remote sensing

MDPI

Article

Assessing the Reliability of Multi-Source Gridded Precipitation Products for Use as Inputs to Hydrological and Hydroclimatic Applications: A Case Study of the Helmand River Basin, Afghanistan

Saad Mohammad Qurishi [1], Ahmed M. Al-Areeq [1,2,*], Hatim O. Sharif [3] and Shakhawat Chowdhury [1,4]

[1] Department of Civil and Environmental Engineering, King Fahd University of Petroleum & Minerals (KFUPM), Dhahran 31261, Saudi Arabia
[2] Interdisciplinary Research Center for Membranes and Water Security, King Fahd University of Petroleum & Minerals (KFUPM), Dhahran 31261, Saudi Arabia
[3] Department of Civil and Environmental Engineering, University of Texas at San Antonio, San Antonio, TX 78249, USA
[4] Interdisciplinary Research Center for Construction and Building Materials, King Fahd University of Petroleum & Minerals (KFUPM), Dhahran 31261, Saudi Arabia
* Correspondence: ahmed.areeq@kfupm.edu.sa

Highlights

What are the main findings?

- A cross-family evaluation of satellite, reanalysis, gauge-based, and multi-source precipitation products reveals clear differences in performance associated with temporal scale, season, and elevation, with monthly and wet-season estimates substantially more robust than daily, dry-season, and high-elevation estimates.
- MSWEP and ERA5-Land emerge as the most robust precipitation inputs, maintaining the best overall balance of temporal agreement, error, bias, rainfall-event detection, and spatial consistency across the Helmand River Basin.

What are the implications of the main findings?

- By identifying where and when precipitation products are reliable, the findings reduce input uncertainty for hydrological, drought, climate-risk, and water-resources assessments in the data-scarce Helmand River Basin.
- The integrated framework provides a transferable basis for selecting fit-for-purpose precipitation products in sparsely gauged, topographically complex basins, where reliability cannot be inferred from a single metric or timescale.

Check for updates

Academic Editor: Konstantinos X. Soulis

Received: 20 August 2026
Revised: 22 October 2026
Accepted: 22 October 2026
Published: 25 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Abstract

Reliable precipitation datasets are essential for basin-scale water resources assessment, climate analysis, flood and drought-related studies, whereas estimation remains challenging in arid to semi-arid, topographically complex, and data-scarce regions. This issue is particularly important in the Helmand River Basin (HRB), Afghanistan, where uncertainty in precipitation input can directly affect hydrological and hydroclimatic assessments. This study evaluated eight multi-source gridded precipitation products against gauge observations from 18 hydro-meteorological stations across the HRB. The evaluation covered daily and monthly timescales, wet and dry season conditions, rainfall-event thresholds, and elevation settings using continuous, categorical, and integrated multi-metric analyses. The results indicate that precipitation product performance in the HRB varies with temporal scale, seasonal regime, and elevation setting. In general, monthly and wet-season evaluations showed better agreement than daily and

dry-season evaluations; however, this pattern reflects the combined influence of temporal scale, native product resolution, and spatial extraction method. Among the evaluated datasets, MSWEP and ERA5-Land showed the most balanced overall performance, combining better temporal agreement, lower error, lower bias, and higher rainfall-event detection across stations and elevation classes. CHIRPS also performed well, particularly at monthly and seasonal scales, whereas IMERG and PERSIANN-CDR showed moderate performance. Conversely, CMORPH and GPCC Products were the least reliable because of persistent underestimation in the former and inflated variability and overestimation in the latter. The results further indicate that dataset skill is not spatially uniform, with weaker performance generally associated with high-elevation stations. Overall, this study establishes a basin-specific framework for precipitation dataset selection in the HRB and provides a reproducible basis for selecting fit-for-purpose precipitation inputs for future water resources and hydroclimatic studies in this data-scarce region.

Keywords: data-scarce region; multi-metric assessment; gridded precipitation; reliability evaluation; multi-timescales

1. Introduction

Precipitation is a fundamental component of the water cycle and a key variable in basin-scale water resources assessment, climate analysis, disaster prevention and mitigation [1]. However, precipitation exhibits high spatial and temporal variability, and its measurement becomes difficult in regions characterized by complex terrain, distinct climatic contrasts, and sparse observations [2]. Although gauge records remain the primary reference source, their spatial representativeness is often limited, particularly in mountainous, remote, and data-scarce settings. Consequently, gridded precipitation products derived from satellite retrievals, reanalysis systems, gauge interpolation, and multi-source merging are widely used because they provide spatially continuous and temporally consistent estimates over large areas [3]. However, their practical value depends on regional reliability. Product performance may vary with input data, correction procedure, spatial resolution, temporal aggregation, seasonality, and terrain complexity. This issue is particularly important in arid to semi-arid basins with high elevation gradients, where precipitation uncertainty can directly influence basin-scale interpretation and subsequent applications. Evidence from different regions demonstrates that gridded precipitation performance is not fixed across settings. Multi-product evaluations in China, Sicily, Texas, West and Central Africa, and other basins have shown that accuracy varies with climate regime, terrain, season, timescale, event type, and evaluation method. These studies also indicate that gauge-based, bias-corrected, or merged products often outperform raw datasets, although no single product is consistently superior across all regions or applications [4–7]. Recent evaluations further show that product performance can differ strongly between daily, monthly, seasonal, and extreme-event scales, especially in mountainous and data-scarce regions [8,9]. This dependence becomes more pronounced in arid, mountainous, and data-scarce environments. A multidimensional assessment over the Tibetan Plateau showed that dataset rankings changed with terrain, timescale, uncertainty treatment, and evaluation dimension, demonstrating that a single metric may provide only a partial view of product quality [10]. Likewise, a recent comparison of merged precipitation products using fully independent gauges showed that merging can improve performance, particularly in observation-scarce regions, but also revealed persistent weaknesses in rain/no-rain detection and the sensitivity of some products to methodological assumptions [11]. These findings are consistent with recent evidence that merged and ensemble products, including MSWEP and IMERG-based datasets, often improve temporal agreement but may still show bias under complex terrain, sparse gauges, or limited local correction [12,13]. Together, these studies highlight

the need to evaluate precipitation products under the specific hydroclimatic, terrain, and data availability conditions in which they are intended to be used. In Afghanistan, the available evidence remains limited. The main national-scale evaluation compared 13 global gridded precipitation datasets against 75 observation stations across daily, monthly, wet-season, and dry-season scales. That study showed that performance varied by basin, climate zone, elevation, and timescale, and identified CFSR and PERSIANN as the best overall products across the country. It also indicated that the HRB had its own leading products, headed by CFSR, followed by CHIRPS-v2, IMERG, PERSIANN-CDR, and PERSIANN-CCS [14]. This provides national-scale evidence and identifies leading products for the HRB. However, its primary objective was to compare product performance across Afghanistan rather than to develop a systematic evaluation framework for basin-specific product selection. The closest studies in the HRB have addressed related but different questions. For example, one study evaluated three global climate datasets over the HRB and applied bias correction and trend analysis. The study found that CRU was the best raw precipitation dataset while Bi-LSTM was the most accurate bias-correction method. However, that analysis considered only three datasets and focused on correction and climate trends rather than broad product intercomparison for precipitation reliability [15]. Other HRB studies have examined hydroclimatic conditions and drought, using water-yield modelling or station-based SPI analysis. These studies are useful for basin background, but they do not provide a multi-product, observation-based evaluation of precipitation-product reliability in the HRB [16,17]. Taken together, these studies indicate that gridded precipitation performance is region-specific and must be evaluated under local climatic and physiographic conditions. Although an Afghanistan-wide evaluation is available and several HRB studies have used precipitation products for correction, drought assessment, or hydrological analysis, no study has systematically evaluated precipitation products for the HRB using an integrated, multi-criteria framework that jointly examines temporal agreement, error characteristics, rainfall-event detection, seasonal behavior, and elevation-dependent performance. Consequently, uncertainty remains regarding which precipitation products provide the most reliable and balanced performance across the multiple evaluation dimensions required for basin-scale hydrological and hydroclimatic applications in the HRB. Such an evaluation is particularly important in a large arid to semi-arid basin where high elevation gradients, pronounced spatial rainfall variability, and sparse observations amplify precipitation uncertainty. In such settings, differences among precipitation datasets can lead to materially different conclusions about basin water availability, runoff behavior, and drought severity [14–17]. In this context, the present study evaluates the reliability of multi-source gridded precipitation products against gauge observations in the HRB in order to identify the datasets that most faithfully represent precipitation across the basin. The evaluation compares products from the main precipitation data families using basin station records across daily and monthly timescales, wet and dry seasons, rainfall occurrence thresholds, and elevation settings. The resulting evaluation provides a transparent and reproducible basis for selecting precipitation products according to the performance requirements of different hydrological and hydroclimatic applications. By integrating complementary evaluation criteria, it establishes a systematic framework for assessing precipitation-product reliability beyond single-metric or single-timescale assessments.

2. Materials and Methods

2.1. Study Area

The HRB is the largest river basin in Afghanistan and constitutes a major hydroclimatic unit of southern and southwestern Afghanistan. The basin covers about 367,000 km^2 and lies approximately between 29–35°N and 60–69°E [18,19]. As shown in Figure 1, the HRB extends from high-elevation headwater regions in the east and northeast toward

broad lowland areas in the southwest, while the hydro-meteorological stations used in this study are distributed across the basin to represent these contrasting environments. Elevation declines from nearly 4000 m above sea level (a.s.l) in the upper basin to around 430 m a.s.l. in the lower reaches [19,20]. This topographic gradient highly controls both precipitation and temperature patterns. The upper basin is relatively cool and receives greater precipitation, including an important winter contribution, whereas the downstream lowlands are markedly warmer and drier, with higher moisture limitations and higher atmospheric evaporative demand. Climatically, the HRB is predominantly arid to semi-arid, and its precipitation regime is both strongly seasonal and spatially uneven. Mean annual precipitation decreases sharply from about 300 mm in the northeastern highlands to nearly 50 mm in the southwestern lowlands, indicating a pronounced basin-wide hydroclimatic gradient [15]. Most precipitation occurs during the cold season, while the dry season is long and dominant over much of the basin. Hydrologically, the basin is sustained by several major river systems, including the Helmand, Arghandab, Tarnak, Farah, and Harut rivers, which are fed mainly by rainfall and snowmelt from the surrounding highlands [21]. Among them, the Helmand River is the principal river course and extends for about 1050–1100 km before reaching the Sistan plain and the Hamun wetlands [15,19]. These topographic, climatic, and hydrological characteristics define the HRB as a spatially diverse basin with high elevation and moisture gradients, providing an appropriate regional setting for basin-scale hydroclimatic analysis. The study area map (Figure 1) was prepared using ArcMap version 10.8.2 (Esri, Redlands, CA, USA).

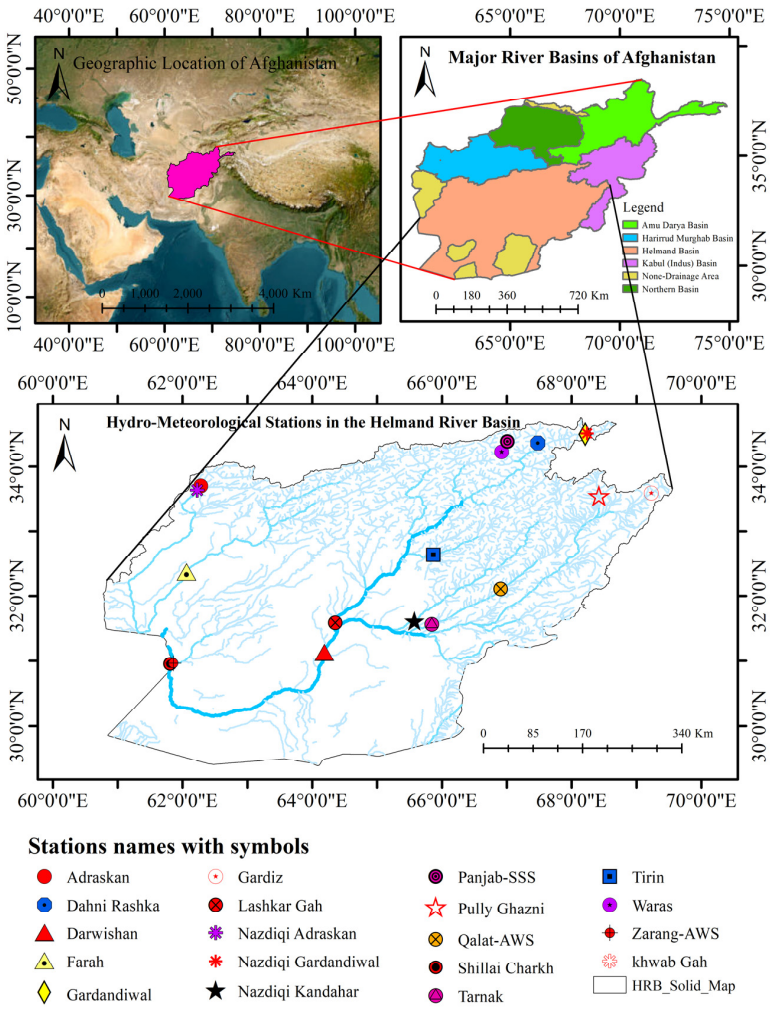


Figure 1. Study area.

