pressure, incoming shortwave and longwave radiation, humidity, and wind speed derived from the downscaled ERA5 [23], the latest climate reanalysis dataset produced by ECMWF as input. Thereby, daily streamflow and the characteristics of flood events (spatial distribution and seasonality) are extracted and studied. This dataset provides an indispensable high-quality data resource for global streamflow monitoring and flood risk assessment. The data record is publicly accessible at https://www.reachhydro.org/home/records/grfr (accessed on 20 July 2024).

The Global Flood Awareness System (GloFAS) [19], an operational system under the European Commission's Copernicus Emergency Management Service, is designed to predict and monitor global floods. It offers forecasts up to 30 days in advance and seasonal outlooks up to 4 months ahead. The system generates daily streamflow forecasts using a coupled H-TESSEL land surface scheme and the LISFLOOD model forced by ECMWF IFS meteorological forecasts. The land surface scheme, referred to as the Hydrology Tiled ECMWF Scheme for Surface Exchanges over Land (H-TESSEL) and used operationally in the Integrated Forecast System (IFS), computes the surface and subsurface runoff. A simplified version of LISFLOOD [24] is used for routing the runoff produced by the land surface scheme through the river network and computing the groundwater fluxes. It is capable of representing features that impact the timing and magnitude of river discharge, such as lakes, reservoirs, and human water use. Thereby, a total of 463 large lakes (>100 km^2) and 667 reservoirs are incorporated into it. During the modeling process, the surface runoff from HTESSEL is used as input into the LISFLOOD river channel routing module to generate a seamless 40-year global coverage daily streamflow. The data record is publicly accessible at https://cds.climate.copernicus.eu (accessed on 20 July 2024).

WaterGAP 2.2d [20] serves as a pivotal global hydrological model, adept at quantifying the human impact on groundwater and surface water resources, along with the intricate dynamics of water flow and storage, thereby providing a comprehensive assessment of terrestrial water resources worldwide. The model is esteemed for its reliability and superior performance in tackling critical water issues and adapting to diverse climatic regions globally. WaterGAP encompasses three integral components: the Global Water Use Model, the Groundwater–Surface Water Use Linkage Model (GWSWUSE), and the WaterGAP Global Hydrological Model (WGHM). The WGHM leverages the gswp3-w5e5 climate forcings (a concatenation of two datasets–one for the period prior to 1979 and one for the period starting in 1979) to model the global water cycle, which involves precipitation, evaporation, streamflow, soil moisture, snowmelt, groundwater, and surface water processes. GSWP3 [25] version 1.09 is a bias-adjusted and downscaled version of

Twentieth Century Reanalysis version 2. W5E5 v2.0 [26] is a bias-adjusted version of the current version of the European Reanalysis ERA5 [23]. It operates on a daily simulation time step, simulating the effects of both human water use and reservoirs and computing hydrological processes for each grid cell by a so-called fractional routing scheme to yield monthly simulated streamflow. The data record is publicly accessible at https://doi.pangaea.de/10.1594/PANGAEA.948461 (accessed on 20 July 2024).

3.2. Methodology

This study aims to evaluate and compare the accuracy and reliability of three grid-based datasets (GRFR V1.0, GloFAS, and WGHM) by the following steps:

(1) Data collection: The observed runoff data at the four gauging stations (Hequ, Fugu, Wubu, and Longmen, see Table 2) were all sourced from the official website of the Yellow River Conservancy Commission (http://www.yrcc.gov.cn/, accessed on 20 July 2024). These data reflect the flow conditions after human interventions. To ensure the consistency and reliability of the comparative analysis across different time scales, we uniformly selected the time series data from 2006 to 2015 for analysis. In addition, three grid runoff products were collected by the corresponding website, with streamflow units for each grid expressed in m^3/s, consistent with the observed streamflow units at hydrological gauging stations.

(2) Simulated streamflow extraction: Based on the longitude and latitude values of the hydrological stations, each station was matched to a unique pixel within the given grid-based global runoff dataset, and the streamflow value of that pixel was extracted. Finally, the extracted streamflow values from the given grid-based runoff dataset were subsequently compared to the observed streamflow at the corresponding hydrological stations.

Table 2. Characteristics of four gauging stations.

Characteristics	Hequ	Fugu	Wubu	Longmen
Latitude	39.37° N	39.04° N	37.45° N	35.67° N
Longitude	111.15° E	111.08° E	110.72° E	110.58° E
Period of daily streamflow	2006–2015	2006–2015	2006–2015	2006–2015
Catchment Area (km^2)	397,658	404,039	433,514	497,552

3.3. Assessment Criteria

The simulated streamflow from each dataset is compared with the observed streamflow over the same period at various scales. Here, the percent bias (PBIAS), correlation coefficient (CC) and Kling–Gupta Efficiency (KGE) [27] are used as the assessment criteria for the performance of the datasets. The evaluation metrics and their corresponding formulas are shown in Table 3.

Table 3. Evaluation metrics and formulas.

Evaluation Metric	Formula	Range of Values
Percent bias	$PBIAS = \frac{\sum(Q_{sim} - Q_{obs})}{\sum Q_{obs}}$	$-\infty \sim 0$
Correlation Coefficient	$CC = \frac{\sum(Q_{obs} - \overline{Q_{obs}})(Q_{sim} - \overline{Q_{sim}})}{\sqrt{\sum(Q_{obs} - \overline{Q_{obs}})^2}\sqrt{\sum(Q_{sim} - \overline{Q_{sim}})^2}}$	$-1 \sim 1$
Kling–Gupta Efficiency	$KGE = 1 - \sqrt{(1 - cc)^2 + (1 - \alpha)^2 + (1 - \beta)^2}$ $\alpha = \alpha_s/\alpha_o, \beta = \mu_s/\mu_o$	$-\infty \sim 1$

In Table 3, Q_{sim} is the simulated streamflow value, Q_{obs} is the observed streamflow value, and $\overline{Q_{obs}}$ and $\overline{Q_{sim}}$ refer to the average observed and simulated streamflow value, respectively. The correlation coefficient (CC) quantifies the fit between simulated and observed values, with values closer to 1 indicating better agreement. Percent bias (PBIAS) assesses the degree of overestimation or underestimation in the simulated streamflow, where values closer to 0 reflect minimal deviation between the observed streamflow value and the simulated streamflow value. The Kling–Gupta Efficiency (KGE), an improved version of the Nash–Sutcliffe Efficiency (NSE) [28], decomposes the NSE into three independent components (linear correlation, bias ratio, and variability). In this research, KGE is utilized as a comprehensive assessment criterion to measure the agreement between the observed and simulated streamflow. It combines linear correlation (CC), variability (α), and bias ratio (β) into a multi-objective: u_s and α_s are the mean and standard deviation of simulated streamflow value, while u_o and α_o are the mean and standard deviation of observed streamflow value [27]. A KGE value of 1 (CC = 1, α = 1, β = 1) signifies perfect consistency between observed and simulated values.

