remote sensing

MDPI

Article

Event-Based and Error-Component Evaluation of GPM IMERG and GSMaP_Gauge for Monitoring Hourly Extreme Precipitation over North China

Lirong Tan [1], Dong Xie [1,2], Li Zhu [1], Weiwei Cui [1,2], Yunqian Wang [1,2], Hao Guo [1,2,*] and Philippe De Maeyer [3,4]

[1] School of Geography and Tourism, Qufu Normal University, Rizhao 276826, China; tanlr2305@qfnu.edu.cn (L.T.)
[2] Sino-Belgian Joint Laboratory for Geo-Information, Rizhao 276826, China
[3] Department of Geography, Ghent University, 9000 Gent, Belgium
[4] Sino-Belgian Joint Laboratory for Geo-Information, 9000 Gent, Belgium
* Correspondence: guohao@qfnu.edu.cn

Highlights

What are the main findings?

- IMERG showed a persistent dry bias, whereas GSMaP's smaller net bias masked large compensating hit, missed, and false components.
- GSMaP detected wet hours more frequently but substantially overestimated event duration; both products underestimated mean event intensity and hourly precipitation peaks.

What are the implications of the main findings?

- Neither net precipitation bias nor detection probability alone provides a sufficient basis for product selection because compensating errors can produce apparently favorable aggregate performance.
- Bias correction for hydrological and hazard applications should jointly address precipitation occurrence and event duration, accumulation, and intensity.

Abstract

Hourly satellite precipitation products are often evaluated using aggregate statistics that can conceal compensating errors in precipitation occurrence and event duration, accumulation, and intensity. Using all matched observations from 522 meteorological stations in North China between 1 October 2019 and 26 September 2022, this study evaluated GPM IMERG Final Run V07 and GSMaP_Gauge V8. The framework combined precipitation-amount and occurrence metrics, event characteristics defined using local 90th percentiles of wet-hour precipitation, and additive and multiplicative error decomposition. IMERG showed a persistent dry bias, with a mean annualized total error of −348.2 mm yr^{-1}. GSMaP had a higher probability of detection than IMERG (77.6% versus 38.6%) but also a higher false alarm ratio (67.5% versus 29.7%); its mean total error of +99.7 mm yr^{-1} masked substantial compensation among the error components. For extreme precipitation events, IMERG underestimated annualized event frequency by 33.9% but reproduced mean duration comparatively closely (−11.0%). GSMaP overestimated event frequency by 31.6% and duration by 198.6%. Its near-unbiased mean event total (+1.1%) resulted from prolonged duration despite underestimation of mean and peak intensity by 67.3% and 67.4%, respectively. Within the approximately three-year evaluation period examined here, these results show

Check for updates

Academic Editors: Yi Chang and Zhaoxin Cai

Received: 17 August 2026
Revised: 14 October 2026
Accepted: 21 October 2026
Published: 24 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.

that product selection should be application-specific and that event-scale correction should address precipitation occurrence, duration, accumulation, and intensity jointly.

Keywords: hourly precipitation; extreme precipitation event; GPM IMERG; GSMaP_Gauge; error-component decomposition; event-based evaluation; North China

1. Introduction

Extreme precipitation is a high-impact hydroclimatic phenomenon that can trigger flooding, urban waterlogging, landslides, and cascading infrastructure failures [1,2]. North China is particularly vulnerable because intense summer rainfall affects densely populated plains, mountain-front cities, and major transport and water-management systems. Long-term observations indicate substantial changes in extreme precipitation across North China. Recent analyses of hourly station records from 1983 to 2022 showed a significant increase in warm-season extreme hourly precipitation amount, primarily driven by an increase in event frequency, with particularly pronounced changes in some plain and mountainous areas [3]. A complementary analysis of summer extreme rainfall during 1979–2021 further identified widespread increases in extreme-rainfall frequency and amount across North China, together with an interdecadal transition toward a wetter extreme-rainfall regime around the early 2000s [4]. These changes, together with the severe impacts of recent events, underscore the need for reliable high-temporal-resolution precipitation monitoring. The 2021 Henan "7.20" rainstorm and the July 2023 regional event over North China illustrate the severe impacts of short-duration rainfall extremes [2,5–7].

Rain gauges provide direct point-scale measurements but cannot fully represent spatial rainfall variability where station density is uneven [8]. Weather radar offers high-resolution spatial information, yet estimates are affected by beam blockage, attenuation, ground clutter, range degradation, and reflectivity–rainfall conversion uncertainty [9]. Satellite precipitation products complement these observations through spatially continuous coverage, although retrieval algorithms, sampling frequency, and the mismatch between grid-cell and point measurements introduce additional uncertainty [8,10].

GPM IMERG and GSMaP are among the most widely used global satellite precipitation products [10,11]. IMERG integrates passive-microwave, infrared, and gauge information at 0.1° and 30-min resolution [12]. GSMaP provides hourly precipitation at 0.1°, and its gauge-adjusted product is widely used for retrospective hydrometeorological analysis [11,13]. Because the products use different retrieval and adjustment strategies, similar total precipitation can arise from markedly different combinations of detected, missed, and falsely detected precipitation.

Previous evaluations have examined IMERG and GSMaP across regions, seasons, terrain classes, and precipitation intensities [14–18]. Studies based on daily extreme indices, including R95p, Rx1day, CDD, CWD, R10mm, and SDII, have also evaluated the products' ability to represent extreme precipitation [19–22]. These aggregated indices, however, do not resolve event onset and termination, weak rainfall tails, peak attenuation, or compensation between missed and false precipitation. Event-based analysis and error-component decomposition can therefore reveal process errors that remain hidden when aggregate measures such as correlation, relative bias, and accumulated precipitation are considered individually [23–25].

North China provides a stringent test because monsoon rainfall interacts with plains, mountain ranges, and strong south–north precipitation gradients [26–28]. Existing regional assessments remain dominated by accumulated precipitation or conventional extreme

indices [25,28,29]. Few studies have combined hourly event characteristics with additive and multiplicative error decomposition to compare the latest gauge-adjusted IMERG and GSMaP products. Event-based error decomposition has previously been used to diagnose the sources of IMERG errors across mainland China [25]. Building on this established diagnostic basis, the present study extends the analysis from a single-product diagnosis to a direct comparison of IMERG Final Run V07 and GSMaP_Gauge V8, asking how two widely used satellite precipitation products can produce contrasting representations of hourly precipitation and extreme-event processes despite apparently favorable performance in some aggregate metrics. By jointly examining precipitation occurrence, error components, product-specific event characteristics, and observation-referenced matched events, the study links aggregate precipitation performance to the combinations of detection, duration, accumulation, and intensity errors that produce it.

Accordingly, this study evaluates GPM IMERG Final Run V07 and GSMaP_Gauge V8 using hourly observations from 522 stations. It addresses four questions: (1) how accurately do the products reproduce annual, seasonal, and hourly precipitation; (2) how do they trade detection probability against false alarms; (3) how do they represent event frequency, duration, accumulation, mean and peak intensity, and threshold-exceedance characteristics, and how well do they reproduce the same observation-defined extreme events in a matched-event analysis; and (4) how are their net errors partitioned among hit, missed, and false components, and what systematic scaling and random-dispersion characteristics occur during hit hours?

2. Materials and Methods

2.1. Study Area

North China is located in eastern China, approximately between 29–42°N and 110–122°E. The study domain includes Beijing, Tianjin, Hebei, and adjacent areas and comprises the Taihang Mountains in the west, the Yanshan Mountains in the north, the North China Plain in the east, and hilly terrain in the south. Elevation generally decreases from northwest to southeast [27].

The region has a temperate monsoon climate with pronounced spatial and seasonal precipitation variability. Most annual rainfall occurs from May to September [26,27] consistent with the pronounced warm-season maximum shown in Figure 1c. Precipitation generally decreases from the southeastern plain and coastal areas toward the northwestern inland and mountainous areas. Monsoon circulation, convection, and terrain jointly favor short-duration heavy rainfall [26,27].

2.2. Datasets

2.2.1. Observation Data

Hourly precipitation observations were obtained from national meteorological stations operated by the National Meteorological Information Center of the China Meteorological Administration [30,31]. The aligned archive contains 26,198 hourly time steps from 1 October 2019 00:00 to 26 September 2022 13:00. Of the original 555 stations, 522 passed the continuity and data-quality screening and were retained. The observations had undergone standard quality-control procedures, including checks for internal consistency and abnormal values [31]. The station network is denser in the North China Plain and along the eastern foothills of the Taihang Mountains, whereas stations are relatively sparse in western mountainous areas (Figure 1).

2.2.2. Satellite Data

GPM IMERG Final Run V07 combines passive-microwave and infrared observations and applies monthly gauge adjustment. It provides global precipitation estimates at 0.1° and 30-min resolution and is intended for retrospective analysis [10,12,32,33]. GSMaP_Gauge V8 is a gauge-adjusted JAXA product at 0.1° and hourly resolution [11,13]. Because both products incorporate gauge information, comparisons against station observations are not fully independent; this issue is considered when interpreting the results.

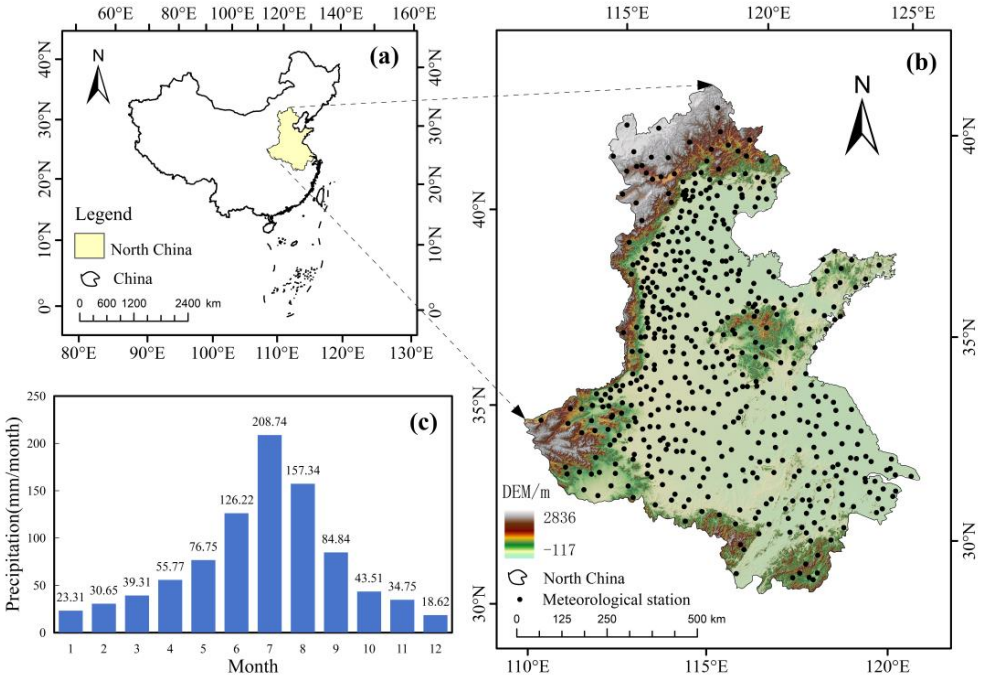


Figure 1. Study area in North China, showing (**a**) the location of the study domain within China, (**b**) topography and the distribution of meteorological stations, and (**c**) station-mean monthly precipitation during the study period from 1 October 2019 to 26 September 2022.