2.2. Data Source

2.2.1. Gauge-Based Precipitation Observations

Gauge-based precipitation observations were obtained from the Ministry of Energy and Water (MoEW), in Afghanistan, using records from 18 automatic hydrometeorological stations that were distributed across the HRB as shown in Figure 1. The station records used in the main evaluation cover 2008–2025, with dataset length varying among stations according to data availability. The observations were originally recorded at 15 min intervals and provided by MoEW in aggregated tabular form, including daily maximum, daily total precipitation, monthly, and yearly max and total. Some station-year files provided by MoEW include remarks indicating that missing observations had been filled using a neighboring-station average method before data release. However, the source files do not document the affected observations or the complete gap-filling procedure. No additional gap filling or modification was performed in this study. Daily analyses used the available daily observations directly, whereas monthly analyses used the native monthly totals supplied by MoEW. Seasonal and annual analyses were derived directly from the native monthly records supplied by MoEW. Accordingly, no additional completeness criteria were imposed, and monthly totals were not recalculated from the daily observations. All evaluations were restricted to the valid overlapping observation periods with each gridded product. The geographic characteristics, observation record, available days, and missing data percentages of the gauge stations are summarized in Table 1.

Table 1. Geographic characteristics of gauge station data used in this study.

No.	Station	Latitude (°N)	Longitude (°E)	Elevation (m a.s.l.)	Observation Period	Available Days	Missing Days	Missing (%)
1	Adraskan	33.7046	62.2802	1343	2008–2019	4048	28	0.69
2	Dahan-i-Rashqa	34.3616	67.4758	2554	2009–2019	3864	29	0.75
3	Darwishan	31.1304	64.1883	706	2008–2017	3531	0	0
4	Farah	32.3645	62.0618	653	2009–2019	3864	30	0.78
5	Gardiz	33.5936	69.2287	2309	2008–2021	4626	25	0.54
6	Khwabgah	30.7977	61.7710	495	2009–2019	3895	30	0.77
7	Lashkargah	31.5829	64.3548	776	2008–2025	6330	11	0.17
8	Nazdik-i-Adraskan	33.640525	62.2256595	1382	2008–2019	4017	29	0.72
9	Nazdik-i-Gardandiwal	34.5103	68.2396	3040	2009–2019	3864	60	1.55
10	Nazdik-i-Kandahar	31.6155	65.5761	981	2009–2019	3681	72	1.96
11	Panjab-SSS	34.3858	67.0085	2837	2012–2021	2556	45	1.76
12	Pul-i-Ghazni	33.5559	68.4200	2194	2009–2018	3348	43	1.28
13	Qalat-AWS	32.1063	66.9069	1571	2012–2021	3347	146	4.36
14	Shila-i-Charkh	30.9521	61.8075	489	2008–2019	4048	86	2.12
15	Tarnak	31.5602	65.8432	1027	2009–2019	3711	63	1.7
16	Tirin	32.6327	65.8666	1297	2013–2019	2372	10	0.42
17	Waras	34.2250	66.9199	2426	2008–2019	3986	70	1.76
18	Zaranj-AWS	30.9636	61.8567	487	2012–2018	1977	56	2.83

Only overlapping periods between each point-based gauge observation and gridded precipitation product were used for comparison. This approach ensured that all performance metrics were calculated from temporally consistent gauge–product pairs without additional modification of the source observations. For the extended historical analysis (1979–2009), archived monthly precipitation records from 14 MoEW stations were additionally used. These records were obtained as archived monthly summaries provided by

MoEW and served only to examine long-term seasonal and interannual variability. Unlike the recent 2008–2025 observation dataset, the source files do not document the historical gap-filling or quality-control procedures. Accordingly, the historical analysis is presented as supplementary evidence and was not used for the primary product ranking.

2.2.2. Multi-Source Gridded Precipitation Products

This study employed a set of gridded precipitation products drawn from four major data families: satellite-based, reanalysis, gauge-based gridded, and multi-source merged datasets. The products were selected to represent the main classes of precipitation information currently used in hydrological and hydroclimatic studies, while also capturing differences in retrieval philosophy, correction strategy, spatial resolution, temporal resolution, and record length. Such diversity is important because the performance of gridded precipitation products is often conditioned by the way precipitation is observed, modeled, merged, or bias-corrected, particularly in regions with high topographic and climatic gradients such as the HRB. The selected datasets were not intended to represent the complete range of available precipitation products, but rather to provide representative coverage of the principal precipitation-product families currently used in hydrological and hydroclimatic studies. Accordingly, dataset selection was guided by four criteria: representation of the major product families, current scientific use, temporal availability for long-term evaluation, and methodological diversity in precipitation estimation. Selection was therefore based on predefined study-design criteria rather than previously reported performance rankings. The principal characteristics of all datasets, including their native resolution, temporal availability, native coverage period, full dataset names, and primary sources, are summarized in Table 2. For the extended historical analysis presented in Section 4.2, the overall analysis window was 1979–2009, but the effective monthly period varied among products. The gauge observations, ERA5-Land, GPCC-FD-M, and TerraClimate covered January 1979–December 2009. MSWEP covered February 1979–December 2009 because January 1979 was unavailable. CHIRPS covered January 1981–December 2009, PERSIANN-CDR January 1983–December 2009, and both IMERG and CMORPH January 1998–December 2009. Annual, seasonal, and anomaly series were derived from the available monthly records within these product-specific periods.

The selected products include CHIRPS v3, MSWEP v2.8, ERA5-Land, IMERG Final Run v07, CMORPH CDR, PERSIANN-CDR, GPCC datasets, and TerraClimate. Collectively, these datasets span fine to coarse-resolution grids and provide both daily and monthly precipitation datasets, thereby allowing evaluation across multiple temporal scales relevant to hydrological applications. Their inclusion was intended to ensure that the intercomparison does not rely on a single product type but instead reflects the broader range of gridded precipitation datasets commonly used in practice. Within this set, CHIRPS v3 and IMERG Final Run v07 represent gauge-corrected products that combine satellite-derived precipitation estimates with station information to improve spatial realism and bias performance. CMORPH CDR and PERSIANN-CDR provide long-term satellite climate data records, which are particularly valuable for assessing temporal consistency and large-scale precipitation behavior. ERA5-Land represents the reanalysis family, offering dynamically consistent and spatially continuous precipitation estimates derived from land-surface modeling and atmospheric forcing. MSWEP v2.8 represents a multi-source ensemble framework that integrates several precipitation inputs into a single product, whereas the GPCC products were treated as distinct datasets. For the daily evaluation, GPCC Full Data Daily (GPCC-FD; 1.0°) was used through December 2020, after which GPCC First Guess Daily (GPCC-FG; 1.0°) was used from January 2021 onward following product availability. This procedure produced the continuous daily GPCC evaluation series, denoted GPCC-FD/FG-D. Each product was evaluated only within its

native availability period; no averaging or blending of the two daily GPCC products was performed. Potential discontinuity at the January 2021 transition was assessed by comparing December 2020–January 2021 precipitation changes with those from 12 historical December–January transitions (2008–2020) at the same station grid cells. The changes fell within the historical ranges at 17 of the 18 station locations, with only Qalat-AWS showing a decrease beyond its historical range (−29.34 mm versus a historical minimum of −23.12 mm). This comparison revealed no widespread anomalous change at the transition, although product-related differences cannot be ruled out from the boundary assessment alone. The monthly, wet-season, and dry-season evaluations used GPCC Full Data Monthly Version 2022 (GPCC-FD-M; 0.25°). Accordingly, differences between daily and monthly GPCC results were not interpreted as temporal-scale changes within a single product. TerraClimate was included as a widely used monthly hydroclimatic dataset with relatively fine spatial resolution and broad application in environmental and water-resources studies. Collectively, these products provide a representative cross-section of contemporary gridded precipitation datasets and establish a robust basis for benchmarking precipitation performance over the HRB.

Table 2. Gauge observations and eight gridded precipitation products summaries from multiple families (gauge-based, satellite, reanalysis, and multi-source merged), including their abbreviations, native resolutions, availability periods, and primary sources.

Dataset	Abbreviation	Data Type	Spatial Resolution	Temporal Resolution	Time Coverage	Data Source
Automatic Hydrometeorology Stations	AHS	Observation	Point	Daily	1979–2025	Ministry of energy and water (MoEW), Afghanistan
Climate Hazards Group InfraRed Precipitation with Stations	(CHIRPS) v3	Satellite + gauge merged	0.05° × 0.05°	Daily and Monthly	1981–Present	[22,23]
Multi-Source Weighted-Ensemble Precipitation v2.8	(MSWEP)	Multi-source merged	0.1° × 0.1°	Daily and Monthly	1979–Present	[24]
ERA5-Land reanalysis	ERA5-Land	Reanalysis	0.1° × 0.1°	Monthly and daily	1950–Present	[25,26]
GPM IMERG Final Run v07	IMERG	Satellite (gauge-corrected)	0.1° × 0.1°	Daily and Monthly	1998–Present	[27,28]
NOAA CMORPH Climate Data Record (CDR)	CMORPH	Satellite CDR (bias-corrected)	0.25° × 0.25°	Daily	1998–Present	[29,30]
Climate Data Record (PERSIANN-CDR)	PERSIANN	Satellite CDR	0.25°	Daily	1983–Present	[31]
Global Precipitation Climatology Centre Full Data Daily	GPCC-FD-D	Gauge-based gridded	1.0° × 1.0°	Daily	1982–2020	[32]
TerraClimate	TerraClimate	TerraClimate	1/24° (~4 km)	Monthly	1958–Present	[33]
GPCC First Guess Daily Product	GPCC-FG-D	Gauge-based gridded (near-real-time)	1.0° × 1.0°	Daily	2009–Present	[34]
GPCC Full Data Monthly Product Version 2022	GPCC-FD-M	Gauge-based gridded	0.25° × 0.25°	Monthly	1891–2020	[35]