4. Results

4.1. Simulated Streamflow Performance on Daily Scale

Figure 2 presents scatterplots of daily observed streamflow compared to the GRFR V1.0 and GloFAS datasets, and Table 4 summarizes the performance statistics for the two datasets.

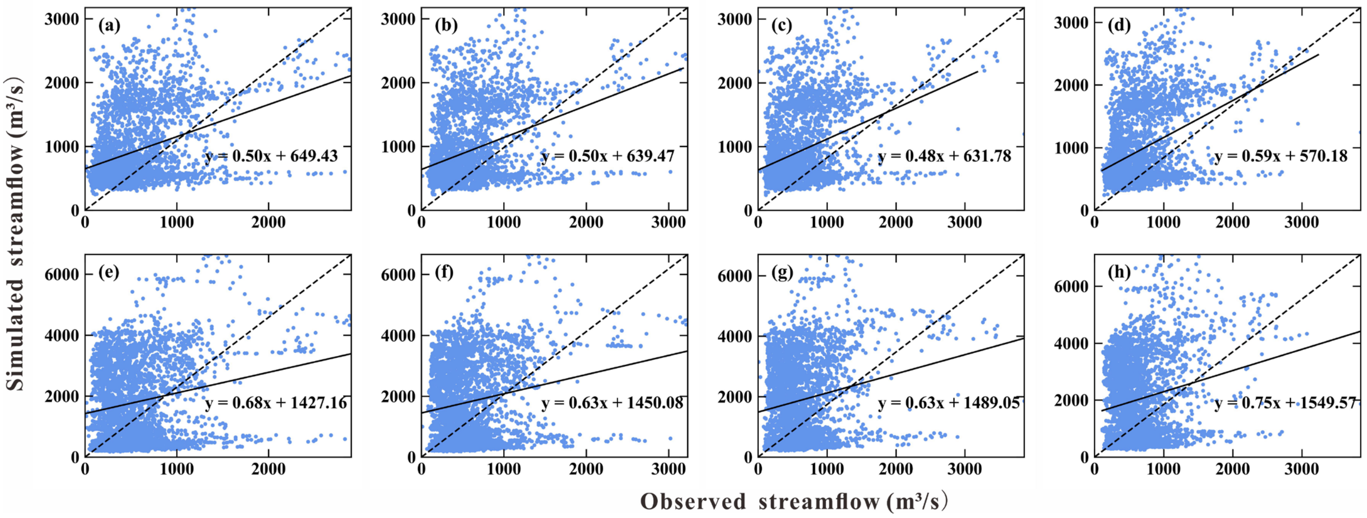

Figure 2. Daily streamflow simulation of two datasets at four selected stations from 2006 to 2015: (**a**) GloFAS dataset at Hequ, (**b**) GloFAS dataset at Fugu, (**c**) GloFAS dataset at Wubu, (**d**) GloFAS dataset at Longmen, (**e**) GRFR V1.0 dataset at Hequ, (**f**) GRFR V1.0 dataset at Fugu, (**g**) GRFR V1.0 dataset at Wubu, (**h**) GRFR V1.0 dataset at Longmen.

Table 4. Daily streamflow evaluation results.

Station	GRFR v1.0			GloFAS		
	KGE	CC	PBIAS	KGE	CC	PBIAS
Hequ	−2.31	0.20	2.23	−0.02	0.35	0.67
Fugu	−2.18	0.20	2.12	0.04	0.36	0.60
Wubu	−2.04	0.20	2.02	0.13	0.37	0.50
Longmen	−2.40	0.21	2.16	0.13	0.42	0.48

As can be seen from Table 4, the two datasets exhibit significant differences in daily streamflow performance. Based on the correlation coefficient (CC), the GloFAS dataset performs the best ($0.35 \leq \text{CC} \leq 0.42$, mean CC of 0.38) compared to the GRFR V1.0 dataset ($0.20 \leq \text{CC} \leq 0.21$, mean cc of 0.20). Regarding the percent bias (PBIAS), the GloFAS dataset performs relatively well at Wubu and Longmen stations, whereas the GRFR V1.0 dataset exhibits poor results across all four hydrological stations. Furthermore, both the GloFAS and GRFR V1.0 datasets consistently overestimate the total streamflow of the study region, as indicated by positive PBIAS values at all stations, with the GRFR V1.0 dataset exhibiting particularly severe overestimation.

In terms of the KGE statistic, neither dataset adequately reproduces the daily runoff time series. The highest KGE value across both datasets is 0.133, with half of the KGE values being negative. Overall, the GloFAS dataset outperforms the GRFR V1.0, with average KGE values of 0.07 and −2.23, respectively. The GloFAS dataset performs better at all four stations, achieving the highest KGE at Longmen and Wubu stations. In comparison, the GRFR V1.0 dataset performs poorly at all four stations, consistently yielding negative KGE values.

4.2. Simulated Runoff Performance on Monthly Scale

Monthly streamflow values are derived by aggregating daily streamflow values. Figure 3 presents the statistical performance metrics of the three products at selected stations, and Figure 4 depicts the monthly trends of the observed streamflow compared with the simulated streamflow of the three datasets.