Station timestamps were converted from Beijing Time (UTC + 8) to UTC before temporal matching. For IMERG, the two 30-min precipitation rates within each hourly interval were integrated as $(0.5r_1 + 0.5r_2)$ to obtain the corresponding hourly precipitation amount. The satellite hourly estimates were labeled by the start of each accumulation interval, whereas the gauge observations were indexed by the interval end time. Accordingly, a satellite estimate labeled 00:00 UTC was paired with the gauge observation labeled 01:00 UTC, both representing precipitation during 00:00–01:00 UTC. To ensure consistent temporal alignment, satellite timestamps were adjusted to the corresponding interval end times before matching. Satellite values were extracted from the 0.1° grid cells containing each station and matched to the station observations using exact timestamps rather than row positions. Calculations used pairwise-valid samples. To assess the sensitivity of the hourly evaluation to potential temporal displacement, the temporally aligned satellite series were additionally shifted by −1 and +1 h relative to the corrected zero-lag reference, and POD, FAR, and MIS were recalculated using common station–time samples available under all three lag settings.

2.3. Statistical Metrics for Precipitation Errors

Precipitation-amount performance was quantified using relative bias (RB), the Pearson correlation coefficient (CC), and root mean squared error (RMSE) (Table 1). RB measures

systematic wet or dry bias, CC describes linear correspondence, and RMSE summarizes the magnitude of hourly differences [13].

Table 1. Definitions of the evaluation metrics.

Metric	Unit	Formula	Range	Optimum Value
Relative bias (RB)	%	$RB = \left[\frac{\sum(S_i - O_i)}{\sum O_i}\right] \times 100\%$	(-∞, +∞)	0
Correlation coefficient (CC)	-	$CC = \frac{\sum_{i=1}^{n}(S_i - \overline{S})(O_i - \overline{O})}{\sqrt{\sum_{i=1}^{n}(S_i - \overline{S})^2}\sqrt{\sum_{i=1}^{n}(O_i - \overline{O})^2}}$	[−1, 1]	1
Root mean squared error (RMSE)	mm h^{-1}	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(S_i - O_i)^2}$	[0, +∞)	0
Probability of detection (POD)	-	$POD = \frac{H}{H+M}$	[0, 1]	1
False alarm ratio (FAR)	-	$FAR = \frac{F}{H+F}$	[0, 1]	0
Miss rate (MIS)	-	$MIS = \frac{M}{H+M}$	[0, 1]	0

Note: S_i denotes the satellite precipitation estimate, O_i denotes the station observation, $\overline{S}$ and $\overline{O}$ denote their respective means during the study period, and n is the number of paired samples. H denotes the number of hits, M denotes misses, and F denotes false alarms.

Precipitation occurrence was evaluated with probability of detection (POD), false alarm ratio (FAR), and miss rate (MIS), using 0.1 mm h^{-1} as the wet-hour threshold. POD is the fraction of observed wet hours detected by a product, FAR is the fraction of product wet hours not observed by gauges, and MIS is the fraction of observed wet hours missed by the product [34]. POD and MIS are complementary, whereas POD and FAR jointly describe the sensitivity–false-alarm trade-off.

For precipitation-occurrence metrics, a wet hour was defined as precipitation $\geq$ 0.1 mm h^{-1}. This threshold was retained because it is consistent with the minimum measurable precipitation used in large-sample hourly gauge evaluations over China and has been widely adopted as a rain/no-rain threshold in hourly satellite-precipitation assessments [35–37]. Precipitation intensity and extreme-event characteristics were evaluated separately using the local 90th-percentile event framework described in Section 2.4. All hourly metrics were calculated separately at each station using pairwise-valid samples. Regional values reported in the text were arithmetic means of the station-level metrics across the 522 retained stations, and seasonal metrics were calculated separately using samples within each season. POD, FAR, and MIS were multiplied by 100 for presentation as percentages. For annual and seasonal spatial evaluation, precipitation was accumulated separately at each station and averaged over the available years. Taylor diagrams summarized the spatial correlation and normalized standard deviation of station-level multi-year mean precipitation relative to STN, whereas the scatterplots compared the corresponding station-level accumulations.

To assess the sensitivity of the hourly occurrence evaluation to potential temporal displacement, the satellite series were additionally shifted by −1 and +1 h relative to the fixed station timestamps. POD, FAR, and MIS were recalculated for each lag using the common station–time samples available under all three settings. The lag analysis was used to assess temporal sensitivity and to verify the revised alignment of the satellite and gauge precipitation records.

Uncertainty in the regional summary statistics was quantified using station-level bootstrap resampling with 5000 replicates. In each replicate, stations were sampled with replacement and the corresponding regional statistic was recalculated using the same

aggregation rule as in the main analysis. The 2.5th and 97.5th percentiles of the bootstrap distributions were used as the 95% confidence intervals. For direct comparisons between IMERG and GSMaP, a paired bootstrap was applied using stations with valid values for both products, with identical resampled station indices retained for the two products in each replicate. Paired differences were defined as GSMaP − IMERG.

2.4. Identification and Characteristics of Extreme Precipitation Events

Extreme precipitation events (EPEs) were identified from the quality-controlled hourly series using an event-identification algorithm. For event segmentation, precipitation hours with $p \geq 0.1$ mm h^{-1} were treated as wet, consistent with the event-identification algorithm (Figure 2). Consecutive wet hours formed a precipitation event, and a dry interval of at least 1 h separated adjacent events (minimum inter-event time, MIT = 1 h) [38]. Only events lasting at least 3 h were retained. For each station and each satellite series, the local extreme threshold was independently defined as the 90th percentile of wet-hour precipitation. A retained event was classified as an EPE if it contained at least one hour exceeding the corresponding local threshold [25,39].

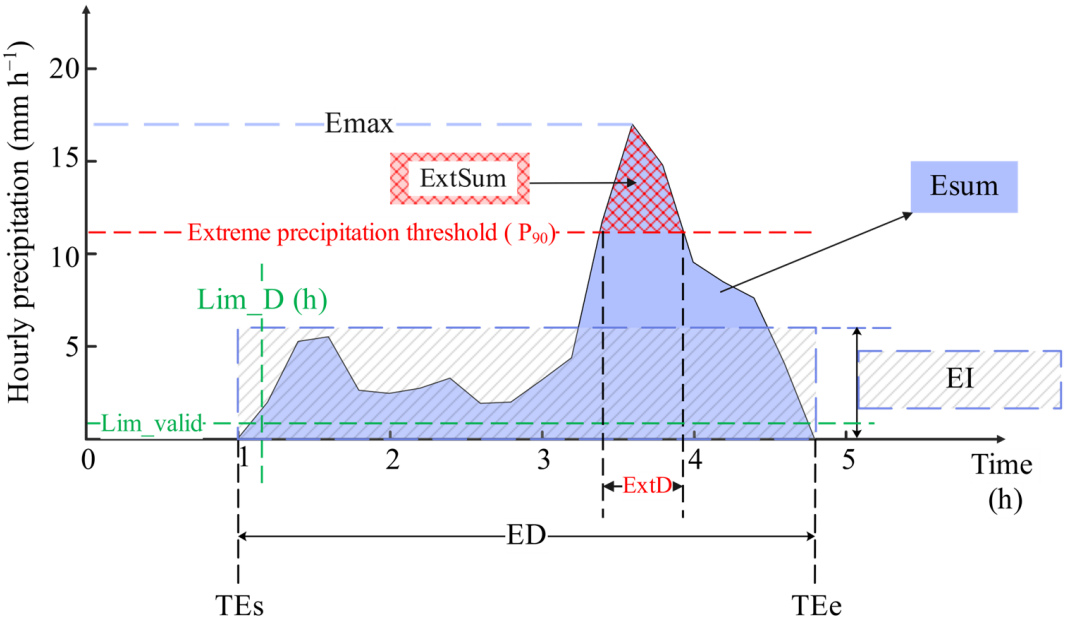


Figure 2. Schematic representation of an extreme precipitation event and its characteristics. Lim_valid is the wet-hour threshold (0.1 mm h^{-1}), and Lim_D is the minimum event duration (3 h). TEs and TEe denote the event start and end times, respectively; ED is event duration; Esum is event precipitation total; EI (=Esum/ED) is mean event intensity; Emax is peak hourly precipitation; ExtD is the number of hours exceeding the local P90 threshold; and ExtSum is the precipitation accumulated during those hours.

For each EPE, five basic characteristics were calculated: event frequency (*EF*), event duration (*ED*; h), event precipitation total or severity (*Esum*; mm), mean event intensity (*EI* = *Esum*/*ED*; mm h^{-1}), and peak hourly precipitation (*Emax*; mm h^{-1}) [25,39]. *EF* was annualized using the effective duration of the hourly record.