3. Methodology

3.1. Data Preparation and Processing

To establish a consistent basis for intercomparison, the gauge observations and gridded precipitation datasets were organized into a common point-based and time-consistent frame-

work over the HRB. The gauge records were treated as the reference dataset, whereas the gridded precipitation products were downloaded in NetCDF format and extracted at the coordinates of the observation stations using station latitude and longitude. Because gauge observations represent point measurements whereas gridded precipitation products represent areal grid-cell averages, the spatial matching procedure was selected according to characteristics of each temporal scale. For daily precipitation, values were obtained from the nearest grid cell in order to preserve the original day-to-day rainfall signal and avoid artificial smoothing of short-term variability. Because the daily evaluation included rainfall-occurrence metrics (POD, FAR, CSI, and HSS), the nearest-grid extraction preserved the original daily precipitation signal and avoided interpolation-induced smoothing, which has been shown to modify rain–no-rain boundaries and influence categorical verification statistics [36]. For monthly precipitation, the bilinear interpolation method was applied to derive station-representative values from the surrounding grid cells, because monthly accumulated precipitation exhibits a smoother spatial structure after temporal accumulation, making interpolation more representative of station-scale monthly precipitation [37,38]. All extracted series were then arranged within a unified temporal framework. Comparisons were conducted using only the overlapping dates available for each station–product pair. Therefore, the effective evaluation period varied according to the availability of each gauge record and precipitation product. The coverage periods of the observations and products are summarized in Tables 1 and 2. Additional preprocessing included date formatting, separation of year and month fields, and screening of records prior to comparison. Where products were directly available at both daily and monthly resolutions, each native series was used as provided. For products available only on the daily scale, monthly totals were generated by aggregating the extracted daily values, and annual totals were subsequently derived from the monthly series for interannual assessment. Precipitation data extraction, processing, and statistical analyses were performed using Python (version 3.13.5). For the elevation-based analysis, the 18 stations were grouped into three classes: <1000 m (7 stations), 1000–2000 m (5 stations), and >2000 m (6 stations). Performance metrics were first calculated separately for each station and then averaged within each elevation class. Within-class variability was quantified using the standard deviation of the station-level integrated scores. The elevation-group comparison was intended to summarize station-level integrated performance across elevation classes and characterize elevation-related patterns rather than test for statistical differences among groups. Across products and evaluation conditions, the standard deviations of the integrated scores on the 0–100 scale ranged from 4.37 to 24.38 for the <1000 m class, from 3.24 to 26.33 for the 1000–2000 m class, and from 2.58 to 14.96 for the >2000 m class. Because elevation is associated with precipitation amount, hydroclimatic regime, and station location, while product spatial resolution also varies among datasets, the observed differences were interpreted descriptively and should not be regarded as independent evidence that terrain controls product performance. To define the seasonal framework consistently with the period used in Figure 2, the seasonal framework was derived from the gauge observations over 2008–2019 using the monthly precipitation climatology computed through a station-first averaging approach, in which monthly means were first calculated separately for each station and then averaged across the stations contributing data in each calendar month, so that stations, rather than individual observations received equal weight in the basin estimate [39]. This climatology was used only to define the wet and dry seasons and was not used as the common-period benchmark for product ranking. As shown in Figure 2, the basin exhibits a distinct seasonal precipitation regime, with a mean monthly precipitation of about 15.87 mm, peak monthly values of 40.29 mm in February and 38.77 mm in March, and very low summer amounts of only 1.48–2.37 mm during June to September. This seasonal contrast is further reflected in the mean precipitation of the two defined seasons, with the wet season averaging 27.53 mm and the dry season only 4.21 mm. The error bars

further indicate notable inter-station variability during the wet-season months, highlighting the uneven distribution of precipitation across the basin. Considering this clear climatological separation, November to April was defined as the wet season, whereas May to October was classified as the dry season.

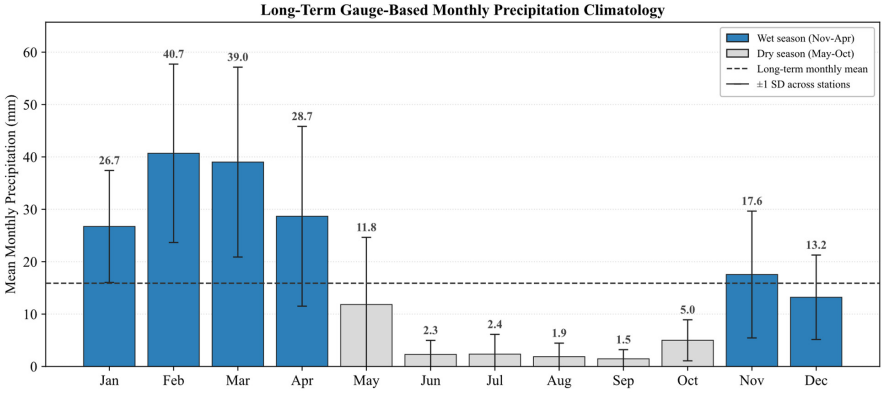


Figure 2. Gauge-based monthly precipitation climatology for the HRB, used to identify the wet and dry seasons from long-term observed precipitation patterns.

This observation-based seasonal partition provides a basin-specific and physically consistent framework for the subsequent temporal evaluation of gridded precipitation products. A schematic overview of the analytical workflow adopted in this study is presented in Figure 3.

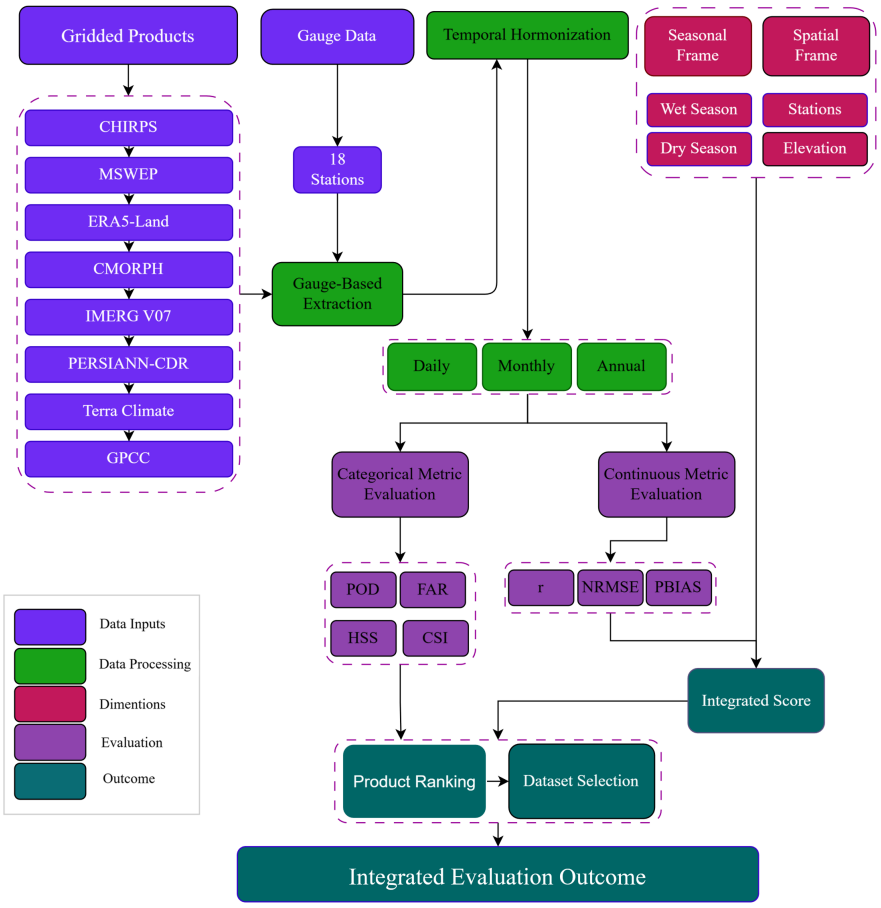


Figure 3. A schematic overview of the integrated methodological approach proposed in this study.

3.2. Evaluation Metrics

The agreement between gridded precipitation estimates and gauge observations was evaluated using two complementary groups of metrics: continuous statistical indices and threshold-based categorical indices. The continuous evaluation employed the Pearson correlation coefficient (r), root mean square error (RMSE), normalized root mean square error (NRMSE), and percent bias (PBIAS). RMSE was retained to report absolute error magnitude, whereas NRMSE was used when comparing errors across different temporal scales [4,5,11,40]. These metrics were selected because precipitation-product performance is inherently multi-dimensional and cannot be represented reliably by a single index. The correlation coefficient (r) measures the strength of temporal agreement and indicates how effectively a product reproduces the observed sequence of fluctuations and variability. RMSE quantifies the overall magnitude of deviation between estimated and observed precipitation, with greater sensitivity to larger errors that are more hydrologically significant. PBIAS further identifies the direction and magnitude of systematic deviation, thereby indicating whether a product persistently overestimates or underestimates precipitation relative to the gauge record. Considered together, these metrics characterize temporal agreement, relative and absolute error magnitude, and systematic bias across daily, monthly, wet-season, and dry-season conditions. To further evaluate daily rainfall occurrence, threshold-based categorical metrics were also adopted. The observed and gridded daily precipitation series were converted into binary rain/no-rain events using thresholds of 0.1, 0.5, and 1.0 mm day^{-1}, representing increasingly strict event definitions [11]. Based on the contingency counts of hits (h), misses (m), false alarms (f), and correct negatives (c), four categorical metrics were calculated: probability of detection (POD), false alarm ratio (FAR), critical success index (CSI), and Heidke skill score (HSS) [41]. POD was used to measure the ability of each product to detect observed rainfall events, whereas FAR quantified the tendency to falsely report rainfall occurrence. CSI was employed to assess overall event-detection skills by jointly accounting for hits, misses, and false alarms, while HSS measured the overall classification skill relative to chance agreement. These categorical indices complement the continuous statistics by resolving an additional aspect of product performance that cannot be inferred from rainfall magnitude agreement alone, namely the reliability of daily rainfall-event identification. The mathematical expressions, interpretation ranges, and best values of all selected metrics are summarized in Table 3.

Table 3. Statistical and categorical metrics used in this study.

Metrics	Equation	Range	Best Value	What It Captures
Pearson correlation coefficient (r)	$r = \frac{\sum_{i=1}^{n}(O_i-\overline{O})(G_i-\overline{G})}{\sqrt{\sum_{i=1}^{n}(O_i-\overline{O})^2}\sqrt{\sum_{i=1}^{n}(G_i-\overline{G})^2}}$	[−1, 1]	1	Temporal agreement
Root mean square error (RMSE)	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(G_i - O_i)^2}$	[0, ∞)	0	Error magnitude
Normalized root mean square error (NRMSE),	$NRMSE(\%) = \frac{RMSE}{\overline{O}} \times 100$	[0, ∞)	0	Relative error magnitude
Percent bias (PBIAS)	$PBIAS = \frac{\sum_{i=1}^{n}(G_i-O_i)}{\sum_{i=1}^{n}O_i} \times 100$	(−∞, ∞)	0	Systematic bias
Probability of detection (POD)	$POD = \frac{h}{h+m}$	[0, 1]	1	Rainfall-event detection
False alarm ratio (FAR)	$FAR = \frac{f}{f+h}$	[0, 1]	0	False rain detection
Heidke skill score (HSS)	$HSS = \frac{2(hc-fm)}{(h+m)(m+c)+(h+f)(f+c)}$	(−∞, 1]	1	Overall classification skill
Critical success index (CSI)	$CSI = \frac{h}{f+h+m}$	[0, 1]	1	Overall event detection skill

In these expressions, O_i and G_i denote observed and gridded precipitation at time i, respectively, $\overline{O}$ and $\overline{G}$ are the corresponding mean values, and n is the number of paired

observations used in the comparison. For the categorical metrics, h, m, f, and c denote hits, misses, false alarms, and correct negatives, respectively. To summarize the continuous evaluation, an integrated performance score was calculated from Pearson correlation, NRMSE, and the absolute value of PBIAS. For each station, temporal scale, season, and product group, the three metrics were independently normalized to a 0–100 scale using min–max normalization. For a metric in which larger values indicate better performance, such as r, the normalized score was calculated as

$$S_x = 100\frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

For metrics in which smaller values indicate better performance, namely NRMSE and |PBIAS|, the score was calculated as

$$S_x = 100\frac{x_{max} - x}{x_{max} - x_{min}} \quad (2)$$

Here, x denotes the metric value, whereas x_{min} and x_{max} denote the minimum and maximum values among the evaluated products within the corresponding station temporal-scale–season group. After normalization, all three metrics were transformed to a common 0–100 scale in which higher scores consistently indicate better performance. Consequently, the normalized scores can be combined directly using equal weighting to calculate the integrated performance score:

$$S_{int} = \frac{S_r + S_{NRMSE} + S_{|PBIAS|}}{3} \quad (3)$$

The integrated score therefore ranges from 0 to 100, with higher values indicating better overall performance within the corresponding comparison group. The integrated score is calculated from the normalized metric scores rather than the raw metric values. Consequently, large raw NRMSE values that may arise under very dry conditions are first transformed to the common 0–100 scoring scale before being equally weighted with the normalized correlation and |PBIAS| scores. If all products had the same value for a metric ($x_{max} = x_{min}$) that metric was assigned a neutral score of 50 to avoid division by zero and to ensure that it did not influence the integrated ranking.

3.3. Trend Estimation and Significance Testing

For the extended 1979–2009 analysis, basin-average monthly precipitation was first calculated for each product by averaging the available station values for each month. Annual totals and wet-season (November–April) and dry-season (May–October) totals were then obtained by summing the corresponding monthly values. Annual anomalies were calculated by subtracting each product's mean annual precipitation from its annual value. Monotonic trends were evaluated using Kendall's tau, representing the rank-based Mann–Kendall trend test rather than linear regression [42,43]. Trend magnitude was estimated using the non-parametric Theil–Sen slope estimator, defined as the median of all pairwise slopes [44,45]. Two-sided p-values were used, and trends were considered statistically significant when ($p < 0.05$). Available non-missing annual and seasonal values were used for each product; therefore, the number of observations varied according to product availability. The resulting trend slopes and significance levels are presented in Section 4.2.