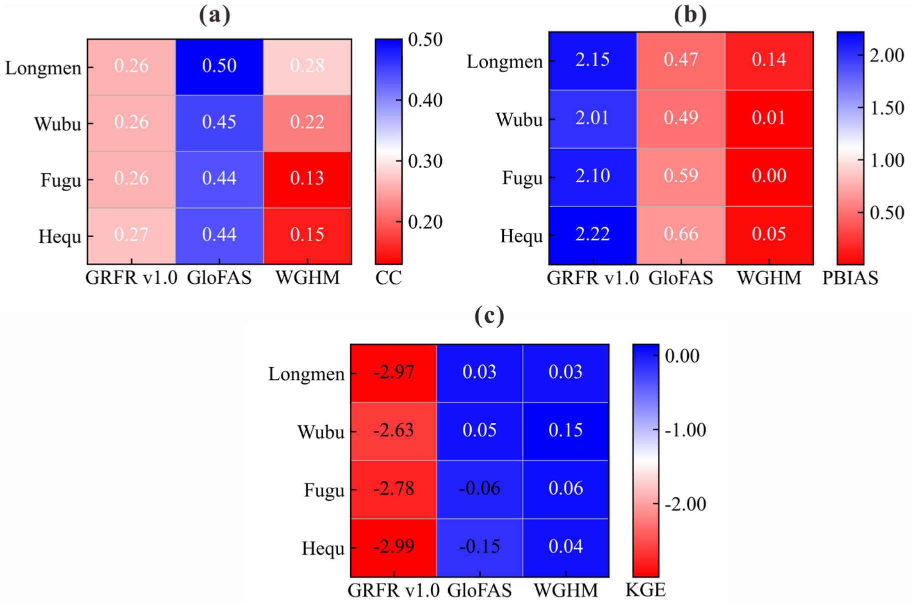

Figure 3. (**a**) Correlation coefficient (CC), (**b**) percent bias (PBIAS), and (**c**) Kling–Gupta Efficiency (KGE) of three products at four selected stations on monthly scale.

From Figure 3, it can be seen that the three datasets exhibit varying performance in monthly streamflow simulation compared to daily scale results. Based on the correlation coefficient (CC), the GloFAS dataset performs the best ($0.44 \leq \text{CC} \leq 0.50$, mean CC of =0.46), followed by GRFR V1.0 ($0.26 \leq \text{CC} \leq 0.27$, mean CC of =0.26) and WGHM ($0.13 \leq \text{CC} \leq 0.28$, mean CC of =0.20), showing improved correlation compared to the daily scale. Regarding the percent bias (PBIAS), this phenomenon for the other two datasets is consistent with the daily scale. The WGHM dataset performs the best, exhibiting the smallest degree of bias and positive PBIAS values at all four stations, indicating a systematical overestimate of total streamflow. PBIAS performance among the three datasets is relatively consistent across stations, with minimal variation. On average, the performance of the Wubu and Longmen stations is slightly better than that of the Hequ and Fugu stations.

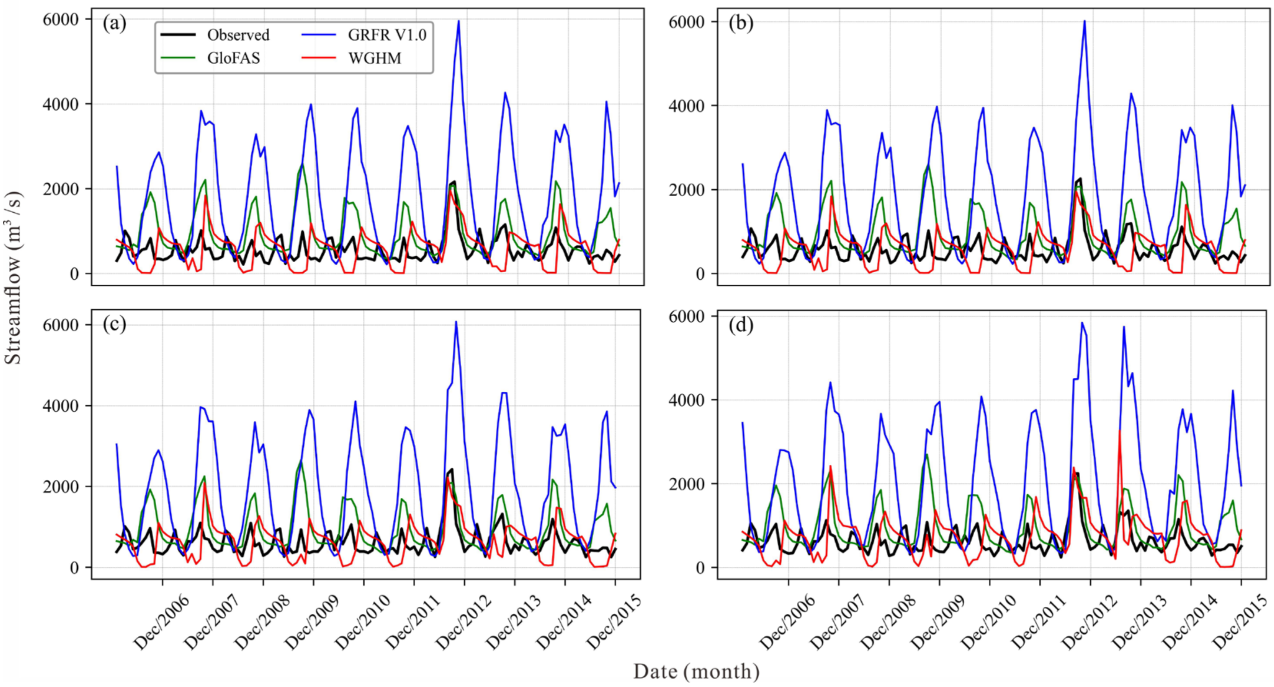

Figure 4. Monthly streamflow trends at four selected stations from 2006 to 2015: (**a**) Hequ, (**b**) Fugu, (**c**) Wubu, and (**d**) Longmen.

In terms of the KGE statistic, the performance gradually improves. Among the three datasets, the highest KGE value is 0.15, while the lowest is −2.99. Overall, the WGHM dataset demonstrates the best performance, while the GRFR V1.0 dataset demonstrates the worst performance. The average KGE values of the three datasets (WGHM, GloFAS, and GRFR V1.0) are 0.07, −0.03, and −2.84, respectively. The GloFAS and GRFR V1.0 datasets perform better at the Wubu and Longmen stations than the Hequ and Fugu stations, whereas the GRFR V1.0 dataset consistently performs poorly across all stations.