Three threshold-exceedance characteristics were also calculated: extreme duration (*ExtD*; number of hours above the local P_{90} threshold), extreme precipitation amount (*ExtSum*; accumulated precipitation during those hours), and extreme mean intensity

(*ExtAvg* = *ExtSum*/*ExtD*) [39]. Characteristics were first averaged across events at each station and then summarized across stations.

$$ED = T_{E_e} - T_{E_s} + 1 \tag{1}$$

$$Emax = \max_{T_{E_s} \le t \le T_{E_e}} P_t \tag{2}$$

$$Esum = \sum_{t=T_{E_s}}^{T_{E_e}} P_t \tag{3}$$

$$EI = \frac{Esum}{ED} = \frac{\sum_{t=T_{E_s}}^{T_{E_e}} P_t}{T_{E_e} - T_{E_s} + 1} \tag{4}$$

where T_{E_s} and T_{E_e} denote the first and last hourly time steps of an event, respectively, and P_t is hourly precipitation. The +1 term accounts for inclusive hourly indexing. *ED* is expressed in hours, *Esum* in millimeters, and *EI* and *Emax* in millimeters per hour.

$$ExtD = \sum_{t=T_{E_s}}^{T_{E_e}} I(P_t > P_{90}) \tag{5}$$

$$ExtSum = \sum_{t=T_{E_s}}^{T_{E_e}} P_t I(P_t > P_{90}) \tag{6}$$

$$ExtAvg = \frac{ExtSum}{ExtD} \tag{7}$$

where P_t is hourly precipitation at time t, P_{90} is the local 90th-percentile threshold, and $I(\cdot)$ is an indicator function equal to 1 when the condition is satisfied and 0 otherwise. *ExtD* represents the number of threshold-exceedance hours, *ExtSum* is the accumulated precipitation during those hours, and *ExtAvg* is their mean precipitation intensity [39]. Relative errors of regional means were calculated as 100 × (Product − STN)/STN. Because EPEs were identified independently in each series, the comparison quantifies differences in event-characteristic distributions rather than one-to-one matching of individual storms.

To further assess whether the satellite products detected the same extreme precipitation events observed by the gauges, we conducted an additional observation-referenced event-matching analysis. Gauge-defined EPEs were used as the reference, and the gauge-derived local 90th-percentile threshold was applied consistently to both satellite products. Satellite and gauge events sharing at least one hourly timestamp were matched on a one-to-one basis, primarily according to their temporal overlap. Missed and false events were identified, and timing, duration, accumulation, and intensity errors were evaluated for the matched events. Full matching procedures and diagnostic metrics are provided in the Supplementary Materials.

2.5. Error-Component Decomposition

2.5.1. Additive Error Components

To identify the sources of satellite precipitation error, paired hourly satellite estimates and station observations were separated into hit, missed, and false precipitation samples using the same precipitation-occurrence threshold of 0.1 mm h^{-1} (Lim_valid). A hit occurred when both the satellite estimates and station observations exceeded 0.1 mm h^{-1}. A missed precipitation sample occurred when the station observation exceeded the threshold, but the satellite estimate did not. A false precipitation sample occurred when the satellite estimate exceeded the threshold, but the station observation did not. Among pairwise-valid

samples, a correct negative occurred when neither the satellite estimates nor the station observation exceeded 0.1 mm h^{-1}. Correct-negative samples were excluded from the precipitation-error decomposition [23].

Let S_i and O_i denote the satellite estimate and station observation at hour i, respectively, and let H, M, and F denote the sets of hit, missed, and false samples. Within the error-component framework, annualized total error (TE) was defined as the sum of the hit, missed, and false components:

$$TE = HE + ME + FE \tag{8}$$

Correct-negative samples were excluded and therefore contributed zero to TE by definition.

$$HE = \frac{1}{Y}\sum_{i\in H}(S_i - O_i) \tag{9}$$

$$ME = -\frac{1}{Y}\sum_{i\in M} O_i \tag{10}$$

$$FE = \frac{1}{Y}\sum_{i\in F} S_i \tag{11}$$

where Y is the number of study years. HE represents the precipitation-intensity bias when precipitation was detected by both datasets. ME is the negative contribution caused by missed observed precipitation, whereas FE is the positive contribution caused by false satellite precipitation. All four components are expressed in mm yr^{-1}. A small TE does not necessarily indircate small component errors because negative missed or hit errors may be offset by positive false precipitation errors.

Total error (TE) and relative bias (RB) describe the same net precipitation difference in absolute and relative forms, respectively. TE is the annualized sum of satellite-minus-observation differences and satisfies TE = HE + ME + FE, whereas RB normalizes the same summed difference by the total observed precipitation. Accordingly, RB = 100 × N_γ × TE/$\Sigma_i O_i$, where N_γ is the number of study years. TE is therefore expressed in mm yr^{-1}, while RB is expressed as a percentage.

The dependence of the error components on precipitation intensity was examined using class boundaries of 0.1, 0.2, 0.4, 0.8, 1.6, 3.2, 6.4, 12.8, 25.6, and 51.2 mm h^{-1}. The precipitation amounts associated with hit, missed, and false samples were summarized separately within each intensity class to determine whether the errors were concentrated in light, moderate, or heavy precipitation. Hit and missed samples were classified according to observed precipitation O_i, whereas false samples were classified according to satellite precipitation S_i. As a complementary terrain diagnostic, terrain slope was derived from the DEM used for Figure 1 and averaged within a station-centered 0.1° × 0.1° window. Stations were grouped into four mean-slope classes (0–2°, 2–5°, 5–10°, and ≥10°), and the same station-level TE, HE, ME, and FE diagnostics used in the elevation-stratified analysis were summarized within each class.

2.5.2. Multiplicative Error Model

For hit samples, a multiplicative error model was additionally used to separate systematic differences in precipitation intensity from random uncertainty:

$$S_i = aO_i^{\beta}\exp(\varepsilon_i) \tag{12}$$

which was transformed into logarithmic form as

$$\ln(S_i) = \alpha + \beta\ln(O_i) + \varepsilon_i, \alpha = \ln(a) \quad (13)$$

The residual term ε_i was assumed to have a mean of zero and a standard deviation of σ, and the scale coefficient a is related to α through $\alpha = \ln(a)$. The parameters α and β were estimated by ordinary least squares using hit samples, for which both S_i and O_i were positive. The intercept α and slope β jointly describe systematic departures from the 1:1 relationship. The slope β describes how satellite estimates scale with observations: $\beta < 1$ indicates compression of the satellite-estimated precipitation-intensity range relative to observations, $\beta > 1$ indicates expansion, and $\beta = 1$ indicates proportional change. In simple terms, a β value close to zero means that the satellite estimate changes very little as observed precipitation becomes stronger, so the observed range from light to heavy precipitation is compressed into a much narrower range in the satellite product. The intercept α represents an overall multiplicative offset in log space and should be interpreted together with β. However, $\beta = 1$ alone does not imply unbiased estimates; the 1:1 relationship additionally requires $\alpha = 0$ ($a = 1$). The parameter σ quantifies residual random dispersion around the fitted relationship in log space and does not represent the total uncertainty of the satellite product. The ideal values are $\alpha = 0$, $\beta = 1$, and $\sigma = 0$ [23].

3. Results

3.1. Evaluation of Precipitation Errors

Observed precipitation shows a strong seasonal cycle, with maximum totals in summer, intermediate values in spring and autumn, and minimum values in winter (Figure 1c).

IMERG, GSMaP, and station observations reproduced this broad spatial gradient, but their magnitudes differed (Figure 3). IMERG systematically underestimated annual precipitation, especially at wet southern stations. GSMaP was closer to the station totals overall, with modest underestimation at lower totals and overestimation in parts of the wetter range.

Seasonal patterns were consistent with the annual results. IMERG underestimated precipitation in all seasons, with the largest absolute deficits in summer. GSMaP displayed a more balanced pattern, although wet biases occurred in parts of central and northern North China in summer and in southern North China in winter.

Taylor-diagram results showed clear seasonal differences (Figure 4). IMERG had high spatial correlations in spring and winter (0.95–0.96) and lower correlations in summer and especially autumn. GSMaP reproduced the magnitude of summer spatial variability more closely than IMERG, but its spatial correlation was also weakest in autumn. Seasonal performance therefore did not vary solely with precipitation amount.

The scatterplots reinforce these findings (Figure 5). IMERG departed increasingly from the 1:1 line at high totals, particularly in summer and autumn. GSMaP was generally closer to the 1:1 line in summer but retained regional and seasonal biases. Because these panels compare station-level annual or seasonal accumulations, they describe spatial agreement rather than hourly temporal correlation.

Hourly metrics confirmed distinct product error structures (Figure 6). IMERG had a mean RB of −41.3% (range: −100.0% to −25.2%); only 4 stations (0.8%) fell between −100% and −60%. GSMaP had a mean RB of +12.5%, and 514 stations (98.5%) fell between −20% and +40%. Mean CC values were 0.504 for IMERG and 0.501 for GSMaP, while mean RMSE values were 0.850 and 0.849 mm, respectively. The two products had nearly identical mean CC and RMSE, whereas GSMaP had a much smaller net bias.