4. Results and Discussion

4.1. Annual Variability

The available-station annual precipitation series exhibits substantial interannual variability over 2008–2025, showing substantial year-to-year variation in the available-station series. The observed series contains alternating wetter and drier years, while the lower values after 2020 coincide with a sharp reduction in station availability. Therefore, this apparent decline is not interpreted as a basin-wide climatic signal. As further indicated in Figure 4, the gridded products differ less in their representation of the broad temporal sequence than in their ability to reproduce annual magnitude.

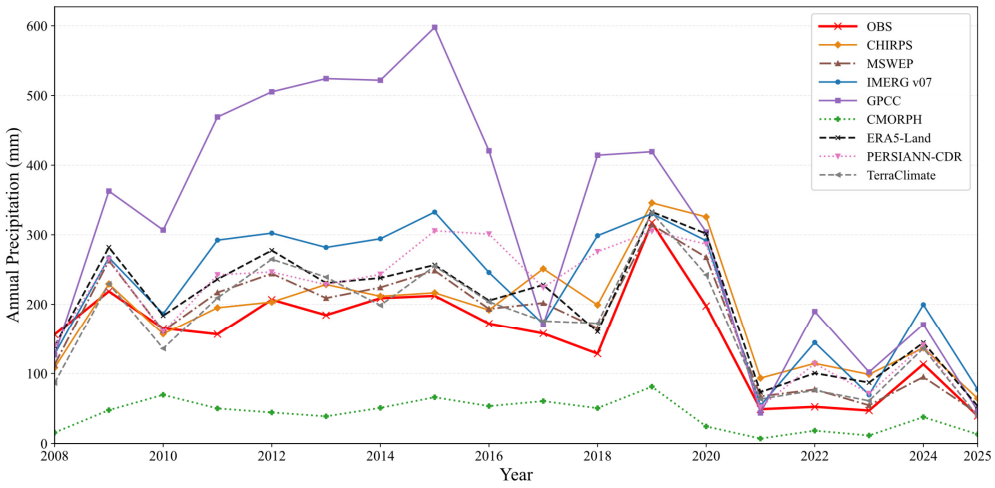


Figure 4. Year-to-year variation in annual precipitation averaged over the stations available in each year across the HRB during (2008–2025), comparing gauge-based observations with gridded precipitation products.

Several datasets broadly preserve the observed wet and dry progression; however, their departures from the available-station gauge-based annual totals remain considerable. GPCC products show the strongest positive deviation and the widest annual range, particularly during (2011–2018), indicating substantial basin-scale overestimation. By contrast, CMORPH persistently underestimates annual precipitation across nearly the entire record and shows limited capacity to reproduce the available-station gauge-based magnitude. CHIRPS, MSWEP, IMERG v07, ERA5-Land, PERSIANN-CDR, and TerraClimate generally reflect the main direction of interannual change more plausibly, although their agreement with the observed magnitude is uneven and varies by year. The central implication of this figure is therefore not a final ranking of products, but a distinction between temporal coherence and magnitude agreement. A product may follow the annual wet and dry sequence while still failing to reproduce the corresponding precipitation totals with sufficient accuracy. This basin-scale perspective provides an initial diagnostic of interannual behavior, whereas the relative performance of the products is quantified more rigorously in the subsequent statistical analyses.

4.2. Extended Historical Seasonal and Trend Consistency

The extended monthly record for 1979–2009 provides a supplementary check on whether the gridded products preserve the observed long-term seasonal pattern, interannual variability, and annual to seasonal direction of change in the HRB (Figure 5). Overall, most products reproduce the main seasonal cycle, with a clear late winter to early spring peak and very low summer precipitation, but they differ in magnitude. CHIRPS, MSWEP, ERA5-Land, and GPCC datasets capture the main wet season peak in March, whereas

IMERG and TerraClimate shift the peak to February, and CMORPH strongly underestimates precipitation throughout the year. The annual variability and anomaly panels illustrate how each product reproduces the observed temporal behavior within its available record. Differences in magnitude and timing are evident among the products. The trend panels summarize the estimated trend direction and statistical significance within each product's effective record. The observed series shows a non-significant annual decline ($p = 0.198$) and a significant dry season decrease ($p = 0.044$), whereas the corresponding product-specific trend estimates show varying degrees of agreement with the observed tendency. CMORPH is the only dataset with significantly increasing annual ($p = 0.045$) and wet-season trends ($p < 0.001$), which contrasts with the observed tendency.

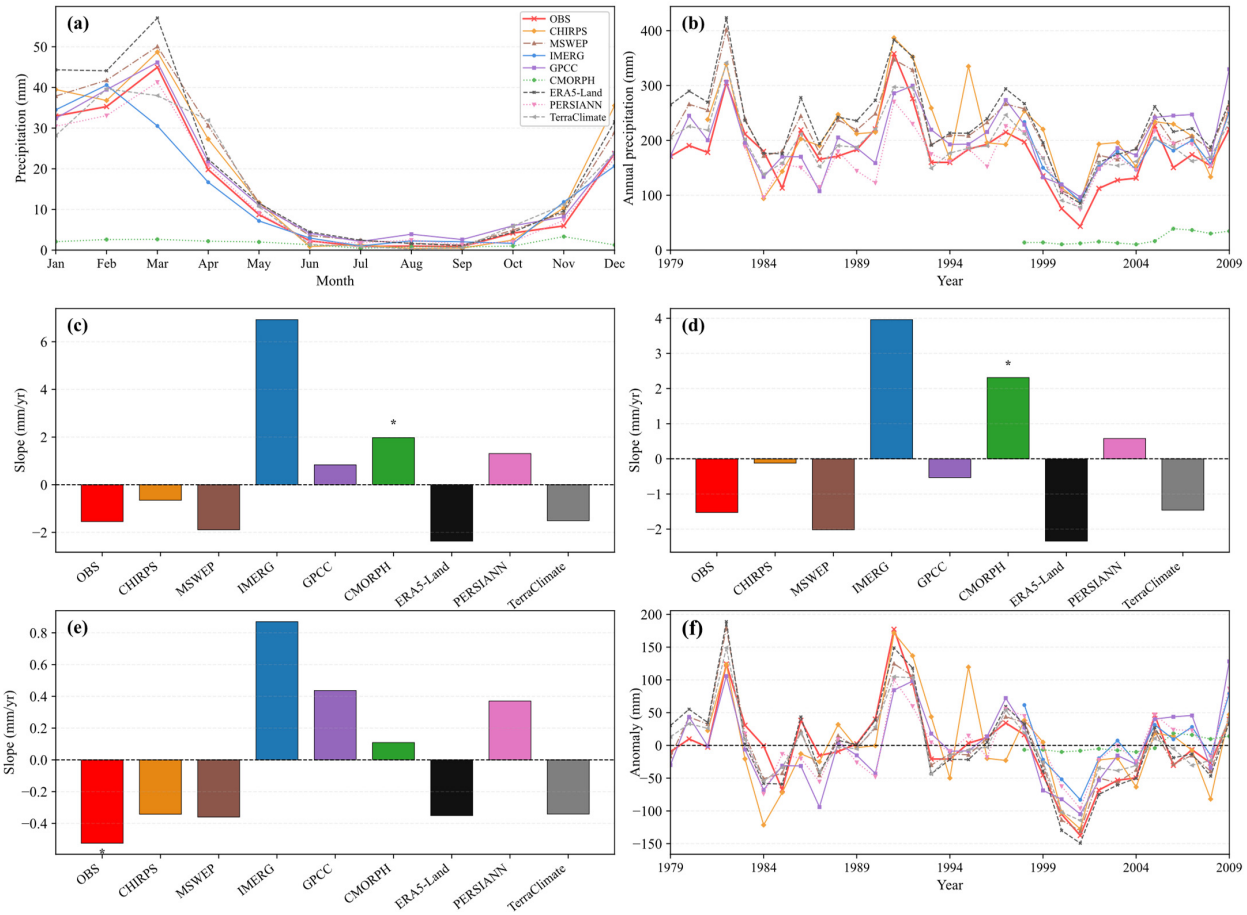


Figure 5. Extended historical precipitation characteristics (1979–2009): (**a**) monthly climatology; (**b**) annual precipitation variability; (**c**) annual trends; (**d**) wet-season trends; (**e**) dry-season trends; and (**f**) annual precipitation anomalies. * denotes a statistically significant trend ($p < 0.05$).

Collectively, this extended historical analysis provides supplementary descriptive evidence regarding long-term seasonal behavior within the available observation period of each dataset. Because the effective record differs among products, these results are interpreted qualitatively and are not used for direct product ranking or quantitative trend comparison. Instead, they provide additional context for interpreting the long-term behavior of the evaluated datasets and complement, rather than influence, the primary benchmarking results.

4.3. Statistical Comparison

The distributional statistics show that product skill changes substantially with both metrics and temporal scale, confirming that no single index is sufficient to characterize

performance over the HRB. Figure 6 summarizes station-level metric distributions computed from the available overlapping gauge–product records during the 2008–2019 evaluation period. Monthly correlation distributions are consistently higher than their daily and dry-season counterparts across all products. This pattern should be interpreted within the context of the evaluation framework, where temporal resolution, spatial extraction procedure, and, for some datasets, the native product differs between the daily and monthly evaluations. In this respect, ERA5-Land and MSWEP display the highest and most stable monthly agreement, while CHIRPS and IMERG also maintain comparatively high correlation. By contrast, CMORPH remains the weakest in terms of temporal agreement, and GPCC datasets show less stable behavior outside the monthly scale. The dry season yields the lowest correlation for all datasets, indicating that weak and intermittent precipitation conditions are more difficult to reproduce than the more coherent wet season and monthly signals.

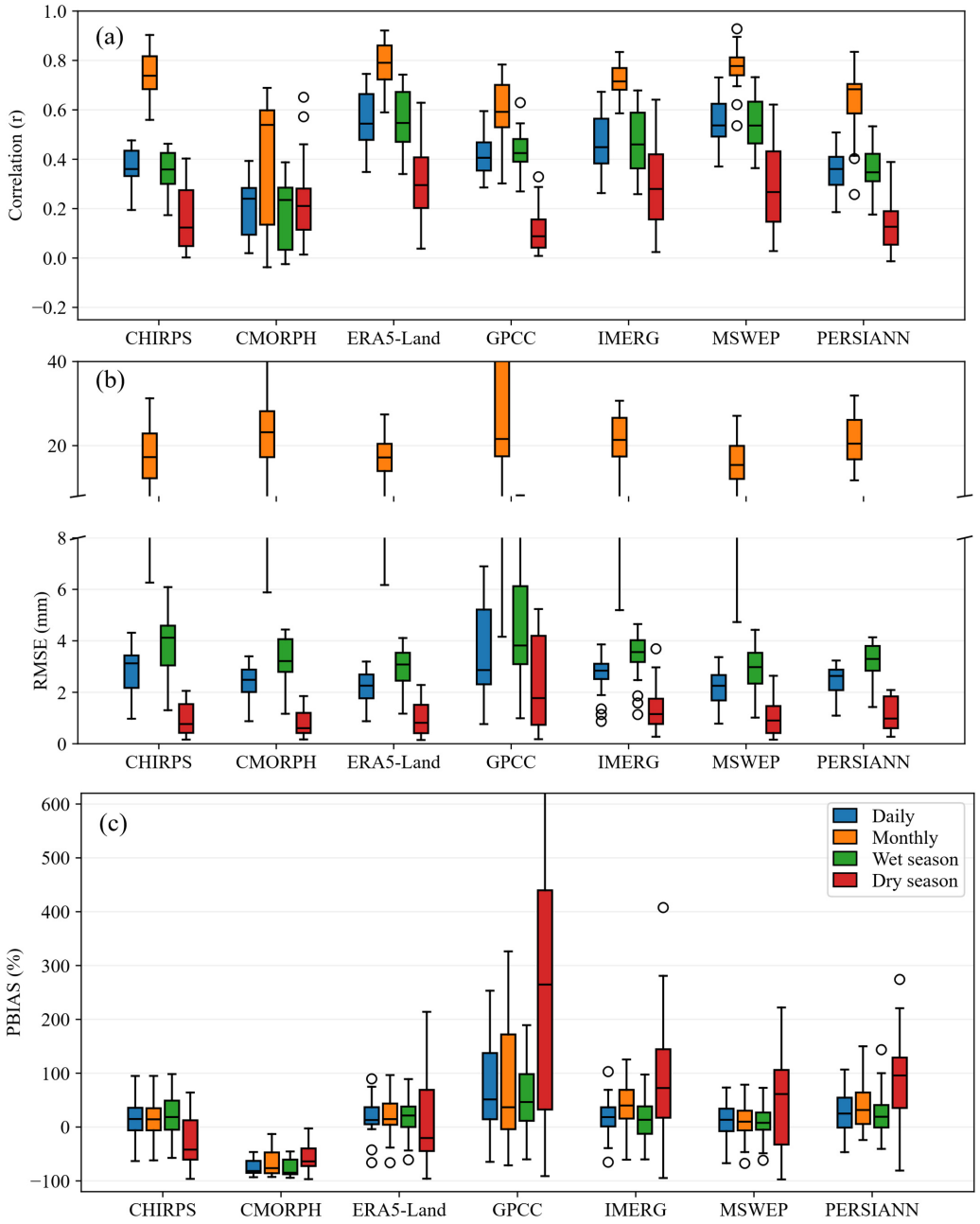


Figure 6. Station-level boxplot distributions of (**a**) correlation, (**b**) RMSE, and (**c**) PBIAS across daily, monthly, wet-season, and dry-season evaluations.