In addition, the ability of the three datasets to capture the time of peak flow in the monthly streamflow simulation is also assessed. As can be seen from Figure 4, most datasets perform better at Wubu and Longmen stations than the Hequ and Fugu stations in simulating peak flow timings. The GloFAS and GRFR V1.0 datasets could accurately capture most of the peak flow timings at all four stations but show overestimation, while the WGHM dataset usually shows a delay of approximately two months in capturing peak flows.

4.3. Simulated Streamflow Performance on Annual Scale

Figure 5 presents the statistical performance metrics of the three datasets at selected stations, and Figure 6 depicts the annual trends of the observed streamflow compared with the simulated streamflow of the three datasets. As shown in Figure 5, the three datasets show improved performance in annual streamflow simulation relative to daily and monthly scales.

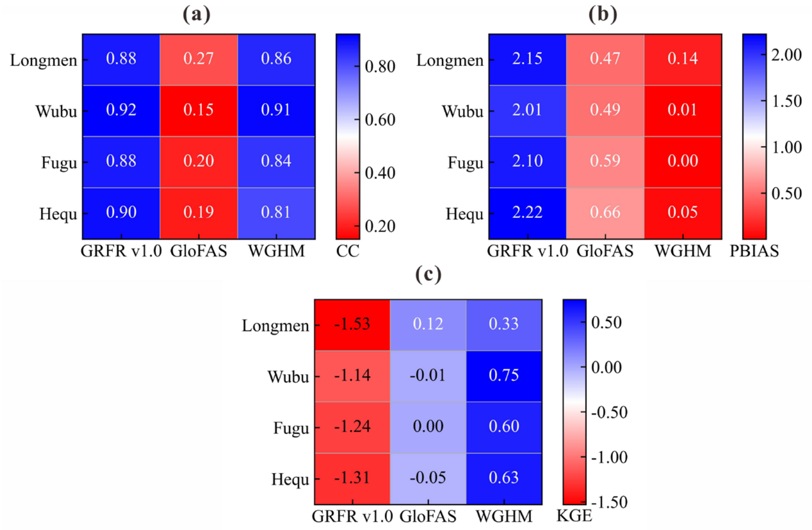

Figure 5. (**a**) Correlation coefficient (CC), (**b**) percent bias (PBIAS), and (**c**) Kling–Gupta Efficiency (KGE) of three products at four selected stations on yearly scale.

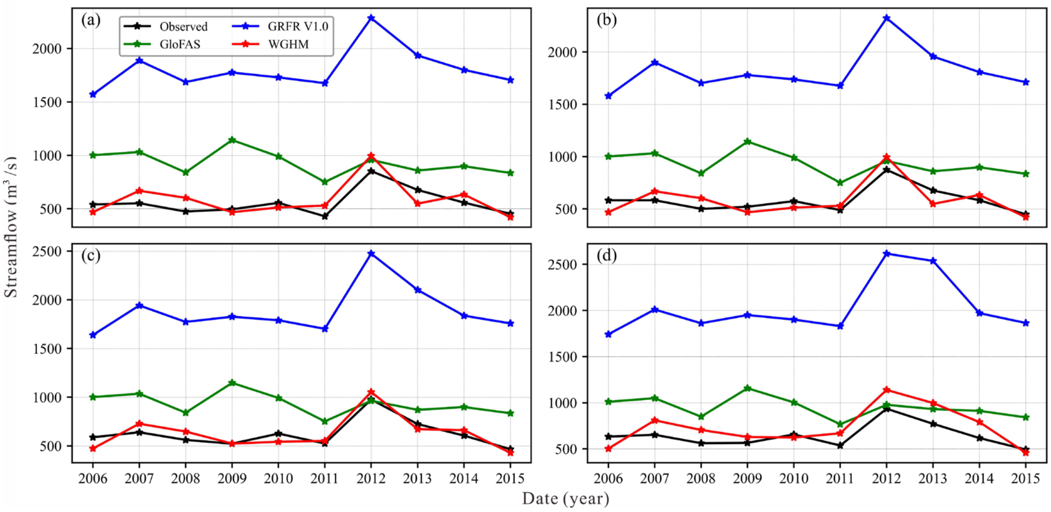

Figure 6. Annual streamflow at four selected stations from 2006 to 2015: (**a**) Hequ, (**b**) Fugu, (**c**) Wubu, and (**d**) Longmen.

Based on the correlation coefficient (CC), the GRFR V1.0 dataset exhibits the best performance ($0.88 \leq \text{CC} \leq 0.92$, mean CC of 0.90), followed by the WGHM dataset ($0.81 \leq \text{CC} \leq 0.91$, mean CC of 0.86), and the GloFAS dataset ($0.15 \leq \text{CC} \leq 0.27$, mean CC of 0.20). The three datasets perform better at the annual scale than at the daily and monthly scales. As the percent bias of the annual scale streamflow is consistent with that of the monthly scale, it is not discussed further here.

In terms of the KGE statistic, approximately half of the KGE statistics are positive, with the highest KGE value among the three datasets being 0.75 and the lowest being -1.53. Overall, the WGHM dataset outperforms the others, while the GRFR V1.0 dataset performs the worst. The average KGE values for the WGHM, GloFAS, and GRFR V1.0 datasets are 0.58, 0.02, and -1.31, respectively. The performance of all datasets exhibits similar differences across stations as the monthly scale, except that the KGE values improve at each station, especially for the GRFR V1.0 dataset, despite remaining negative at each station.

As can be seen in Figure 6, although the base streamflow is different at each station, a general trend of a 'steep decrease' in 2011 followed by a 'steep increase' in 2012 is observed. Both the GRFR V1.0 and WGHM datasets could reasonably reproduce the interannual streamflow trend to capture the two inflection points. However, the GloFAS dataset shows a peak inflection point in 2009 across all stations, which deviates from the observed interannual streamflow trends. Among the three datasets, the WGHM dataset performs the best.