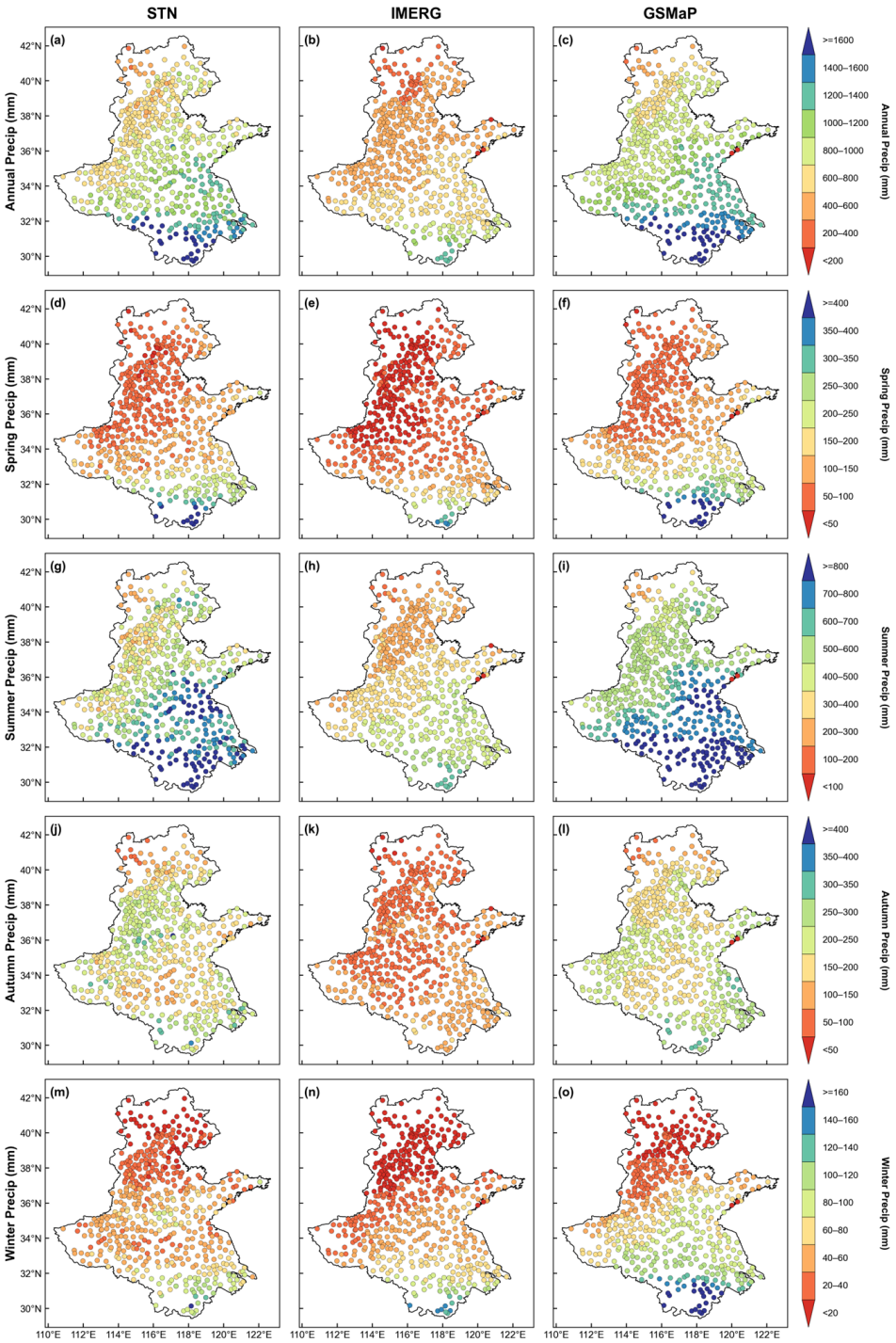


Figure 3. Spatial distribution of average annual and seasonal cumulative precipitation in North China from 1 October 2019 to 26 September 2022: annual precipitation from (**a**) STN, (**b**) IMERG, and (**c**) GSMaP; spring precipitation from (**d**) STN, (**e**) IMERG, and (**f**) GSMaP; summer precipitation from (**g**) STN, (**h**) IMERG, and (**i**) GSMaP; autumn precipitation from (**j**) STN, (**k**) IMERG, and (**l**) GSMaP; and winter precipitation from (**m**) STN, (**n**) IMERG, and (**o**) GSMaP.

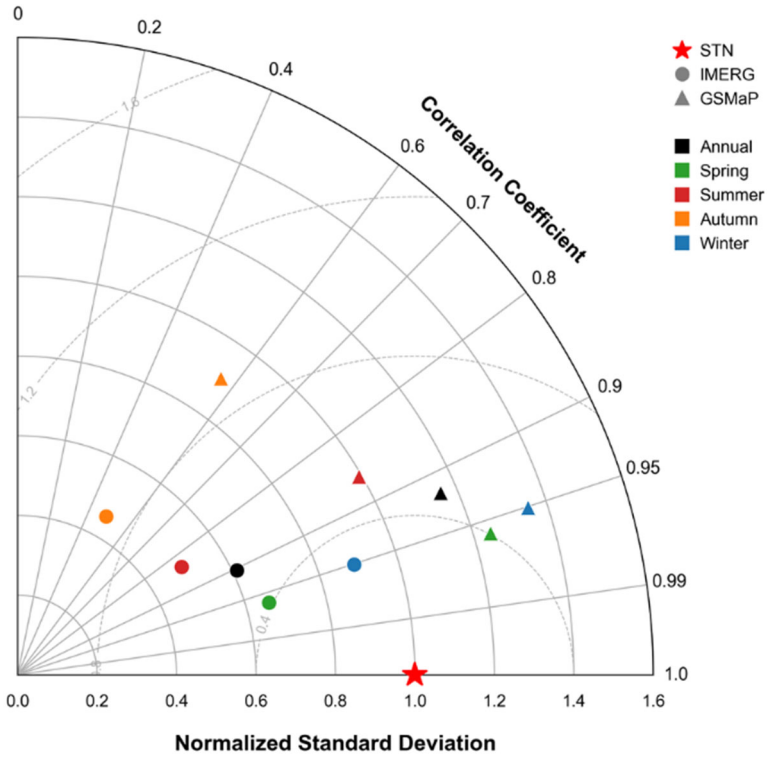


Figure 4. Taylor diagram of multi-year mean precipitation in North China during 2019–2022.

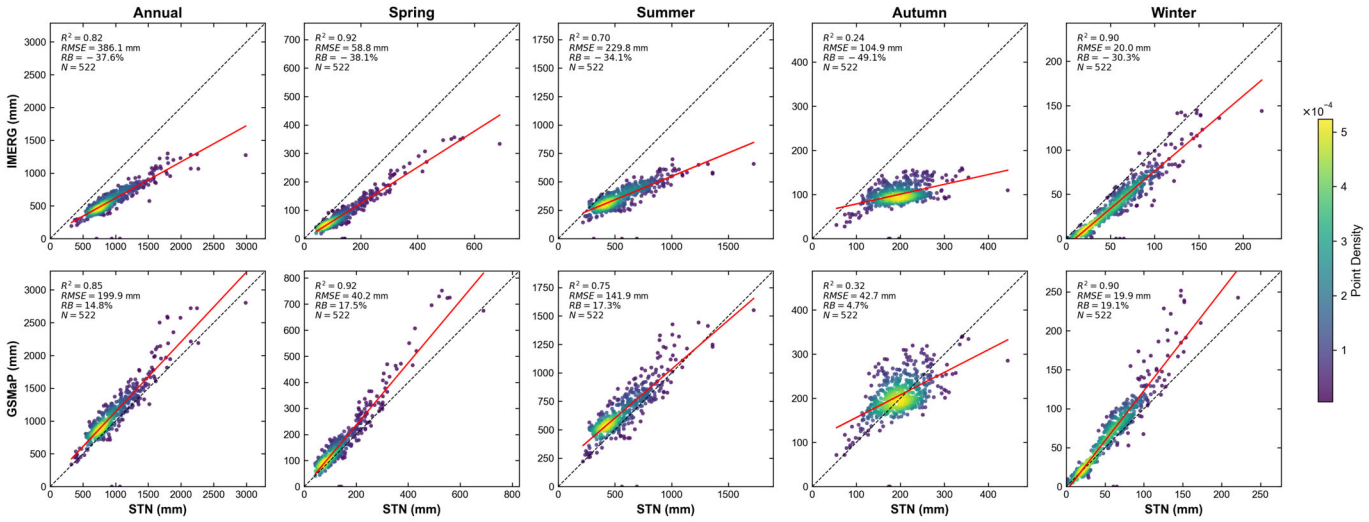


Figure 5. Scatterplot of multi-year mean precipitation in North China during 2019–2022. The solid red lines represent ordinary least-squares regression fits to station-level annual and seasonal precipitation accumulations, and the black dashed lines indicate the 1:1 relationship. R^2 denotes the coefficient of determination. Point colors indicate local scatter-point density.

GSMaP detected wet hours much more frequently than IMERG (Figure 7). Mean POD was 77.6% for GSMaP and 38.6% for IMERG, and the corresponding MIS values were 22.4% and 61.4%, respectively.

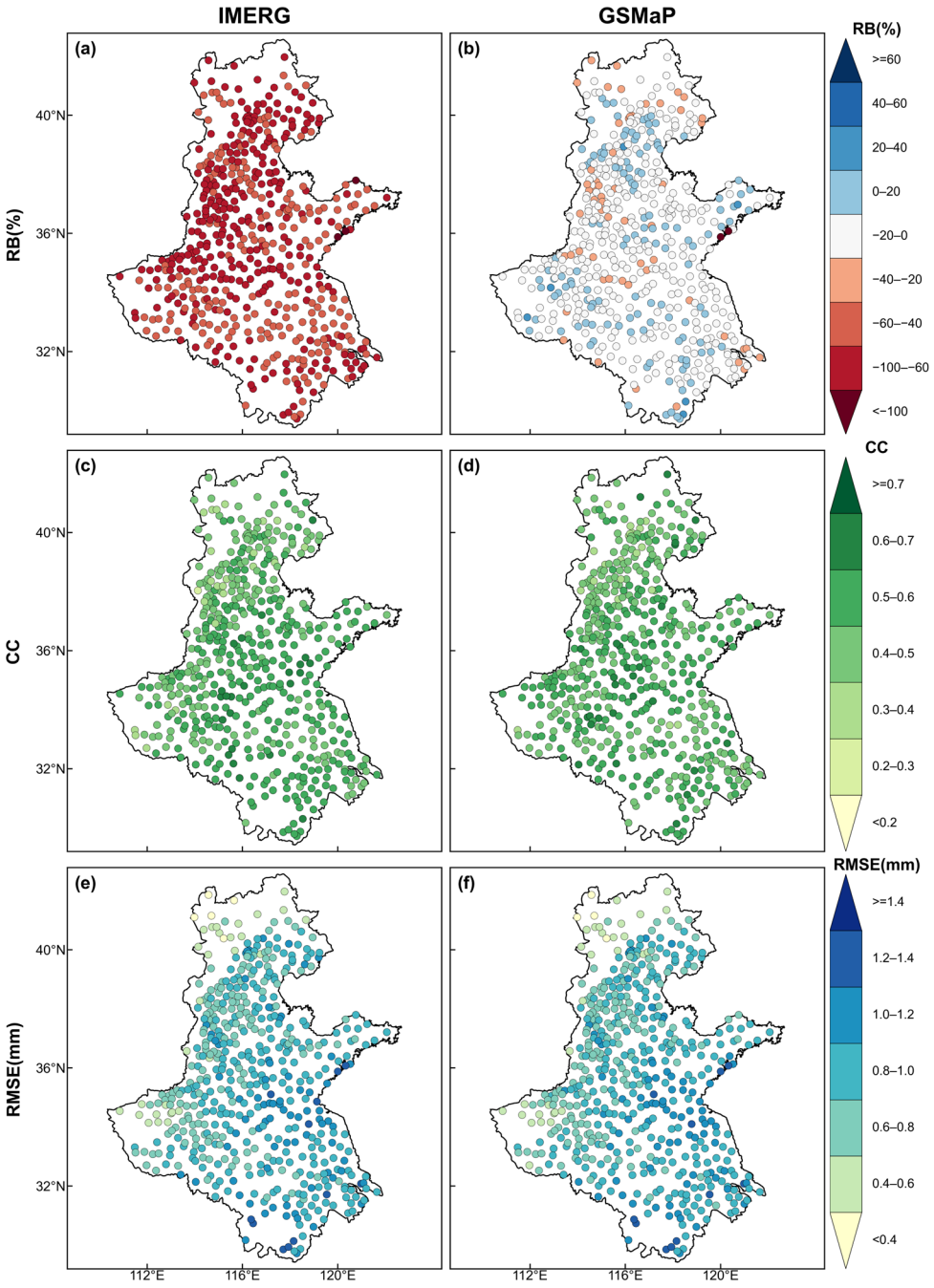


Figure 6. Spatial distributions of hourly precipitation error metrics: RB for (**a**) IMERG and (**b**) GSMaP; CC for (**c**) IMERG and (**d**) GSMaP; and RMSE for (**e**) IMERG and (**f**) GSMaP.