A different pattern emerges for absolute and normalized errors. Although monthly RMSE is clearly larger across all products because monthly totals accumulate precipitation over longer periods, raw RMSE values were interpreted only within the same temporal scale. Therefore, larger monthly RMSE was not taken as evidence of poorer performance. Figure 7 presents the corresponding NRMSE distributions for scale-normalized comparison. Monthly NRMSE is lower than daily NRMSE for all products, confirming that the larger monthly RMSE does not necessarily indicate greater relative error. The very large dry-season NRMSE values primarily reflect normalization by very small mean observed precipitation and should therefore be interpreted as normalized error indicators rather than absolute error magnitudes. MSWEP and ERA5-Land maintain comparatively lower and more compact error distributions, whereas the GPCC products show the broadest and highest error distributions, particularly at the monthly and wet-season scales.

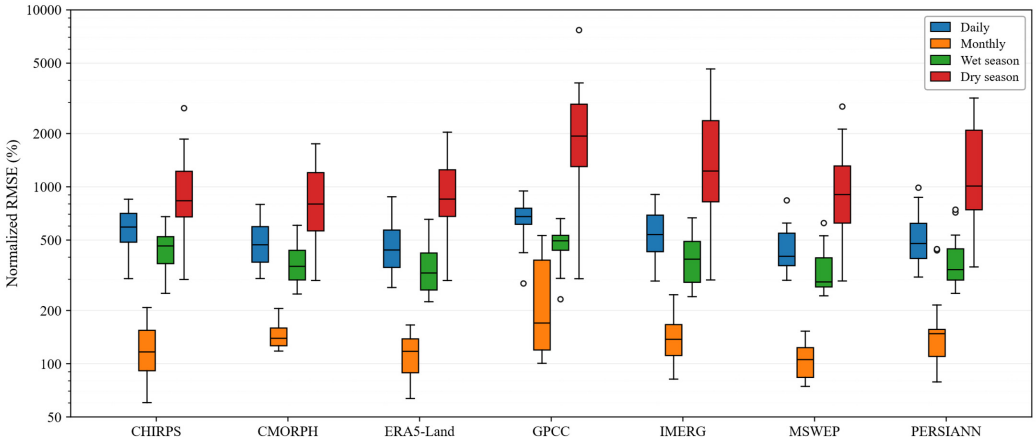


Figure 7. Station-level NRMSE distributions across products and temporal scales (2008–2019).

The bias distributions provide the sharpest product separation. CMORPH is consistently characterized by negative PBIAS across all timescales, confirming systematic underestimation, whereas GPCC datasets exhibit high positive bias and the widest dispersion, particularly in the dry season, indicating substantial overestimation and weak robustness. IMERG and PERSIANN also tend toward positive bias, especially under monthly and dry season conditions, while MSWEP and ERA5-Land maintain comparatively more moderate bias behavior. Overall, these distributions indicate that the more reliable products are not those with high correlation alone, but those that preserve a more balanced relationship between temporal agreement, absolute and normalized error, and systematic bias. This central finding highlights that product performance over the HRB is both metric-dependent and scale-dependent. These distributional contrasts are further examined in the Taylor diagram analysis presented in the next subsection.

4.4. Multi-Metric Performance Patterns

The joint behavior of correlation, normalized variability, and centered root-mean-square difference provides a stricter measure of product realism than any single metric alone. In this multi-metric space, better performance is indicated by closer proximity to the reference state, where correlation is higher, normalized standard deviation is closer to unity, and centered error is smaller. As shown in Figure 8, the clearest product separation occurs across timescale and season.

Accordingly, products located closer to the reference point exhibit a more balanced combination of temporal agreement, normalized variability, and centered error. The annual and wet season monthly panels exhibit tighter clustering around the reference state than the corresponding daily panels. This pattern should be interpreted within the context of the

evaluation framework, where temporal resolution, spatial extraction method, and, for some datasets, the native product differs between the daily and monthly evaluations. In these panels, ERA5-Land, MSWEP, CHIRPS, and, to a lesser extent, IMERG occupy comparatively favorable positions, whereas GPCC datasets remain clearly displaced because of inflated variability and larger centered error. CMORPH is also consistently separated from the main cluster, reflecting suppressed variability and weak reproduction of the observed signal. By contrast, the highest degradation appears under dry-season conditions, especially in the daily domain, where the product spread increases and the distance from the reference state become larger. A similar, though less pronounced, deterioration is evident in the monthly dry-season panel. Overall, the Taylor-space comparison indicates that product performance over the basin is highly scale-dependent and favors datasets that maintain a better balance between temporal agreement, variability fidelity, and centered error.

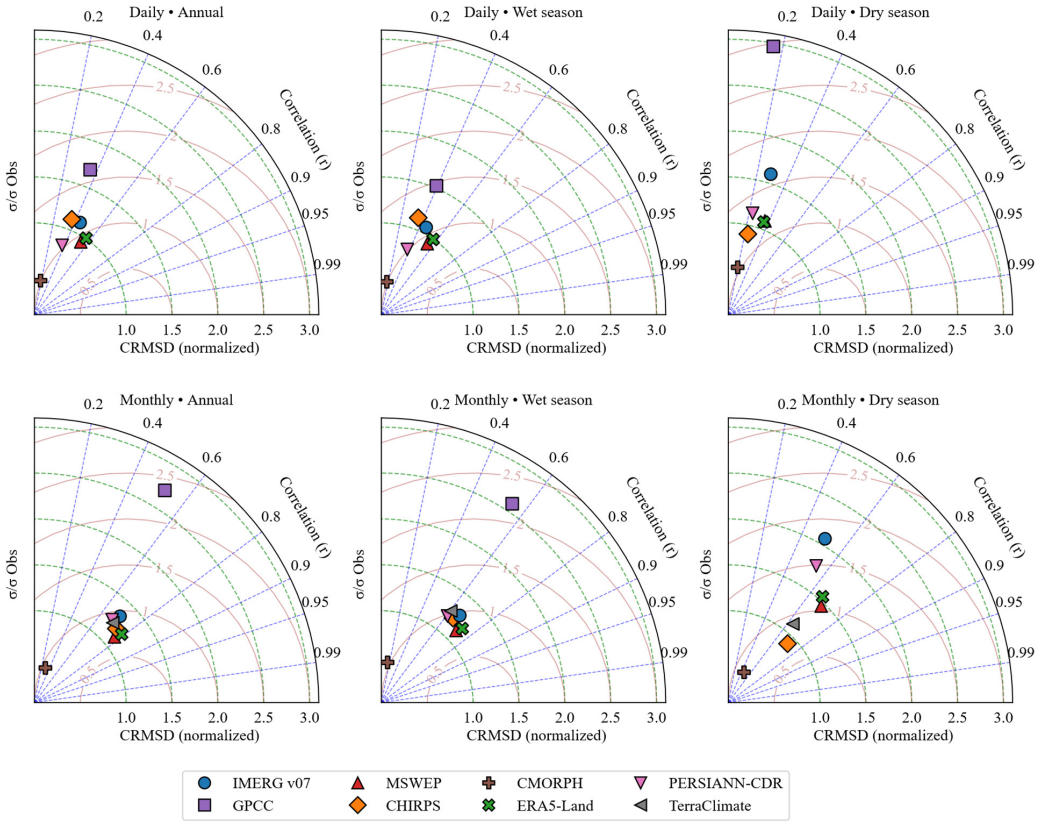


Figure 8. Taylor diagrams showing the relative performance of gridded precipitation products across temporal scales and seasons.

4.5. Integrated Performance Patterns

The integrated heatmap sharpens the main message of the preceding analyses by showing not only which products perform well, but also whether that skill is spatially stable across stations and elevation classes (Figure 9). Because the integrated scores are derived by normalizing the performance metrics separately within each station–temporal-scale–season group, their absolute magnitudes are intended only for within-group comparison and should not be interpreted as being directly compared across different temporal scales or seasons. Cross-scale differences in precipitation-product performance are therefore interpreted from the directly comparable continuous metrics (Section 4.3) and Taylor-diagram analysis (Section 4.4), whereas the integrated heatmap identifies the products that exhibit consistently strong relative performance within each evaluation condition.

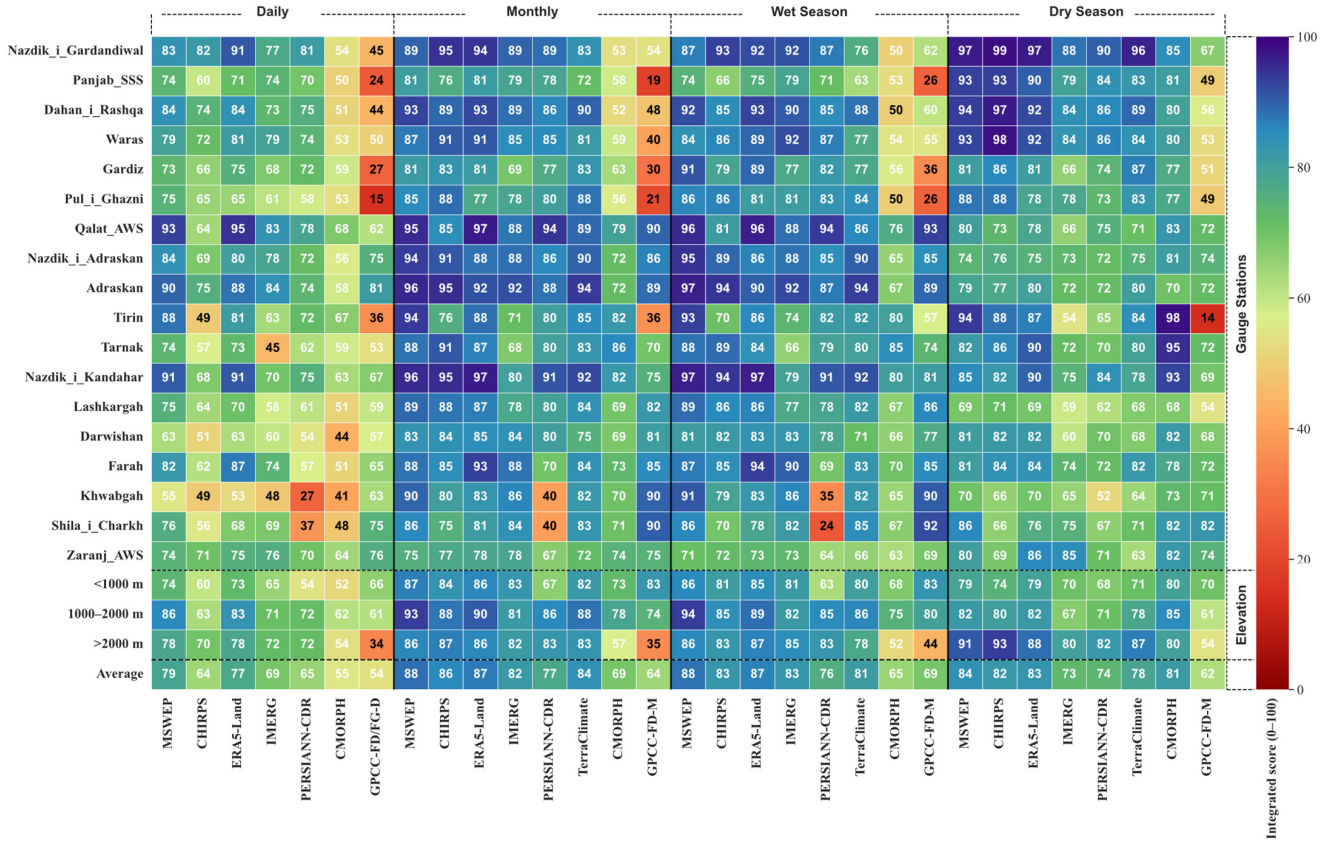


Figure 9. Integrated performance scores of gridded precipitation products across daily, monthly, wet-season, and dry-season scales, summarized by gauge stations and elevation classes. Scores are relative within each evaluation group and are not directly comparable across temporal scales or seasons.