4.4. Simulated Streamflow Performance in Dry/Wet Periods

Through consulting relevant data and observing the change characteristics of observed streamflow at hydrological stations, the middle reaches of the Yellow River basin are divided into two periods from 2006 to 2015: October–May as the dry period and June–September as the wet period. To compare the accuracy of the three datasets during the two periods, the statistical performance metrics of the three datasets is presented, as shown in Figure 7.

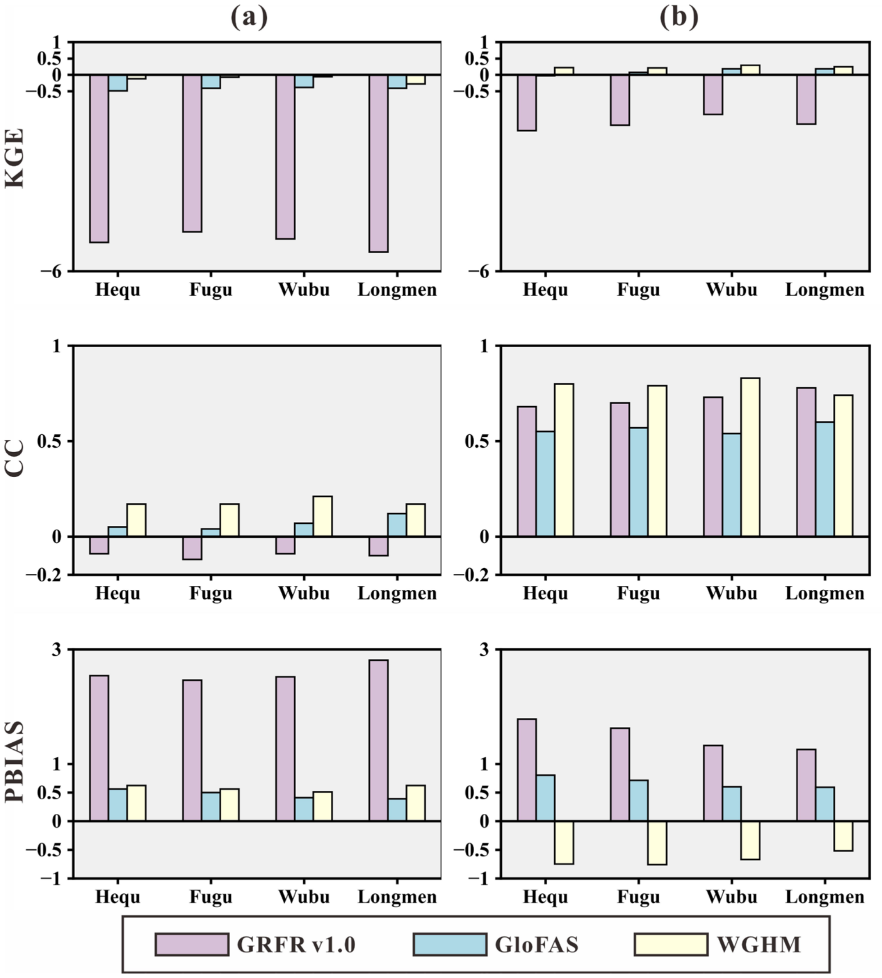

Figure 7. Correlation coefficient (CC), percent bias (PBIAS), and Kling–Gupta Efficiency (KGE) of three products at four selected stations during two periods from 2006 to 2015: (**a**) dry period; (**b**) wet period.

During the dry period, the three datasets generally perform poorly, with predominantly negative KGE values and low CC values, indicating limited accuracy in streamflow simulation. All three datasets tend to overestimate the total streamflow. During the wet period, the performance of the datasets improves, with most KGE values exceeding 0 and CC values typically above 0.5, indicating enhanced streamflow simulation capability. However, the GloFAS dataset shows a higher degree of overestimation at all stations during the wet period compared to the dry period, while the WGHM dataset tends to underestimate streamflow. In conclusion, for both hydrological scenarios, most datasets perform better during the wet period and could simulate streamflow changes more effectively.

Both the GloFAS and GRFR v1.0 datasets exhibit overestimation during the wet period, which is consistent with monthly and annual scales. Differently, the WGHM dataset shifts from overestimation during the dry period to underestimation in the wet period, with the degree of overestimation being more pronounced. Compared to the other two datasets, the WGHM dataset performs better regardless of the period but exhibits variability among stations, with Wubu and Longmen stations consistently achieving the best results.

5. Discussion

Although existing datasets are valuable for modeling streamflow in the middle reaches of Yellow River Basin and can aid in streamflow prediction, they exhibit some issues, including low accuracy, significant percent bias, and poor performance at specific time scales and stations. These issues are likely attributed to factors such as the quality of meteorological input data, missing or simplified model processes, and incorrect model parameterization [29–31]. As a result, these model outputs are often limited to specific rivers and regions, where they can only be qualitatively assessed or modeled.

Accurate streamflow estimation is highly dependent on the quality of the meteorological input data, particularly the precipitation forcing [31], as it is the primary and direct water source for the surface water cycle. The GRFR V1.0 dataset used the global precipitation dataset MSWEP version 2.2, with a spatial resolution of 0.1° [32], as its primary precipitation forcing during the modeling. However, previous studies [33,34] have shown that the precipitation data perform poorly in the Yellow River basin. The GloFAS dataset used ERA5 from the ECMWF (European Centre for Medium-Range Weather Forecasts) as input precipitation forcing, which represents a significant improvement over the previous ERA-Interim. But a recent study [35] found that precipitation data in the ERA5 are significantly overestimated in the Yellow River Basin, particularly in the highland climate zones. This may contribute to the overestimation of simulated streamflow in the GloFAS dataset. During the period from 2006 to 2015, WGHM utilized the precipitation forcing data from the bias-corrected dataset W5E5 v2.0. It is important to note that in the Chinese region, this dataset is constrained by a limited number of observational stations and comparatively lower data quality. In a word, the simulated streamflow of a single hydrological model forced by different meteorological forcing datasets will inherently exhibit biases, let alone the simulations produced by various hydrological models forced by different meteorological datasets.