Higher GSMaP sensitivity was accompanied by more false precipitation. Mean FAR was 67.5% for GSMaP and 29.7% for IMERG. Thus, GSMaP captured a substantially larger fraction of observed wet hours, but almost two-thirds of its detected wet hours were false relative to the station threshold. IMERG produced fewer false detections but missed nearly two-thirds of observed wet hours.

The two products should therefore not be described as universally better or worse. Their occurrence performance reflects different operating points: GSMaP favors sensitivity, whereas IMERG is more conservative. For operational use, the preferable product depends on the relative cost of missed precipitation and false alarms.

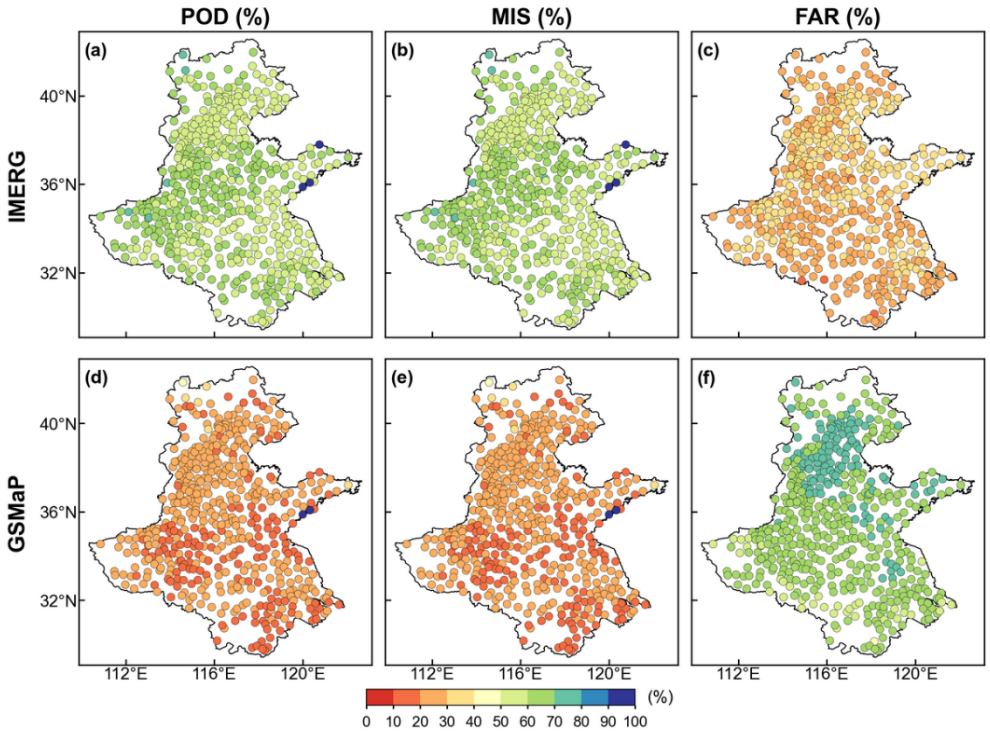


Figure 7. Spatial distributions of wet-hour occurrence metrics calculated from all matched hours during the study period: (**a**) IMERG POD; (**b**) IMERG MIS; (**c**) IMERG FAR; (**d**) GSMaP POD; (**e**) GSMaP MIS; and (**f**) GSMaP FAR. All metrics are expressed as percentages.

The contrast persisted in every season (Table 2; Figures S2–S5). POD peaked in summer for both products (53.40% for IMERG and 84.47% for GSMaP) and was lowest in winter (21.6% and 68.19%). GSMaP FAR remained higher in all seasons (59.01–71.20%) than IMERG FAR (24.51–32.65%).

Table 2. Seasonal wet-hour occurrence metrics for IMERG and GSMaP.

Season	POD (%)		MIS (%)		FAR (%)	
	IMERG	GSMaP	IMERG	GSMaP	IMERG	GSMaP
Spring	32.46 [31.93, 33.01]	75.90 [75.16, 76.61]	67.54 [66.99, 68.08]	24.10 [23.39, 24.84]	28.34 [27.77, 28.89]	66.54 [66.03, 67.07]
Summer	53.40 [52.87, 53.90]	84.47 [83.90, 84.94]	46.60 [46.09, 47.17]	15.53 [15.06, 16.12]	32.65 [32.20, 33.12]	71.20 [70.75, 71.64]
Autumn	33.53 [33.04, 34.01]	73.77 [73.19, 74.29]	66.47 [65.99, 66.98]	26.23 [25.72, 26.81]	24.51 [24.02, 25.01]	59.26 [58.64, 59.86]
Winter	21.60 [20.90, 22.30]	68.19 [66.90, 69.44]	78.40 [77.71, 79.09]	31.81 [30.54, 33.10]	25.59 [24.32, 26.95]	59.01 [58.43, 59.59]

Note: Values are regional means; numbers in square brackets indicate 95% confidence intervals.

Among the 522 CMA stations, 25 (4.79%) were identified as potential overlaps with the available international-exchange station network. Because product-specific station-ingestion records were unavailable, these stations represent potential rather than confirmed overlaps with the gauge networks used for product adjustment. Excluding these stations produced only minor changes in the principal evaluation metrics (Table S2). For IMERG, POD/FAR/MIS changed from 38.614%/29.691%/61.386% to 38.617%/29.682%/61.383%, whereas the corresponding GSMaP values changed from 77.605%/67.529%/22.395% to 77.498%/67.639%/22.502%. Changes in RB, CC, and RMSE were also small. The principal contrast between IMERG and GSMaP therefore remained unchanged after excluding the

potentially overlapping stations. Gauge adjustment does not, however, imply that the hourly precipitation sequence is fully constrained by gauge observations. For IMERG Final, the monthly gauge adjustment primarily constrains accumulated precipitation, while hourly occurrence, timing, and intensity still depend on the underlying satellite retrieval and temporal processing. Hourly-scale errors can therefore persist even when accumulated precipitation is improved. Similarly, gauge adjustment in GSMaP_Gauge does not eliminate subdaily retrieval and timing errors. Thus, the persistence of substantial POD, FAR, and hourly precipitation errors is not inconsistent with the use of gauge-adjusted products.

The paired station-bootstrap analysis showed that the 95% confidence intervals for all 12 seasonal differences in POD, MIS, and FAR between GSMaP and IMERG excluded zero (Table S4). Across all four seasons, GSMaP had higher POD and FAR and lower MIS than IMERG, confirming that the contrasting wet-hour occurrence characteristics were robust across the station network.

3.2. Representation of Extreme-Precipitation Event Characteristics

Across the stations, the station observations yielded a mean annualized EPE frequency of 14.68 events station^{-1} yr^{-1}. The corresponding frequencies were 9.70 events station^{-1} yr^{-1} for IMERG (−33.9%) and 19.32 events station^{-1} yr^{-1} for GSMaP (+31.6%) (Table S1). Mean event duration was 8.01 h for the station observations, compared with 7.13 h for IMERG (−11.0%) and 23.92 h for GSMaP (+198.6%).

IMERG underestimated Esum, EI, and Emax by −36.1%, −35.2%, and −52.3%, respectively. GSMaP reproduced mean Esum closely (+1.1%) but underestimated EI and Emax by 67.3% and 67.4%. The apparently accurate GSMaP event total therefore resulted from compensating errors: rainfall was distributed over events that were almost three times too long and substantially too weak.

Threshold-exceedance characteristics showed the same compensation. IMERG slightly underestimated ExtD (−7.1%) but underestimated ExtSum and ExtAvg by 48.9% and 46.1%. GSMaP underestimated ExtSum by only 9.3%, yet overestimated ExtD by 175.9% and underestimated ExtAvg by 71.0%. Spatial differences were consistent with these regional summaries. IMERG showed widespread deficits in event frequency, accumulation, and intensity. In contrast, GSMaP showed widespread duration excess together with strong deficits in peak and mean intensity (Figure 8, Figures S6 and S7). Because EPEs were identified independently in each series, these differences characterize event distributions rather than matched storm-by-storm errors.

The observation-referenced event matching broadly confirmed contrasting event representations by the two products. Of 22,902 gauge-defined EPEs, IMERG and GSMaP matched 7392 and 8006 events, corresponding to E-POD values of 32.28% and 34.96%, respectively; their event-level false-alarm ratios were 32.13% and 18.57%. For successfully matched events, IMERG reproduced event duration relatively closely (median error = 0 h), whereas GSMaP substantially broadened the events (median error = +11 h). For GSMaP, the median fraction of matched satellite-event precipitation occurring outside the gauge-defined event window was 28.76%, and its median accumulation error changed from −15.88% within the fixed gauge window to +27.30% when its extended event boundary was used. This indicates that temporal broadening materially affected its apparent event-total precipitation. Full event-matching diagnostics are provided in Table S3.

Paired bootstrap comparisons of the eight annual EPE characteristics likewise yielded 95% confidence intervals that excluded zero (Table S4). Relative to IMERG, GSMaP had higher EF, ED, Esum, ExtSum, and ExtD but lower EI, Emax, and ExtAvg, indicating that the principal contrasts in event characteristics were robust across the station network.