In contrast, the dry season panel is more selective: although MSWEP and ERA5-Land still retain the joint highest average score 84 and 83, CHIRPS and CMORPH remain comparatively high 82 and 81, whereas IMERG declines to 73 and GPCC-FD-M drops sharply to 62; TerraClimate and PERSIANN-CDR are 78 and 74, respectively, indicating that weak rainfall conditions expose differences that are less evident at wetter or aggregated scales. The figure also reveals an important spatial pattern: stations such as Nazdik-i-Kandahar, Farah, Adraskan, Qalat_AWS, and Lashkargah maintain consistently high scores across several products, whereas Gardiz, Pul-i-Ghazni, and some northeastern upland stations show recurrent score decline, indicating an association between station setting and product performance. This elevation-related pattern becomes clearer in the class-based rows, where the >2000 m elevation class shows the weakest integrated performance for several products, with GPCC scores of 36, 36, 45, and 55 and CMORPH scores of 54, 53, 52, and 79 across the daily, monthly, wet-season, and dry-season evaluations, respectively, while low- and mid-elevation bands generally sustain higher scores. An additional insight from the heatmap is that high overall performance must be interpreted as balanced performance rather than isolated strength in any single metric; this explains why MSWEP and ERA5-Land appear most robust across nearly all panels, while CMORPH shows reasonable dry-season scores despite weaker overall performance, and GPCC fails to maintain consistently stable performance when the evaluation is extended across space, season, and elevation. Overall, the heatmap consolidates the basin-scale evidence that precipitation-product performance over the HRB is both scale-dependent and spatially variable across the evaluated station and elevation classes, with MSWEP and ERA5-Land emerging as the most consistent products,

CHIRPS as a strong alternative, and GPCC as the least reliable under integrated evaluation. To evaluate the robustness of the integrated rankings under extremely dry conditions, a sensitivity analysis was performed by recalculating the integrated scores without the normalized RMSE component while retaining the same scoring framework. The exclusion of NRMSE resulted in only minor shifts in the integrated product rankings. At the daily scale, CHIRPS and GPCC-FD/FG-D each moved up one position (fifth to fourth and seventh to sixth, respectively), while PERSIANN-CDR and CMORPH each moved down one position (fourth to fifth and sixth to seventh, respectively). At the monthly scale, GPCC-FD-M moved from eighth to seventh, while CMORPH moved from seventh to eighth. The wet- and dry-season rankings remained unchanged, with MSWEP and ERA5-Land consistently occupying the first and second positions across all four evaluation conditions. These findings indicate that the principal integrated product rankings remain largely stable when NRMSE is excluded, despite the exceptionally high NRMSE values associated with low observed precipitation during the dry season.

4.6. Rainfall Occurrence Detection

The categorical metrics reveal a clear separation in daily rainfall occurrence performance across the evaluated products in Figure 10. At the lowest threshold (0.1 mm day^{-1}), MSWEP and ERA5-Land exhibit the highest relative categorical performance among the evaluated products. MSWEP combines high event detection (median POD = 0.695) with relatively lower false-alarm behavior (FAR = 0.618), while ERA5-Land attains the highest overall categorical skill, with the highest median CSI (0.320) and HSS (0.391). In contrast, CMORPH shows weak occurrence detection from the outset (POD = 0.347) and correspondingly low CSI (0.208) and HSS (0.207). Although PERSIANN and GPCC maintain comparatively high detection rates (POD = 0.689 and 0.623, respectively), their skill is constrained by persistently high false-alarm ratios, particularly for PERSIANN (FAR = 0.738) and GPCC (0.660). This performance structure remains largely consistent as the threshold increases to 0.5 and 1.0 mm day^{-1}, indicating that the relative ranking is reasonably robust rather than highly threshold-dependent.

At 0.5 mm day^{-1}, MSWEP retains the highest median POD (0.672) and HSS (0.464), whereas ERA5-Land remains similarly high with POD = 0.654, CSI = 0.359, and HSS = 0.452. A comparable pattern is preserved at 1.0 mm day^{-1}, where MSWEP continues to show the highest POD (0.651) and HSS (0.488), while ERA5-Land yields the highest CSI (0.363) with nearly identical HSS (0.478). These results indicate that the relatively better-performing products are not simply those with higher detection rates, but those that sustain a more favorable balance between capturing observed rain days and limiting false alarms under increasingly strict event definitions. A second important feature is the different threshold response among products. CHIRPS shows a modest rise in POD from 0.350 to 0.410, but this improvement is offset by an increase in FAR from 0.577 to 0.651, which restricts gains in overall occurrence skill. GPCC follows a similar pattern, with relatively stable POD but consistently high false-alarm ratios. IMERG remains intermediate across all thresholds, with POD declining from 0.569 to 0.488 and HSS increasing only slightly from 0.359 to 0.400, indicating moderate but not leading performance. PERSIANN maintains comparatively high POD, yet its consistently elevated FAR limits both CSI and HSS. The weakest response is observed for CMORPH, whose POD declines sharply from 0.347 at 0.1 mm day^{-1} to 0.115 at 1.0 mm day^{-1}, while CSI and HSS decrease to 0.098 and 0.152, respectively. This behavior indicates limited capability to identify observed rainfall days reliably, particularly when the threshold emphasizes clearer daily events rather than trace precipitation. Overall, these categorical metrics provide an important complement to the continuous statistics. A product may appear acceptable in rainfall magnitude agreement yet still perform weakly in

daily event identification. In this respect, MSWEP and ERA5-Land demonstrate the highest relative rainfall-event detection skill among the evaluated products, whereas CMORPH shows clear limitations and CHIRPS, GPCC, IMERG, and PERSIANN occupy intermediate positions with different trade-offs between detection and false-alarm behavior.

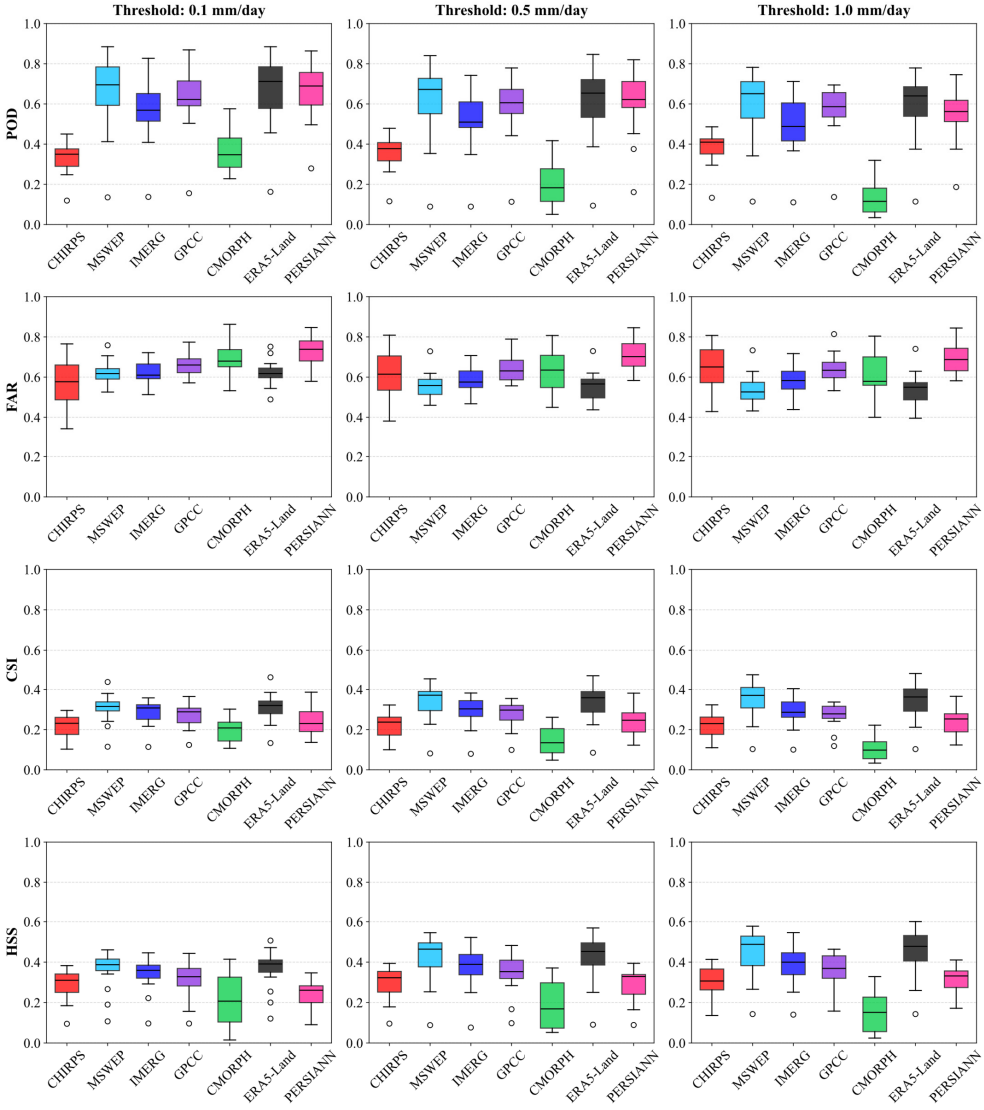


Figure 10. Boxplot distributions of POD, FAR, CSI, and HSS for gridded precipitation products across three daily rainfall thresholds (0.1, 0.5, and 1.0 mm day^{-1}).

4.7. Overall Product Ranking

Product rankings in the HRB vary across temporal scales, seasons, and station/elevation groups, reflecting multiple evaluation dimensions rather than any single metric as summarized in Table 4 and Figure 9. Within each evaluation group, MSWEP and ERA5-Land consistently occupy the leading positions, while CHIRPS performs particularly well in the monthly and wet-season evaluations. Greater product separation occurs under dry-season conditions, where IMERG and GPCC-FD-M show comparatively weaker relative performance. MSWEP and ERA5-Land form the leading performance tier. MSWEP ranks consistently among the strongest products within the daily, monthly, wet-season, and dry-season evaluation groups. MSWEP also demonstrates comparatively strong rainfall-event detection, with POD/HSS values of 0.672/0.464 at 0.5 mm day^{-1} and 0.651/0.488 at 1.0 mm day^{-1}. ERA5-Land likewise maintains strong relative performance within all four evaluation groups, and comparatively high categorical performance, with CSI/HSS

values of 0.359/0.452 at 0.5 mm day^{-1} and 0.363/0.478 at 1.0 mm day^{-1}. These categorical metrics describe rainfall-event detection separately and were not included in the integrated score, which was calculated from *r*, NRMSE, and absolute PBIAS. CHIRPS forms the next tier, showing strong relative performance in the monthly, wet-season, and dry-season evaluations and across several station and elevation groups, but more moderate performance at the daily and rainfall-event scales, mainly because gains in detection are offset by higher false alarms. IMERG and PERSIANN-CDR form an intermediate group, and TerraClimate also shows moderate performance for the monthly and seasonal evaluations, the only temporal scales for which it is available. Although these products preserve useful agreement across some evaluation conditions, their balance across the applicable evaluation metrics is weaker than that of the leading tier. GPCC generally forms the lowest-scoring group, whereas CMORPH indicates more condition-dependent performance. CMORPH is limited by persistent underestimation (daily PBIAS = −81.7%; wet-season PBIAS = −84.8%) and weak rainfall-event detection, whereas GPCC is limited by inflated variability, strong overestimation, very large dry-season bias (PBIAS = 264.7%), and consistently weak relative performance within the corresponding evaluation groups. Therefore, no single product can be assigned a universally superior rank across all evaluation conditions. Overall, the most reliable products are those that maintain balanced performance across metrics, seasons, and station/elevation groups, rather than those that appear strong in only one test. Table 4 summarizes the integrated evidence from Figure 9 together with the complementary evaluation results. The qualitative ratings in Table 4 were derived by synthesizing the integrated score together with the complementary continuous, categorical, spatial consistency, and seasonal evaluation results. These labels summarize the overall body of evidence across all evaluation dimensions and are intended as descriptive interpretations rather than independently calculated metrics derived from predefined numerical thresholds.