In addition to meteorological input data, simplifications in model structure or the omission of key processes may also lead to biases in simulated streamflow. As is known to all, many human activities in the Yellow River basin seriously affect the flow regimes, such as ecological construction, and human water use (particularly the irrigation water use). On the one hand, ecological construction such as soil and water conservation and afforestation projects have made the underlying surface conditions of the basin more complex, thus affecting the processes of interception, infiltration, evapotranspiration, and runoff within the basin and further increasing the difficulty of hydrological simulation. On the other hand, both WGHM and GloFAS took into account human water use, whereas GRFR V1.0 overlooked its impact [18], which may account for its severe overestimation and inferior performance in the middle reaches of the Yellow River. Different models show varying performance due to differences in their underlying physical mechanisms. The GRFR v1.0 dataset used the VIC model for hydrological simulations to generate runoff. However, a previous study [36] showed that the VIC model tends to underestimate evapotranspiration under low upper soil moisture, which further impacted runoff simulation. The GloFAS product combined gridded surface and subsurface runoff from the CHTESSEL model with the LISFLOOD model, allowing lateral connections between grid cells and runoff through stream channels to simulate streamflow. However, the HTESSEL surface model usually calculates Potential Evapotranspiration (PET) by calling upon the surface energy balance a second time, which may overestimate evapotranspiration under drought conditions, thus affecting runoff generation. Additionally, simplifications in model structure may also explain the significant differences in runoff modeling under dry versus wet hydrological scenarios. Runoff generation is based on the saturation-excess mechanism [37,38], which assumes that runoff can be generated only when the entire soil column is saturated. During arid periods, runoff typically follows the infiltration-excess mechanism, where runoff occurs when rainfall intensity exceeds the soil's infiltration capacity [39,40]. During semi-arid and semi-humid periods, rainfall runoff generation mechanisms are more complex in semi-arid and semi-humid periods than in humid periods, as infiltration-excess and saturation-excess runoff interact to varying degrees in these areas. Consequently, all models, whether simple or complex, yield more accurate and reliable results during wet periods. This explains why the three datasets perform better during the wet period.

The modeling process of hydrological models in these grid-based runoff datasets involves numerous parameters, including land surface and model parameters. Land surface parameters, such as vegetation, soil, elevation, and slope, can be obtained from high-resolution satellite products with acceptable accuracy. However, hydrological model parameters are less deterministic due to their significant temporal and spatial variability, which limits the effectiveness of the calibration algorithm at the watershed scale. The GloFAS dataset calibrated only the LISFLOOD model parameters and did not adjust the land surface model (H-TESSEL) parameters, which may trigger runoff bias [41]. Furthermore, the WGHM has been calibrated to align with the observed long-term average annual streamflow at gauging stations. This is why WGHM provides such good results at the annual scale. For a single catchment, discrepancies between simulation and observation can be reduced through model parameter calibration. However, global-scale calibration of model parameters is challenging due to the limited availability of surface observation data, especially in data-poor basins where significant uncertainties may arise. Additionally, all three products input hydrological model simulated runoff into the river routing model for streamflow simulation. The river routing model used in the GRFR v1.0 product was based on the fixed-velocity Muskingum method [18]. However, this method lacks explicit parameterization of floodplains, making it ineffective in capturing changes in flow dynamics caused by human regulations. In contrast, the WGHM and GloFAS products [18,21] have been calibrated for river routing parameters, with WGHM specifically calibrated for the Yellow River Basin. As a result, the errors in river routing simulation for the two products are relatively small.

In conclusion, the complex climate condition and the impact of human activities in the Yellow River Basin pose stringent demands on hydrological modeling. To achieve high-precision simulated streamflow, efforts should be made to enhance the model's performance in non-humid areas, consider the influence of human factors in highly managed regions, and improve the parameters and structure of vegetation–soil–atmosphere interaction models.

Finally, it is noted that another dataset, called Global Runoff Data Centre (GRDC) [42], is a unique collection of river discharge data at a global scale. It contains time series of daily and monthly river discharge data of currently more than 9,800 stations all over the world. In the GRDC dataset, nine stations are located within the Yellow River Basin, of which only three are in the middle reaches—Sanmenxia, Heishiguan, and Huayuankou. However, these three stations lack daily data, and the length of their monthly data does not overlap with our study period. This is why we used the gauge data as the reference, instead of GRDC, in this study. However, considering the similarity between GRDC and gauge observations, this study could also be considered as a comparison study between GRDC and the selected three datasets.

6. Conclusions

We systematically evaluated the accuracy of runoff datasets across daily, monthly, seasonal, and annual time scales and analyzed the underlying causes of performance variations using observed flow data from four key hydrological stations in the middle reaches of the Yellow River Basin—Hequ, Fugu, Wubu, and Longmen—from 2006 to 2015. The key findings are summarized as follows:

(1) The performance of simulated streamflow improves with increasing time scale. Specifically, the GloFAS dataset performs the best at the daily scale, while the WGHM dataset outperforms the others at the monthly, annual, and seasonal scales, providing more reliable streamflow simulations.
(2) All three datasets tend to overestimate the total streamflow across all stations.
(3) The three datasets perform better during the wet period, accurately simulating streamflow changes compared to the dry period.
(4) The ability to capture peak flow timing is influenced by the dataset type, temporal scale, and station characteristics. Overall, the datasets perform better at hydrological stations with higher base streamflow, such as Longmen and Wubu stations.
(5) Simulation errors in the datasets may stem from issues including the quality of meteorological input data, missing or simplified model processes, and improper model parameterization.

The performance of the three grid-based runoff datasets varies in the middle Yellow River Basin. In our findings, the WGHM dataset generally reflects the actual streamflow changes effectively, thus providing a valuable reference for streamflow prediction and hydrological simulation studies in arid inland areas where hydrological stations are sparse. However, to obtain more accurate simulated streamflow and fill the gaps in data-scarce regions, remote sensing streamflow estimation methods can be integrated to generate simulated streamflow at satellite virtual stations [1], facilitating further validation, processing, and optimization of the existing runoff datasets. Future research will focus on refining and enhancing these runoff datasets to achieve more accurate data sources, thereby improving the accuracy and reliability of watershed streamflow simulations.