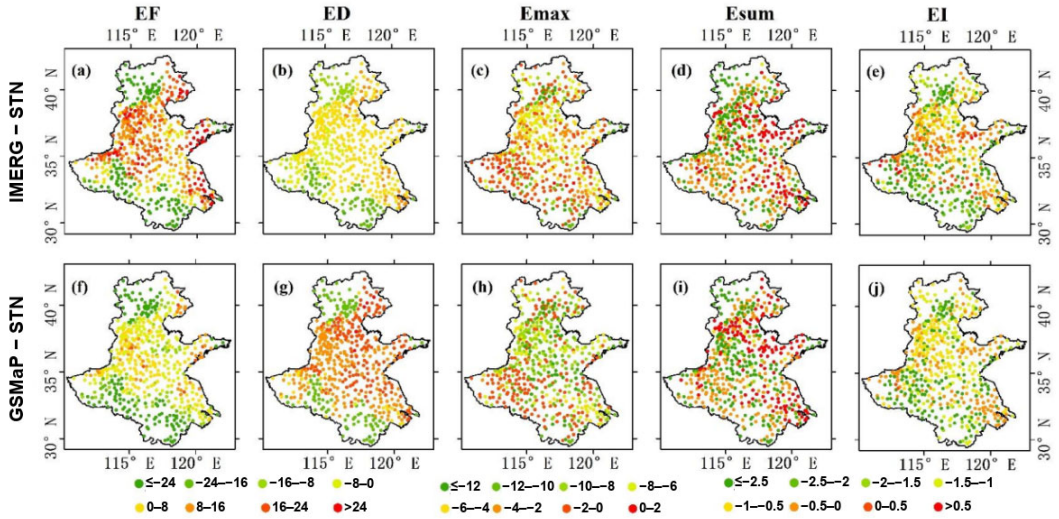


Figure 8. Spatial distributions of station-level differences (product − STN) in mean EPE characteristics. Panels (**a**–**e**) show IMERG − STN, and panels (**f**–**j**) show GSMaP − STN. EF is expressed in events yr^{-1}, ED in h, Esum in mm, and EI and Emax in mm h^{-1}.

3.3. Error-Component Analysis

Mean annualized total error (TE) was −348.2 mm yr^{-1} (median, −316.8 mm yr^{-1}; interquartile range, −388.2 to −269.4 mm yr^{-1}; Figure 9). Mean hit error (HE) and missed error (ME) were −201.1 and −252.9 mm yr^{-1}, whereas false error (FE) contributed +105.8 mm yr^{-1}. False precipitation offset only 23.3% of the combined negative hit and missed components, leaving a persistent dry bias.

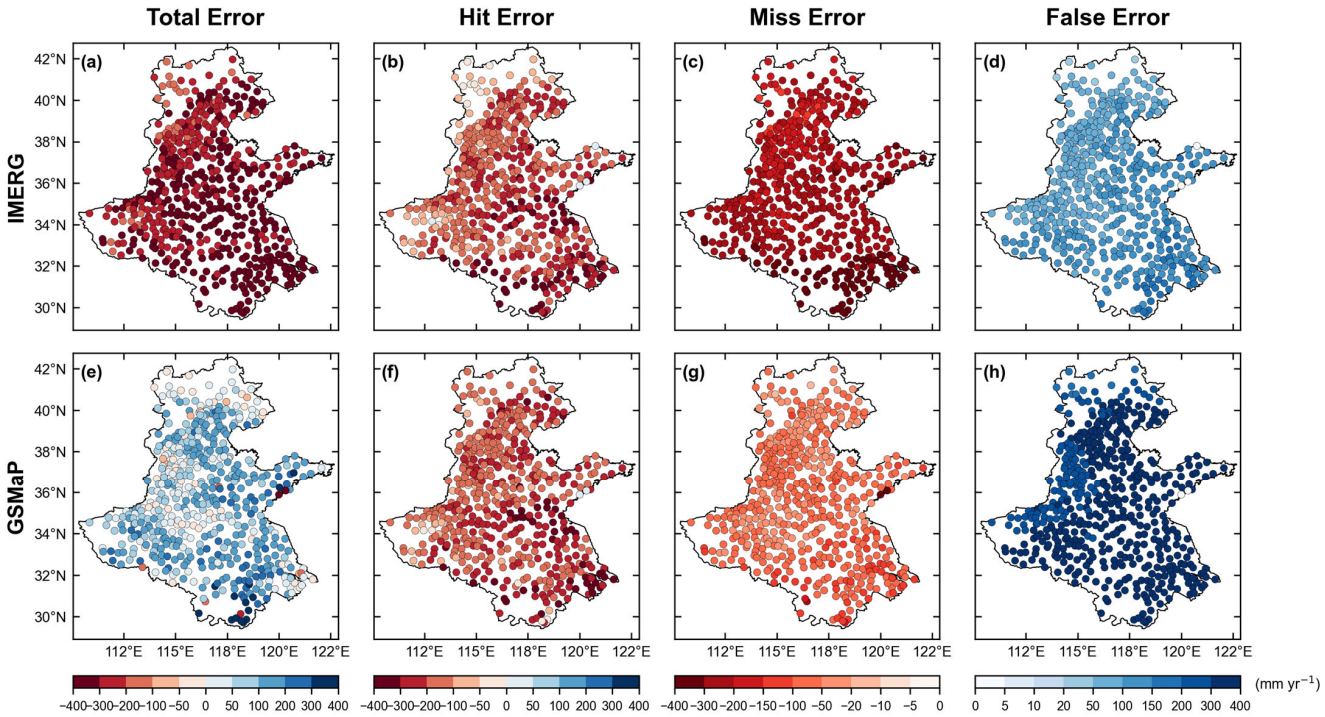


Figure 9. Spatial distributions of mean annualized additive error components for IMERG (**a**–**d**) and GSMaP (**e**–**h**), calculated from all matched hours between 1 October 2019 and 26 September 2022. TE, HE, ME, and FE denote total, hit, missed, and false errors, respectively, and are expressed in mm yr^{-1}.

GSMaP exhibited the opposite compensation structure. Mean HE and ME were −200.0 and −70.8 mm yr^{-1}, while FE reached +370.5 mm yr^{-1}, 1.37 times the combined magnitude of the two negative components. Mean TE was therefore +99.7 mm yr^{-1} (median: +97.9 mm yr^{-1}), and 88.9% of stations had positive TE. GSMaP's smaller absolute net bias than IMERG consequently concealed much larger compensating component errors.

Intensity stratification showed how these errors accumulated (Figure 10). Annualized missed precipitation totaled 243.6 mm station^{-1} yr^{-1} for IMERG and 67.6 mm station^{-1} yr^{-1} for GSMaP, whereas false precipitation totaled 128.4 and 415.6 mm station^{-1} yr^{-1}. Classes $\leq$ 1.6 mm h^{-1} contributed 80.3% and 88.9% of false precipitation, respectively. For hit samples $\geq$ 6.4 mm h^{-1}, IMERG and GSMaP reproduced only 28.8% and 33.5% of observed precipitation. Underestimation in the 25.6 mm h^{-1} class reached 81.6% and 80.4%, increasing to 87.8% and 87.5% in the 51.2 mm h^{-1} class.

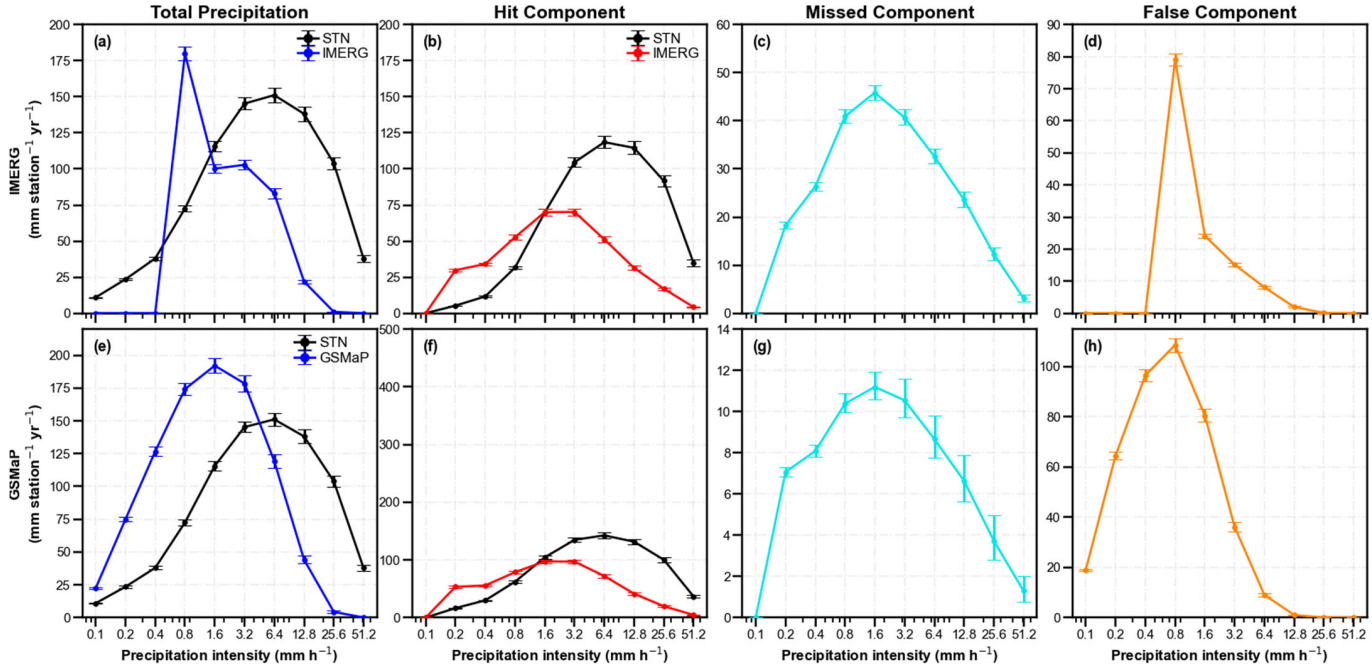


Figure 10. Mean annualized precipitation amounts by intensity class for IMERG (**a**–**d**) and GSMaP (**e**–**h**). Panels (**a**,**e**) compare total station and satellite precipitation; panels (**b**,**f**) compare station and satellite precipitation for hit hours; panels (**c**,**g**) show observed precipitation missed by the satellite product; and panels (**d**,**h**) show false satellite precipitation. Values are averaged across stations and expressed in mm station^{-1} yr^{-1}. Error bars indicate 95% confidence intervals.