Table 4. Comparative ranking of gridded precipitation products in the HRB.

Product	Daily	Rainfall Occurrence	Monthly	Wet-Season	Dry-Season	Spatial Consistency	Main Strength	Main Limitation	Overall Reliability
CHIRPS	Moderate	Moderate	Strong	Strong	Strong	Strong	gauge-adjusted seasonal agreement	dry-event false alarms	Moderate
MSWEP	Strong	Strong	Strong	Strong	Strong	Strong	balanced multi-metric structure	slight dry-season undercapture	High
ERA5-Land	Strong	Strong	Strong	Strong	Strong	Strong	low-error cross-scale consistency	reduced variance in dry conditions	High
IMERG	Moderate	Moderate	Moderate	Moderate	Moderate	Moderate	good temporal variability reproduction	positive-bias amplification	Moderate
CMORPH	Weak	Weak	Weak	Weak	Moderate	Weak	low dry-season RMSE	persistent rainfall underestimation	Low
PERSIANN	Moderate	Moderate	Moderate	Moderate	Moderate	Moderate	stable event detectability	elevated false-alarm tendency	Moderate
GPCC-FD/FG-D	Weak	Moderate	N/A	N/A	N/A	Weak	Gauge-based daily precipitation estimate	inflated variability and overestimation	Low
GPCC-FD-M	N/A	N/A	Weak	Weak	Weak	Weak	Gauge-based Monthly precipitation estimate	inflated variability and overestimation	Low

Table 4. *Cont.*

Product	Daily	Rainfall Occurrence	Monthly	Wet-Season	Dry-Season	Spatial Consistency	Main Strength	Main Limitation	Overall Reliability
TerraClimate	N/A	N/A	Moderate	Moderate	Moderate	Moderate	stable seasonal-spatial representation	no daily event-scale capability	Moderate

Note: The categorical ratings are descriptive syntheses of the evaluation results, not classifications based on predefined numerical thresholds.

4.8. Common-Period and Fixed-Network Sensitivity Analysis

The original evaluation used the valid overlap between each station and each precipitation product. Because these overlaps were not identical, differences in product performance could partly reflect differences in evaluation samples rather than product behavior. We therefore repeated the analysis using a common sample. Here, O denotes the original score or rank, C denotes the common-sample score or rank, and ΔR denotes the change in rank between the two analyses as reported in Table 5. The daily common sample comprised 1699 shared dates from January 2013 to August 2017, whereas the monthly and seasonal analyses used 56 common months within the same period. TerraClimate was not included in the daily comparison because it is available only on the monthly scale. The resulting original and common-sample comparisons are presented in Table 5.

Table 5. Original/common-period (O/C) integrated scores and ranks, with rank change (ΔR), for precipitation products across daily, monthly, wet-season, and dry-season evaluations.

Product	Daily		Monthly		Wet Season		Dry Season	
	Score (O/C)	Rank (O/C; ΔR)	Score (O/C)	Rank (O/C; ΔR)	Score (O/C)	Rank (O/C; ΔR)	Score (O/C)	Rank (O/C; ΔR)
MSWEP	78.52/81.00	1/1 (+0)	88.53/90.29	1/1 (+0)	88.26/89.14	1/1 (+0)	83.68/81.35	1/3 (+2)
CHIRPS	63.71/66.19	5/5 (+0)	85.66/86.71	3/3 (+0)	82.63/83.43	4/4 (+0)	82.52/81.83	3/2 (−1)
ERA5-Land	77.44/80.14	2/2 (+0)	87.33/90.18	2/2 (+0)	86.50/89.04	2/2 (+0)	83.26/82.01	2/1 (−1)
IMERG	68.84/72.61	3/3 (+0)	81.74/85.29	5/4 (−1)	82.72/85.57	3/3 (+0)	72.54/69.50	7/7 (+0)
PERSIANN-CDR	65.22/70.65	4/4 (+0)	77.30/76.27	6/6 (+0)	76.09/72.76	6/6 (+0)	73.87/73.91	6/6 (+0)
CMORPH	55.30/56.43	6/7 (+1)	68.70/70.06	7/7 (+0)	64.82/65.60	8/8 (+0)	80.95/80.13	4/4 (+0)
GPCC	54.37/60.33	7/6 (−1)	63.98/68.71	8/8 (+0)	69.05/72.48	7/7 (+0)	62.05/57.54	8/8 (+0)
TerraClimate	—	—	84.15/83.99	4/5 (+1)	81.48/80.24	5/5 (+0)	78.38/78.02	5/5 (+0)

Most product rankings were retained under the common-period benchmark. In the daily analysis, only CMORPH and GPCC exchanged positions, while the monthly changes were limited to IMERG and TerraClimate. All wet-season rankings remained unchanged. The largest differences occurred in the dry season, where MSWEP changed from first to third, whereas ERA5-Land and CHIRPS improved from second to first and from third to second, respectively. These changes indicate that sample definition can influence individual rankings, but the main performance structure remained broadly stable. Differences in score values without a rank change reflect the recalculation of the integrated metrics within the common sample. The influence of changing gauge availability on the basin reference was evaluated separately using annual station totals. Annual precipitation was first calculated for each station from 12 valid monthly observations and then averaged across stations. The fixed network comprised the largest set of 12 stations with complete annual records throughout 2009–2017. As shown in Figure 11, the annual series based on the fixed network and all stations with complete annual records closely followed one another during the

nine-year period, with a Pearson correlation of 0.988, an RMSE of 3.96 mm, a mean absolute difference of 3.44 mm, and a mean absolute percentage difference of 1.81%.



Figure 11. Gauge-network robustness during 2009–2017: (**a**) annual basin precipitation from all complete-year stations and the fixed 12-station network; (**b**) station availability and percentage differences.

The annual percentage differences ranged from −3.60% to 1.77%, indicating that station-network composition had a limited effect on the annual basin estimate during 2009–2017. Together, these checks demonstrate that the principal product rankings and the annual gauge-based basin estimates are generally robust to the evaluated differences in temporal coverage and station-network composition. However, the fixed-network verification cannot be extended beyond 2019 because the multi-station record becomes incomplete. Accordingly, the available-station annual series after 2019 is interpreted as station-limited evidence rather than as a complete basin-wide climatic record.

5. Limitations of Station Coverage

The gauge network used in this study provides an important reference for evaluating gridded precipitation products in the HRB, but some limitations should be noted. Although the 18 stations cover the basin, their distribution is not fully even. As depicted in Figure 1, many stations are located near the river network and more accessible areas, while some remote desert and upper-basin mountainous areas are less represented. This may reduce the spatial representativeness of the station data in parts of the basin where precipitation is affected by elevation and local terrain. A second limitation relates to record consistency. The main evaluation used recent observed precipitation records from 18 stations during 2008–2025, according to station-specific record availability. A longer historical dataset from 14 stations covering 1979–2009 was used only for supporting seasonal, interannual, and trend consistency analysis. These historical records include gap-filled values provided by the Ministry of Energy and Water (MoEW) in Afghanistan; however, the gap-filling procedure was not documented in the source files. Therefore, the historical period was not used for the main product ranking, and its results were interpreted only as supporting evidence. The independence of the validation data also remains uncertain. CHIRPS, IMERG,

MSWEP, and GPCC use gauge information in their generation or correction [24,27,32,46], but the available documentation does not identify whether the 18 MoEW stations contributed to their input networks. Therefore, independence cannot be verified and is not assumed. Any overlap could reduce the apparent errors of the affected products. Their results are consequently interpreted as agreement with gauge-informed products rather than as fully independent validation. Another limitation is the scale mismatch between gauge observations and grid-cell precipitation estimates. Gauge observations represent rainfall at a single location, whereas gridded products represent average values over a grid cell. For this reason, some differences between station and gridded precipitation are expected, especially in areas with strong spatial variability. In this study, daily values were extracted from the nearest grid cell, and monthly values were derived using bilinear interpolation, but this difference in spatial scale cannot be removed completely. Overall, these limitations do not reduce the value of the study, but they suggest that the results should be interpreted as a station-based basin evaluation, with more caution in sparsely monitored and topographically complex parts of the HRB.

6. Conclusions

This study evaluated the reliability of multi-source gridded precipitation products over the HRB using gauge observations across daily and monthly timescales, wet and dry seasons, rainfall occurrence thresholds, and elevation settings. The results indicate that product performance over the basin is strongly dependent on temporal scale, seasonal condition, and elevation setting, rather than on any single metric alone. Overall, MSWEP and ERA5-Land provided the most consistent and balanced performance across the different evaluations, while CHIRPS emerged as a strong alternative, especially at monthly and seasonal scales. IMERG and PERSIANN-CDR showed moderate skill, whereas CMORPH and GPCC datasets were the least reliable overall because of persistent underestimation in the former and inflated variability and strong overestimation in the latter. In general, monthly and wet-season evaluations showed higher agreement than daily and dry-season evaluations; however, this pattern reflects the combined influence of temporal scale, spatial extraction method, and native product characteristics, rather than temporal aggregation alone. Rainfall-event analysis further confirmed the high relative detection skill of MSWEP and ERA5-Land. The results also indicate that product behavior is not spatially uniform across the HRB, with weaker performance generally associated with high-elevation stations. Taken as a whole, these findings indicate a station-based basis for more informed precipitation dataset selection in the HRB, while highlighting uncertainty in sparsely monitored parts of the basin. They also indicate that balanced performance across error, bias, temporal agreement, event detection, and elevation settings are more important than strength in any single evaluation metric.

Author Contributions: Conceptualization, S.M.Q., A.M.A.-A. H.O.S. and S.C.; methodology, S.M.Q. and A.M.A.-A.; software, S.M.Q.; validation, S.M.Q.; formal analysis, S.M.Q., A.M.A.-A. H.O.S. and S.C.; investigation, S.M.Q., and A.M.A.-A.; resources, S.M.Q.; data curation, S.M.Q.; writing—original draft preparation, S.M.Q.; writing—review and editing, A.M.A.-A. H.O.S. and S.C.; visualization, S.M.Q.; supervision, A.M.A.-A.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The precipitation observations used in this study were obtained from the Hydro-Meteorological Database, Surface Water Resources Department, General Directorate of Water Resources, Ministry of Water and Energy, Afghanistan. These third-party governmental data are subject to redistribution restrictions imposed by the data provider and therefore cannot be publicly shared by the authors. The original data and supporting correspondence can be provided to the journal for verification/checking purposes only, upon request.

Acknowledgments: The authors would like to acknowledge all support provided by the Department of Civil and Environmental Engineering and the Interdisciplinary Research Center for Membranes and Water Security at King Fahd University of Petroleum & Minerals in Saudi Arabia for their support. Article Processing Charge (APC) was provided by the Deanship of Research at KFUPM.