Author Contributions: Conceptualization, C.H.; methodology, C.H. and H.C.; validation, H.C.; formal analysis, H.C.; data curation, H.C.; writing—original draft, H.C.; writing—review and editing, C.H.; project administration, C.H.; funding acquisition, C.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Grant No. 42471023 and U2243205).

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Huang, C.; Li, Y.; Tarpanelli, A.; Wang, N.; Chen, Y. Observing river discharge from space: Challenges and opportunities. *Innov. Geosci.* **2024**, *2*, 100076. [CrossRef]
2. Yang, D.W.; Xu, Z.X.; Li, Z.; Yuan, X.; Wang, L.; Liao, C.; Tian, F.; Tian, L.; Long, D.; Tang, Q. Progress and prospect of hydrological sciences. *Prog. Geogr.* **2018**, *37*, 36–45.
3. Lohse, K.A.; Brooks, P.D.; McIntosh, J.C.; Meixner, T.; Huxman, T.E. Interactions between biogeochemistry and hydrologic systems. *Annu. Rev. Environ. Resour.* **2009**, *34*, 65–96. [CrossRef]

4. Ahmad, T.; Pandey, A.C.; Kumar, A.; Tirkey, A. Understanding the role of surface runoff in potential flood inundation in the Kashmir valley, Western Himalayas. *Phys. Chem. Earth Parts A/B/C* **2023**, *131*, 103423. [CrossRef]
5. Gu, L.; Chen, J.; Yin, J.; Xu, C.Y.; Zhou, J. Responses of precipitation and runoff to climate warming and implications for future drought changes in China. *Earth's Future* **2020**, *8*, e2020EF001718. [CrossRef]
6. Do, H.X.; Gudmundsson, L.; Leonard, M.; Westra, S. The Global Streamflow Indices and Metadata Archive (GSIM)–Part 1: The production of a daily streamflow archive and metadata. *Earth Syst. Sci. Data* **2018**, *10*, 765–785. [CrossRef]
7. Gudmundsson, L.; Do, H.X.; Leonard, M.; Westra, S. The Global Streamflow Indices and Metadata Archive (GSIM)–Part 2: Quality control, time-series indices and homogeneity assessment. *Earth Syst. Sci. Data* **2018**, *10*, 787–804. [CrossRef]
8. Beck, H.E.; van Dijk, A.I.J.M.; De Roo, A.; Miralles, D.G.; McVicar, T.R.; Schellekens, J.; Bruijnzeel, L.A. Global-scale regionalization of hydrologic model parameters. *Water Resour. Res.* **2016**, *52*, 3599–3622. [CrossRef]
9. Gao, H.; Zhao, F. Global-scale hydrological models: Opportunities, challenges, and prospects. *J. Glaciol. Geocryol.* **2020**, *42*, 224–233. (In Chinese)
10. Hou, Y.; Guo, H.; Yang, Y.; Liu, W. Global evaluation of runoff simulation from climate, hydrological and land surface models. *Water Resour. Res.* **2023**, *59*, e2021WR031817. [CrossRef]
11. Gudmundsson, L.; Tallaksen, L.M.; Stahl, K.; Clark, D.B.; Dumont, E.; Hagemann, S.; Bertrand, N.; Gerten, D.; Heinke, J.; Hanasaki, N.; et al. Comparing large-scale hydrological model simulations to observed runoff percentiles in Europe. *J. Hydrometeorol.* **2012**, *13*, 604–620. [CrossRef]
12. Zaitchik, B.F.; Rodell, M.; Olivera, F. Evaluation of the Global Land Data Assimilation System using global river discharge data and a source-to-sink routing scheme. *Water Resour. Res.* **2010**, *46*. [CrossRef]
13. Zhang, X.; Tang, Q.; Zhang, X.; Lettenmaier, D.P. Runoff sensitivity to global mean temperature change in the CMIP5 Models. *Geophys. Res. Lett.* **2014**, *41*, 5492–5498. [CrossRef]
14. Rakovec, O.; Mizukami, N.; Kumar, R.; Newman, A.J.; Thober, S.; Wood, A.W.; Clark, M.P.; Samaniego, L. Diagnostic evaluation of large-domain hydrologic models calibrated across the contiguous United States. *J. Geophys. Res. Atmos.* **2019**, *124*, 13991–14007. [CrossRef]
15. Beck, H.E.; Van Dijk, A.I.J.M.; De Roo, A.; Dutra, E.; Fink, G.; Orth, R.; Schellekens, J. Global evaluation of runoff from 10 state-of-the-art hydrological models. *Hydrol. Earth Syst. Sci.* **2017**, *21*, 2881–2903. [CrossRef]
16. Sikder, M.S.; David, C.H.; Allen, G.H.; Qiao, X.; Nelson, E.J.; Matin, M.A. Evaluation of available global runoff datasets through a river model in support of transboundary water management in South and Southeast Asia. *Front. Environ. Sci.* **2019**, *7*, 171. [CrossRef]
17. Bai, P.; Liu, X.; Yang, T.; Liang, K.; Liu, C. Evaluation of streamflow simulation results of land surface models in GLDAS on the Tibetan plateau. *J. Geophys. Res. Atmos.* **2016**, *121*, 12180–12197. [CrossRef]
18. Yang, Y.; Pan, M.; Lin, P.; Beck, H.E.; Zeng, Z.; Yamazaki, D.; David, C.H.; Lu, H.; Yang, K.; Hong, Y.; et al. Global reach-level 3-hourly river flood reanalysis (1980–2019). *Bull. Am. Meteorol. Soc.* **2021**, *102*, E2086–E2105. [CrossRef]
19. Harrigan, S.; Zsoter, E.; Alfieri, L.; Prudhomme, C.; Salamon, P.; Wetterhall, F.; Barnard, C.; Cloke, H.; Pappenberger, F. GloFAS-ERA5 operational global river discharge reanalysis 1979–present. *Earth Syst. Sci. Data* **2020**, *12*, 2043–2060. [CrossRef]
20. Müller Schmied, H.; Cáceres, D.; Eisner, S.; Flörke, M.; Herbert, C.; Niemann, C.; Peiris, T.A.; Popat, E.; Portmann, F.T.; Reinecke, R.; et al. The global water resources and use model WaterGAP v2. 2d: Model description and evaluation. *Geosci. Model Dev.* **2021**, *14*, 1037–1079. [CrossRef]
21. Liang, X.; Lettenmaier, D.P.; Wood, E.F.; Burges, S.J. A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *J. Geophys. Res. Atmos.* **1994**, *99*, 14415–14428. [CrossRef]
22. David, C.H.; Maidment, D.R.; Niu, G.Y.; Yang, Z.-L.; Habets, F.; Eijkhout, V. River network routing on the NHDPlus dataset. *J. Hydrometeorol.* **2011**, *12*, 913–934. [CrossRef]
23. Hersbach, H.; de Rosnay, P.; Bell, B.; Schepers, D.; Simmons, A.; Soci, C.; Abdalla, S.; Alonso-Balmaseda, M.; Balsamo, G.; Bechtolg, P.; et al. Operational Global Reanalysis: Progress, Future Directions and Synergies with NWP. ERA Report. 2018. Available online: https://www.ecmwf.int/en/elibrary/80922-operational-global-reanalysis-progress-future-directions-and-synergies-nwp (accessed on 5 February 2024).
24. Van Der Knijff, J.M.; Younis, J.; De Roo, A.P.J. LISFLOOD: A GIS-based distributed model for river basin scale water balance and flood simulation. *Int. J. Geogr. Inf. Sci.* **2010**, *24*, 189–212. [CrossRef]
25. Kim, H. Global Soil Wetness Project Phase 3 Atmospheric Boundary Conditions (Experiment 1). Available online: https://search.diasjp.net/en/dataset/GSWP3_EXP1_Forcing (accessed on 5 February 2024).
26. Cucchi, M.; Weedon, G.P.; Amici, A.; Bellouin, N.; Lange, S.; Müller Schmied, H.; Hersbach, H.; Buontempo, C. WFDE5: Bias-adjusted ERA5 reanalysis data for impact studies. *Earth Syst. Sci. Data* **2020**, *12*, 2097–2120. [CrossRef]
27. Gupta, H.V.; Kling, H.; Yilmaz, K.K.; Martinez, G.F. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *J. Hydrol.* **2009**, *377*, 80–91. [CrossRef]