The multiplicative model confirmed pronounced compression of the satellite-estimated precipitation-intensity range relative to station observations (Figure 11). Annual median α, β, and σ were 0.404, 0.170, and 0.550 for IMERG and −0.099, 0.374, and 0.935 for GSMaP. β was <1 at 519 of 522 IMERG stations (99.4%) and 520 of 522 GSMaP stations (99.6%), indicating near-universal range compression. GSMaP retained higher median sensitivity but had 70% greater median random dispersion. Random dispersion was largest in summer, when median σ was 0.590 for IMERG and 0.970 for GSMaP.

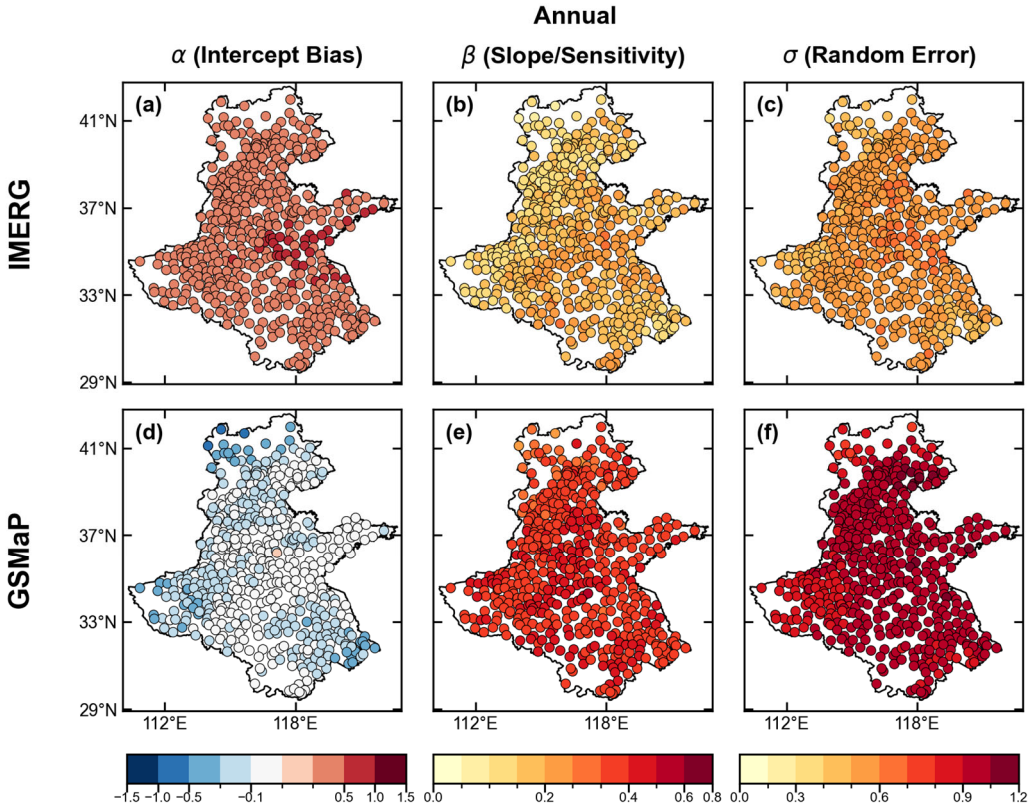


Figure 11. Spatial distributions of the annual multiplicative-error parameters α, β, and σ for IMERG (**a**–**c**) and GSMaP (**d**–**f**). Columns show α, β, and σ, respectively. The parameter α is the log-space intercept, β is the slope describing sensitivity to observed precipitation, and σ is the standard deviation of the log-space residuals.

Elevation-stratified diagnostics showed that these signatures persisted across all twenty 200-m elevation bands (Figure S8). Annual IMERG TE remained negative in every band (-0.052 to -0.022 mm h^{-1}), while GSMaP TE remained positive ($+0.004$ to $+0.020$ mm h^{-1}). Hit and missed components were negative and false components positive throughout the elevation range. The contrast intensified in summer, indicating that the product-level structures were not confined to one terrain class. A complementary slope-stratified analysis showed the same product-level error structure across four mean-terrain-slope classes (Figure S10). IMERG retained negative TE, HE, and ME and positive FE in all slope classes, whereas GSMaP retained positive TE and FE and negative HE and ME. Error-component magnitudes were generally larger in summer, while no consistent monotonic relationship with terrain slope was evident.

4. Discussion

4.1. Contrasting Error Structures in Hourly Precipitation Estimates

The present results demonstrate that satellite precipitation products cannot be evaluated reliably using a single aggregate statistic. IMERG's pronounced negative total bias resulted from the combined effects of underestimated precipitation during correctly detected hours and substantial missed precipitation. By contrast, the relatively small total bias of GSMaP did not indicate uniformly accurate retrieval. It arose from the compensation between large negative hit and missed components and a large positive false-precipitation component. Previous error-component analyses have likewise shown that total bias can conceal substantial and physically distinct contributions from hit, missed, and false precipitation, particularly at hourly and event scales [23–25]. Error compensation should therefore

be regarded as a central characteristic of product performance rather than as a secondary statistical detail.

The occurrence metrics help explain the contrasting error structures of the two products. GSMaP achieved a substantially higher probability of detection and consequently reduced missed precipitation. However, its high false alarm ratio generated widespread false precipitation, primarily at low intensities. IMERG occupied a more conservative detection regime, with fewer false precipitation hours but considerably more missed hours. Previous evaluations over mainland China have similarly reported that GSMaP generally detects a greater proportion of precipitation occurrences [12,17,30,40–42]. However, this greater sensitivity may be accompanied by excessive light precipitation and higher false-alarm rates. These results indicate that the preferred product depends on the intended application. A high probability of detection may be desirable for preliminary flood screening, whereas excessive false precipitation can be problematic for water-balance estimation, drought monitoring, and assessments of wet- and dry-spell persistence. Product selection should therefore consider the relative consequences of missed and falsely detected precipitation rather than relying exclusively on POD, FAR, or total bias.

The contrasting occurrence behavior of IMERG and GSMaP is also consistent with differences in their retrieval, temporal merging, and gauge-adjustment procedures. IMERG V07 intercalibrates passive-microwave retrievals and combines them with infrared information through a quasi-Lagrangian interpolation framework, whereas the Final Run gauge adjustment primarily constrains precipitation amount at the monthly scale rather than independently reconstructing hourly precipitation occurrence and timing. GSMaP V8 uses a different PMW–IR combination framework involving morphing, Kalman filtering, and histogram matching, and its gauge-adjustment procedure incorporates CPC daily gauge precipitation analyses. These differences in temporal propagation, rain/no-rain classification, and gauge adjustment may contribute to the substantially higher POD but also higher FAR of GSMaP and to the more conservative occurrence behavior of IMERG. However, because the present analysis does not isolate individual algorithm components, these mechanisms should be regarded as algorithm-consistent explanations rather than direct causal attribution.

4.2. Event-Scale Error Compensation and Temporal Broadening

The EPE characteristics reveal a second form of error compensation. GSMaP had a mean event accumulation close to the station-derived mean, but this apparent agreement was accompanied by substantially overestimated event duration and strongly underestimated mean and peak intensities. On average, GSMaP events lasted approximately 2.99 times as long as station-derived events, whereas their mean intensity was only about one third of the observed value. These opposing deviations can therefore partly offset each other in event accumulation. Because these statistics are based on independently identified event populations, this relationship should be interpreted as distribution-level compensation rather than an exact event-wise identity. Previous studies have also shown that weak false precipitation in GSMaP can extend wet periods and connect otherwise separate precipitation sequences [40,42].

The observation-referenced common-event analysis provides more direct evidence for this temporal-broadening mechanism. Among the 22,902 gauge-defined EPEs, IMERG and GSMaP matched 7392 and 8006 events, corresponding to event-level POD values of 32.28% and 34.96%, respectively. Although GSMaP exhibited a substantially higher false-alarm ratio at the hourly scale, its event-level false-event ratio was lower than that of IMERG (18.57% versus 32.13%), indicating that many additional GSMaP wet hours occurred around observed events rather than forming independent false EPEs. Consistently, the median

duration error was 0 h for IMERG, but +11 h for GSMaP, and the median fraction of event precipitation occurring outside the corresponding gauge-defined event window was 3.96% for IMERG and 28.76% for GSMaP.

This temporal extension materially affected the apparent event-total accuracy. For GSMaP, the accumulation bias was +27.30% when calculated over its own matched-event boundaries, but changed to −15.88% when precipitation was restricted to the exact gauge-defined event window; meanwhile, matched-event mean and peak intensities were underestimated by 44.74% and 40.26%, respectively. These results indicate that GSMaP's apparently favorable event accumulation partly reflects compensation between temporal extension and attenuated within-event intensity. IMERG exhibited a different error structure, with event duration closer to the observations but more missed events and persistent underestimation of event accumulation and peak intensity. Event-based evaluations in China and elsewhere have likewise shown that IMERG errors involve not only precipitation magnitude but also event detection, timing, duration, and peak intensity [28,43]. Thus, agreement in event-total precipitation alone does not demonstrate accurate representation of the underlying event process.

4.3. Implications for Satellite Precipitation Evaluation and Application

The systematic attenuation of event peaks is consistent with both the spatial scale of satellite estimates and known limitations of satellite precipitation retrieval. A rain gauge measures precipitation at a point, whereas an IMERG or GSMaP grid cell represents an areal estimate over approximately $0.1^\circ \times 0.1^\circ$. Because convective precipitation can vary substantially within such an area, point-scale maxima are not expected to be reproduced directly by grid-cell averages. During highly heterogeneous convective rainfall, substantial sub-grid spatial variability means that a point gauge may sample a localized high-intensity area and consequently record an hourly maximum greater than the corresponding grid-cell mean [44,45]. Previous studies have demonstrated that point-to-grid scale mismatch and gauge-network density can materially affect estimates of satellite-product accuracy, particularly at short temporal scales and under intense precipitation [44,46]. In addition, several retrieval-related factors may affect the representation of intense and rapidly evolving precipitation. These include passive-microwave sampling limitations, infrared-based temporal propagation or interpolation, assumptions regarding cloud and precipitation microphysics, and subsequent gauge adjustment [22,25,40,41]. The negative annual mean hit-precipitation components and the large deficits in the highest hit-precipitation classes found here therefore likely reflect the combined effects of satellite retrieval uncertainty and point-to-grid representativeness differences, rather than retrieval error alone.