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

References

1. Zhao, C.; Xu, M.; Wang, Z.; Li, J.; Zheng, J.; Yuan, M.; Tao, Y.; Shi, L. Spatiotemporal Distribution and Applicability Evaluation of Remote Sensing Precipitation in River Basins Across Mainland China. *Remote Sens.* **2025**, *17*, 3534. [CrossRef]
2. Henn, B.; Newman, A.J.; Livneh, B.; Daly, C.; Lundquist, J.D. An assessment of differences in gridded precipitation datasets in complex terrain. *J. Hydrol.* **2018**, *556*, 1205–1219. [CrossRef]
3. Awange, J.L.; Hu, K.X.; Khaki, M. The newly merged satellite remotely sensed, gauge and reanalysis-based Multi-Source Weighted-Ensemble Precipitation: Evaluation over Australia and Africa (1981–2016). *Sci. Total Environ.* **2019**, *670*, 448–465. [CrossRef] [PubMed]
4. Wan, Y.; Li, D.; Sun, J.; Wang, M.; Liu, H. Evaluation of six latest precipitation datasets for extreme precipitation estimates and hydrological application across various climate regions in China. *Atmos. Res.* **2025**, *315*, 107932. [CrossRef]
5. Yıldız, M.B.; Nunno, F.; Di Marinis, G.; De Granata, F. Journal of Hydrology: Regional Studies Evaluating global precipitation datasets over Sicily: From daily estimates to extreme events. *J. Hydrol. Reg. Stud.* **2026**, *63*, 103062. [CrossRef]
6. Tarkegn, T.G.; Ray, R.L.; Tefera, G.W. Comprehensive evaluations of gridded precipitation datasets across diverse climate zones of Brazos River Basin, Texas, USA. *Remote Sens. Appl. Soc. Environ.* **2026**, *41*, 101947. [CrossRef]
7. Kouakou, C.; Paturel, J.-E.; Satgé, F.; Tramblay, Y.; Defrance, D.; Rouché, N. Comparison of gridded precipitation estimates for regional hydrological modeling in West and Central Africa. *J. Hydrol. Reg. Stud.* **2023**, *47*, 101409. [CrossRef]
8. Dollan, I.J.; Maina, F.Z.; Kumar, S.V.; Nikolopoulos, E.I.; Maggioni, V. An assessment of gridded precipitation products over High Mountain Asia. *J. Hydrol. Reg. Stud.* **2024**, *52*, 101675. [CrossRef]
9. de la Fraga, P.; Del-Toro-Guerrero, F.J.; Vivoni, E.R.; Cavazos, T.; Kretzschmar, T. Evaluation of gridded precipitation datasets in mountainous terrains of Northwestern Mexico. *J. Hydrol. Reg. Stud.* **2024**, *56*, 102019. [CrossRef]
10. Cheng, Y.; Zhang, X.; Wang, K.; Zhang, Y.; Guo, Y.; Shen, Y.-J. Multidimensional evaluation of satellite-based and reanalysis-based precipitation datasets in the Tibetan Plateau. *J. Hydrol.* **2025**, *660*, 133364. [CrossRef]
11. Zhang, H.; Wei, L.; Zhu, Y.; Zhou, J.; Liu, S.; Sun, X.; Kang, X.; Gao, M.; Duan, Z.; Crow, W.T. Comparison of multi-source merged precipitation products using independent gauge observations. *Atmos. Res.* **2025**, *328*, 108427. [CrossRef]
12. Gebrechorkos, S.H.; Leyland, J.; Dadson, S.J.; Cohen, S.; Slater, L.; Wortmann, M.; Ashworth, P.J.; Bennett, G.L.; Boothroyd, R.; Cloke, H. Global scale evaluation of precipitation datasets for hydrological modelling. *Hydrol. Earth Syst. Sci. Discuss.* **2023**, *2023*, 1–33. [CrossRef]
13. Kao, Y.-C.; Tsou, H.-E.; Chen, C.-J. Development of multi-source weighted-ensemble precipitation: Influence of bias correction based on recurrent convolutional neural networks. *J. Hydrol.* **2024**, *629*, 130621. [CrossRef]
14. Samim, A.T.; Nayyer, F.; Hussainzada, W.; Lee, H.S. Intercomparison of gridded global precipitation data for arid and mountainous regions: A case study of Afghanistan. *J. Hydrol. Reg. Stud.* **2024**, *53*, 101837. [CrossRef]
15. Javadi, M.; Jalilehvand, M.; Alizadeh, H.; Zafarmomen, N. Analysis of historical global warming impacts on climatological trends for the partially gauged Hirmand river basin based on multiple data products and bias correction methods. *J. Hydrol. Reg. Stud.* **2025**, *62*, 102886. [CrossRef]
16. Karim, F.; Penton, D.J.; Aryal, S.K.; Wahid, S.; Chen, Y.; Taylor, P.; Cuddy, S.M. Large scale water yield assessment for sparsely monitored river basins: A case study for Afghanistan. *PLoS Water* **2024**, *3*, e0000165. [CrossRef]
17. Mushwani, H.; Hairan, M.H.; Sahak, K.; Arabzai, A.; Safi, L.; Ahmadzai, M.R.; Shirzai, H.; Hashmi, S.K.; Peroz, S. Meteorological Drought Risk Assessment Using Spi Numerical Model: A Case Study of Helmand River Basin, Afghanistan. *Plan. Malays.* **2024**, *22*. [CrossRef]
18. Maleki, S.; Mohajeri, S.H.; Samadi, A.; Hasani, H.; Akbary, M. Remote sensing insights into water allocation and evaporation challenges in the Hirmand River Basin, after the operation of Kamal Khan Dam. *J. Hydrol. Reg. Stud.* **2024**, *56*, 101996. [CrossRef]
19. Galehban, E.; Hamzeh, S.; Veysi, S.; AlaviPanah, S.K.; Yazdanpanah Dero, Q.; Karimi, P. Capability of remote sensing and reanalysis data for water conflict assessment in Helmand River Basin. *Appl. Water Sci.* **2025**, *15*, 294. [CrossRef]
20. Hazarika, A.; Thander, B.; Lahiri, S.K. Lithium resource mapping in Afghanistan's Helmand Basin: A multi-criteria approach of zonation using the Analytic Hierarchy Process. *Appl. Geochem.* **2025**, *190*, 106460. [CrossRef]

21. Goes, B.J.M.; Howarth, S.E.; Wardlaw, R.B.; Hancock, I.R.; Parajuli, U.N. Integrated water resources management in an insecure river basin: A case study of Helmand River Basin, Afghanistan. *Int. J. Water Resour. Dev.* **2016**, *32*, 3–25. [CrossRef]
22. Funk, C.; Peterson, P.; Landsfeld, M.; Pedreros, D.; Verdin, J.; Shukla, S.; Husak, G.; Rowland, J.; Harrison, L.; Hoell, A. The climate hazards infrared precipitation with stations—A new environmental record for monitoring extremes. *Sci. Data* **2015**, *2*, 150066. [CrossRef] [PubMed]
23. Climate Hazards Center. *Climate Hazards Center Infrared Precipitation with Stations Version 3 (CHIRPS v3)*; CHIRPS3 Data Repository; Climate Hazards Center, University of California, Santa Barbara: Santa Barbara, CA, USA, 2025. [CrossRef]
24. Beck, H.E.; Wood, E.F.; Pan, M.; Fisher, C.K.; Miralles, D.G.; Van Dijk, A.I.J.M.; 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]
25. Muñoz Sabater, J. *ERA5-Land Monthly Averaged Data from 1950 to Present*; Copernicus Climate Change Service (C3S) Climate Data Store (CDS); European Centre for Medium-Range Weather Forecasts (ECMWF): Reading, UK, 2019. [CrossRef]
26. Muñoz Sabater, J. *ERA5-Land Hourly Data from 1950 to Present*; Copernicus Climate Change Service (C3S) Climate Data Store (CDS); European Centre for Medium-Range Weather Forecasts (ECMWF): Reading, UK, 2019. [CrossRef]
27. Huffman, G.J.; Stocker, E.F.; Bolvin, D.T.; Nelkin, E.J.; Jackson, T. *GPM IMERG Final Precipitation L3 1 Month 0.1 Degree x 0.1 Degree V07 (GPM_3IMERGM)*; Goddard Earth Sciences Data and Information Services Center (GES DISC): Greenbelt, MD, USA, 2023. [CrossRef]
28. Huffman, G.J.; Stocker, E.F.; Bolvin, D.T.; Nelkin, E.J.; Jackson, T. *GPM IMERG Final Precipitation L3 1 Day 0.1 Degree x 0.1 Degree V07*; Goddard Earth Sciences Data and Information Services Center (GES DISC): Greenbelt, MD, USA, 2023. [CrossRef]
29. Xie, P.; Joyce, R.; Wu, S.; Yoo, S.-H.; Yarosh, Y.; Sun, F.; Lin, R. Reprocessed, Bias-Corrected CMORPH Global High-Resolution Precipitation Estimates from 1998. *J. Hydrometeorol.* **2017**, *18*, 1617–1641. [CrossRef]
30. Xie, P.; Joyce, R.; Wu, S.; Yoo, S.-H.; Yarosh, Y.; Sun, F.; Lin, R. NOAA Climate Data Record (CDR) of CPC Morphing Technique (CMORPH) High Resolution Global Precipitation Estimates, Version 1. In *NOAA CDR Program*; National Centers for Environmental Information (NCEI), NOAA: Asheville, NC, USA, 2019. [CrossRef]
31. Ashouri, H.; Hsu, K.L.; Sorooshian, S.; Braithwaite, D.K.; Knapp, K.R.; Cecil, L.D.; Nelson, B.R.; Prat, O.P. PERSIANN-CDR: Daily precipitation climate data record from multisatellite observations for hydrological and climate studies. *Bull. Am. Meteorol. Soc.* **2015**, *96*, 69–83. [CrossRef]
32. Ziese, M.; Rauthe-Schöch, A.; Becker, A.; Finger, P.; Rustemeier, E.; Hänsel, S.; Schneider, U. *GPCC Full Data Daily Version 2022 at 1.0°: Daily Land-Surface Precipitation from Rain-Gauges Built on GTS-Based and Historic Data*; Global Precipitation Climatology Centre (GPCC.) at Deutscher Wetterdienst: Frankfurt am Main, Germany, 2022. [CrossRef]
33. Abatzoglou, J.T.; Dobrowski, S.Z.; Parks, S.A.; Hegewisch, K.C. TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015. *Sci. Data* **2018**, *5*, 170191. [CrossRef] [PubMed]
34. Schamm, K.; Ziese, M.; Becker, A.; Finger, P.; Meyer-Christoffer, A.; Rudolf, B.; Schneider, U. *GPCC First Guess Daily Product at 1.0°: Near Real-Time First Guess Daily Land-Surface Precipitation from Rain-Gauges Based on SYNOP Data*; Global Precipitation Climatology Centre (GPCC.) at Deutscher Wetterdienst: Frankfurt am Main, Germany, 2013. [CrossRef]
35. Schneider, U.; Hänsel, S.; Finger, P.; Rustemeier, E.; Ziese, M. *GPCC Full Data Monthly Product Version 2022 at 0.25°: Monthly Land-Surface Precipitation from Rain-Gauges Built on GTS-Based and Historical Data*; Global Precipitation Climatology Centre at Deutscher Wetterdienst: Frankfurt am Main, Germany, 2022. [CrossRef]
36. Accadia, C.; Mariani, S.; Casaioli, M.; Lavagnini, A.; Speranza, A. Sensitivity of precipitation forecast skill scores to bilinear interpolation and a simple nearest-neighbor average method on high-resolution verification grids. *Weather Forecast.* **2003**, *18*, 918–932. [CrossRef]
37. Xin, Y.; Yang, Y.; Chen, X.; Yue, X.; Liu, Y.; Yin, C. Evaluation of IMERG and ERA5 precipitation products over the Mongolian Plateau. *Sci. Rep.* **2022**, *12*, 21776. [CrossRef] [PubMed]
38. Deng, X.; Nie, S.; Deng, W.; Cao, W. Statistical evaluation of the performance of gridded monthly precipitation products from reanalysis data, satellite estimates, and merged analyses over China. *Theor. Appl. Climatol.* **2018**, *132*, 621–637.
39. Dandapath, S.; Patil, A.; Shinde, D.S.; Pathak, P.K. Identification and distribution of wet and dry season in the 70 most populated cities in India. *J. Atmos. Sol.-Terr. Phys.* **2025**, *277*, 106649. [CrossRef]
40. Guo, R.; Liu, Y. Evaluation of satellite precipitation products with rain gauge data at different scales: Implications for hydrological applications. *Water* **2016**, *8*, 281. [CrossRef]
41. Cao, B.; Yu, Q.; Bai, Y.; Lu, Z.; Wang, S.; Wang, M.; Zhang, H.; Hu, C. Improving precipitation estimation and hydrological simulation in Tianshan Mountain basins via CNN-SE-EF fusion. *J. Hydrol. Reg. Stud.* **2026**, *64*, 103179. [CrossRef]
42. Mann, H.B. Nonparametric tests against trend. *Econom. J. Econom. Soc.* **1945**, *13*, 245–259. [CrossRef]
43. Frimpong, B.F.; Koranteng, A.; Molkenthin, F. Analysis of temperature variability utilising Mann–Kendall and Sen's slope estimator tests in the Accra and Kumasi Metropolises in Ghana. *Environ. Syst. Res.* **2022**, *11*, 24. [CrossRef]
44. Sen, P.K. Estimates of the regression coefficient based on Kendall's tau. *J. Am. Stat. Assoc.* **1968**, *63*, 1379–1389. [CrossRef]

45. Hirsch, R.M.; Slack, J.R.; Smith, R.A. Techniques of trend analysis for monthly water quality data. *Water Resour. Res.* **1982**, *18*, 107–121. [CrossRef]
46. Funk, C.; Peterson, P.; Harrison, L.; Saldivar, R.; Landsfeld, M.; Pedreros, D.; Shukla, S.; Fink, A.H.; Davenport, F.; Peterson, S. The climate hazards center infrared precipitation with stations, version 3. *Sci. Data* **2026**, *13*, 718. [CrossRef] [PubMed]

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.