28. Nash, J.E.; Sutcliffe, J.V. River flow forecasting through conceptual models part I—A discussion of principles. *J. Ofhydrology* **1970**, *10*, 282–290. [CrossRef]
29. Fekete, B.M.; Vörösmarty, C.J.; Roads, J.O.; Willmott, C.J. Uncertainties in precipitation and their impacts on runoff estimates. *J. Clim.* **2004**, *17*, 294–304. [CrossRef]
30. Müller Schmied, H.; Eisner, S.; Franz, D.; Wattenbach, M.; Portmann, F.T.; Flörke, M.; Döll, P. Sensitivity of simulated global-scale freshwater fluxes and storages to input data, hydrological model structure, human water use and calibration. *Hydrol. Earth Syst. Sci.* **2014**, *18*, 3511–3538. [CrossRef]
31. Weiland FC, S.; Vrugt, J.A.; Weerts, A.H.; Bierkens, M.F. Significant uncertainty in global scale hydrological modeling from precipitation data errors. *J. Hydrol.* **2015**, *529*, 1095–1115. [CrossRef]
32. Beck, H.E.; Wood, E.F.; Pan, M.; Fisher, C.K.; Miralles, D.G.; Van Dijk, A.I.; McVicar, T.R.; Adler, R.F. MSWEP V2 global 3-hourly 0.1 precipitation: Methodology and quantitative assessment. *Bull. Am. Meteorol. Soc.* **2019**, *100*, 473–500. [CrossRef]
33. Yang, Y.; Wu, J.; Bai, L.; Wang, B. Reliability of gridded precipitation products in the Yellow River Basin, China. *Remote Sens.* **2020**, *12*, 374. [CrossRef]
34. Wu, H.; Liu, D.; Huang, Q.; Zheng, H.; Zou, H.; Ye, N. Accuracy Assessment and Alternative Study of Precipitation Products in the Loess Plateau. *J. Hydroelectr. Eng.* **2021**, *40*, 31–40. (In Chinese) [CrossRef]
35. Huang, R.; Yong, B.; Huang, F.; Wu, H.; Shen, Z.; Qian, D. A comprehensive investigation of three long-term precipitation datasets: Which performs better in the Yellow River basin? *Int. J. Climatol.* **2024**, *44*, 1302–1325. [CrossRef]
36. Liang, X.; Wood, E.F.; Lettenmaier, D.P. Surface soil moisture parameterization of the VIC-2L model: Evaluation and modification. *Glob. Planet. Change* **1996**, *13*, 195–206. [CrossRef]
37. Beven, K.J.; Kirkby, M.J. A physically based, variable contributing area model of basin hydrology/Un modèle à base physique de zone d'appel variable de l'hydrologie du bassin versant. *Hydrol. Sci. J.* **1979**, *24*, 43–69. [CrossRef]
38. Niu, G.Y.; Yang, Z.L.; Dickinson, R.E.; Gulden, L.E. A simple TOPMODEL-based runoff parameterization (SIMTOP) for use in global climate models. *J. Geophys. Res. Atmos.* **2005**, *110*. [CrossRef]
39. Kirkby, M. Hillslope runoff processes and models. *J. Hydrol.* **1988**, *100*, 315–339. [CrossRef]
40. Pilgrim, D.H.; Chapman, T.G.; Doran, D.G. Problems of rainfall-runoff modelling in arid and semiarid regions. *Hydrol. Sci. J.* **1988**, *33*, 379–400. [CrossRef]
41. Hirpa, F.A.; Salamon, P.; Beck, H.E.; Gulden, L.E. Calibration of the Global Flood Awareness System (GloFAS) using daily streamflow data. *J. Hydrol.* **2018**, *566*, 595–606. [CrossRef]
42. GRDC: Long-Term Statistics and Annual Characteristics of GRDC Time Series Data/Online Provided by the Global Runoff Data Centre of WMO. Available online: https://grdc.bafg.de/data/data_portal/index.html (accessed on 5 February 2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.