These results also have implications for product correction and hydrological application. Correcting only the accumulated precipitation or overall mean bias may leave the underlying event structure largely unchanged. The relative importance of accumulation accuracy and event-structure accuracy depends on the application. For analyses dominated by integrated water input over relatively long temporal windows, such as monthly-to-seasonal precipitation accounting or broad-scale water-budget assessment, accumulated precipitation may be the primary quantity of interest. In such cases, GSMaP's comparatively small total bias and close mean event accumulation may be advantageous, provided that its large false-precipitation contribution is recognized. By contrast, applications that depend on the temporal organization of rainfall—including event-based runoff simulation, flash-flood or urban-drainage analysis, erosion assessment, and analyses of wet- and dry-spell persistence—require realistic event onset, termination, duration, intensity, and peak precipitation rather than correct accumulation alone.

The two products therefore represent different error trade-offs rather than a universal performance ranking. At the hourly occurrence scale, GSMaP detects a substantially larger proportion of observed precipitation and has a much smaller total accumulation bias, but this greater sensitivity is accompanied by many more false precipitation hours. The event-matching results further show that these additional precipitation occurrences often broaden observed events in time rather than forming independent false EPEs. IMERG exhibits a more conservative hourly occurrence pattern, with fewer false precipitation hours but considerably more missed precipitation, while successfully matched events retain durations closer to the gauge observations. For applications strongly controlled by hourly precipitation peaks, event timing, or event separation, neither product should therefore be selected solely on the basis of aggregate precipitation accuracy or hourly occurrence statistics.

For GSMaP, an accumulation-based correction could preserve excessive event duration, false light precipitation, and suppressed within-event intensity. For IMERG, it could retain missed-event errors and attenuated peaks. Hydrological studies have shown that precipitation occurrence, magnitude, temporal variability, and intensity-related errors do not affect runoff simulations in equivalent ways [47]. Thus, the preferred product and correction strategy should be determined by which precipitation attributes most directly control the target application, rather than by a single overall accuracy metric. A more defensible correction framework would therefore treat precipitation occurrence, false precipitation removal, event separation and termination, event accumulation, and within-event intensity as related but distinct targets. The corrected products should subsequently be evaluated using independently matched precipitation events and, where possible, a hydrological model. The present results identify relevant diagnostic targets for such correction, but they do not directly demonstrate improvements in runoff or flood prediction.

4.4. Limitations and Future Directions

Several limitations should be considered when interpreting the findings. First, the study period covers only approximately three years and is therefore insufficient for evaluating long-term changes or climatic trends. Second, both IMERG Final and GSMaP_Gauge incorporate gauge information during product calibration or adjustment [12,40,41]. Consequently, their evaluation against the Chinese station network cannot be considered fully independent. This limitation does not invalidate the comparison but may affect the interpretation of apparent product skill. Third, the point-to-grid scale mismatch and the uneven spatial distribution of stations introduce uncertainty into the estimated errors, particularly for localized convective extremes and sparsely gauged mountainous areas [28,46]. Future evaluations using denser gauge networks with multiple observations within individual satellite grid cells would allow a more comprehensive assessment of point-to-grid representativeness effects. Fourth, although the observation-referenced event matching provides a direct comparison of the same gauge-defined extreme events, the resulting matched-event statistics remain dependent on the selected gauge-based extreme threshold and temporal matching criterion. Alternative event definitions or matching tolerances may therefore produce somewhat different event catalogues. Finally, the effects of the identified error structures on runoff generation and flood simulation were inferred rather than tested directly. Future work should use longer records, compare gauge-adjusted and unadjusted products, examine the sensitivity of event matching to alternative event definitions, and directly evaluate the hydrological consequences of different precipitation-error structures.

5. Conclusions

Using 522 stations and all available matched hours from 1 October 2019 to 26 September 2022, this study combined conventional accuracy metrics, EPE characteristics, and additive and multiplicative error decomposition to evaluate IMERG Final Run V07 and GSMaP_Gauge V8. The principal conclusions are:

(1) IMERG exhibited a persistent dry bias, whereas GSMaP had a smaller positive net bias. GSMaP's smaller net bias concealed strong compensation, with positive false precipitation offsetting negative hit and missed components. For hit hours, relative underestimation by both products increased with observed precipitation intensity.

(2) GSMaP was more sensitive to wet-hour occurrence (POD = 77.6%) but produced a substantially higher proportion of false wet-hour detections (FAR = 67.5%). IMERG was more conservative (POD = 38.6%; FAR = 29.7%). This trade-off precludes a universal product ranking and should instead be evaluated against application-specific costs.

(3) The event-based analysis showed that IMERG yielded lower event frequency, accumulation, and intensity than the station observations, while its mean event duration was comparatively close to the station-derived value. GSMaP yielded higher event frequency and duration but substantially lower mean and peak intensities. Its near-unbiased mean event accumulation was consistent with compensation between overestimated duration and underestimated intensity. The observation-referenced event matching further showed that IMERG more closely preserved the duration of successfully detected events, whereas GSMaP tended to broaden observed events in time, and this temporal extension substantially affected its apparent event accumulation. Event-scale correction should therefore address occurrence, duration, accumulation, and intensity jointly.

These conclusions should be interpreted within the approximately three-year evaluation period examined here, and longer records will be needed to assess their temporal robustness.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/rs18213654/s1, Text S1. Observation-referenced event-matching analysis. Figure S1. Pie charts of annual mean hourly precipitation occurrence metric classes. Figure S2. Spatial distributions of mean hourly precipitation-occurrence metrics (POD, MIS, and FAR) in spring, calculated from matched hours between 1 October 2019 and 26 September 2022. Figure S3. Spatial distributions of mean hourly precipitation-occurrence metrics (POD, MIS, and FAR) in summer, calculated from matched hours between 1 October 2019 and 26 September 2022. Figure S4. Spatial distributions of mean hourly precipitation-occurrence metrics (POD, MIS, and FAR) in autumn, calculated from matched hours between 1 October 2019 and 26 September 2022. Figure S5. Spatial distributions of mean hourly precipitation-occurrence metrics (POD, MIS, and FAR) in winter, calculated from matched hours between 1 October 2019 and 26 September 2022. Figure S6. Spatial distributions of the basic EPE characteristics derived from station observations, IMERG, and GSMaP between 1 October 2019 and 26 September 2022. Figure S7. Spatial distributions of the threshold-exceedance EPE characteristics derived from station observations, IMERG, and GSMaP between 1 October 2019 and 26 September 2022. Figure S8. Elevation-stratified additive error components for IMERG (left column) and GSMaP (right column) at annual and seasonal scales. Black and red lines denote total and hit errors, respectively, and cyan and orange shading denote missed and false error components. All components are expressed in mm h^{-1}. Figure S9. Seasonal spatial distributions of the multiplicative-error parameters α, β, and σ for IMERG and GSMaP. Columns show spring, summer, autumn, and winter. Rows, from top to bottom, show α for IMERG, α for GSMaP, β for IMERG, β for GSMaP, σ for IMERG, and σ for GSMaP. The parameter α is the log-space intercept, β is the slope describing sensitivity to observed precipitation, and σ is the standard deviation of the log-space residuals. The in-range colour scales match the corresponding parameters in Figure 11;

triangular colour-bar ends mark values outside the annual range. Figure S10. Slope-stratified additive error components for IMERG and GSMaP. Stations were grouped according to mean terrain slope within a station-centered 0.1° × 0.1° window into 0–2° (n = 392), 2–5° (n = 82), 5–10° (n = 27), and ≥10° (n = 21) classes. Panels show annual and seasonal mean TE, HE, ME, and FE. TE, HE, ME, and FE denote total, hit, missed, and false precipitation error components, respectively. Figure S11. Sensitivity of hourly precipitation-occurrence metrics (POD, FAR, and MIS) to ±1 h satellite timestamp offsets for IMERG and GSMaP. Points represent station-mean values, and shaded bands indicate the 25th–75th percentile across stations. Temporal offsets are defined relative to the corrected zero-lag alignment, with gauge observations at time t compared against the temporally aligned satellite series at t + k. Table S1. Mean characteristics of extreme precipitation events (EPEs) identified from station observations, IMERG, and GSMaP, with 95% bootstrap confidence intervals. Table S2. Sensitivity of the principal evaluation metrics to the exclusion of potentially overlapping gauge stations. The full evaluation network contains 522 CMA stations. A total of 25 stations were identified as potential overlaps with the available CMA international-exchange station network and were excluded in the sensitivity test, leaving 497 stations. POD, FAR, and MIS are expressed as percentages; RB denotes relative bias, CC the Pearson correlation coefficient, and RMSE the root mean squared error. Table S3. Observation-referenced event-matching performance of IMERG and GSMaP for gauge-defined extreme precipitation events. Table S4. Paired bootstrap differences between GSMaP and IMERG for seasonal wet-hour occurrence metrics and annual EPE characteristics.

Author Contributions: Conceptualization, H.G.; methodology, D.X. and H.G.; software, L.Z.; validation, L.Z. and H.G.; formal analysis, Y.W.; investigation, L.Z.; resources, H.G.; data curation, L.T.; writing—original draft preparation, L.T.; writing—review and editing, L.T., W.C., H.G. and D.X.; visualization, H.G.; supervision, H.G. and P.D.M.; project administration, H.G. and P.D.M.; funding acquisition, H.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of Shandong Province (Grant No. ZR2025MS536) and the Rizhao Natural Science Foundation (Grant No. RZ2024ZR12).

Data Availability Statement: The hourly station precipitation observations used in this study were obtained from the National Meteorological Information Center of the China Meteorological Administration through the China Meteorological Data Service Centre (https://data.cma.cn/; accessed on 21 July 2025). These station observations are subject to the data-access conditions of the provider and are not redistributed by the authors. The GPM Integrated Multi-satellitE Retrievals for GPM Final Run half-hourly precipitation product, Version 07 (IMERG Final Run V07; dataset short name: GPM_3IMERGHH_07), is publicly available from the NASA Goddard Earth Sciences Data and Information Services Center at https://disc.gsfc.nasa.gov/datasets/GPM_3IMERGHH_07/summary (accessed on 1 January 2026). The gauge-calibrated Global Satellite Mapping of Precipitation standard product, algorithm Version 8 (GSMaP_Gauge V8), is available from the JAXA Global Rainfall Watch at https://sharaku.eorc.jaxa.jp/GSMaP/ (accessed on 3 September 2025). The derived statistical data and analysis code supporting the findings of this study are available from the corresponding author upon reasonable request, subject to applicable data-use restrictions.

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

Abbreviations

The following abbreviations are used in this manuscript:

GPM	Global Precipitation Measurement
IMERG	Integrated Multi-satellitE Retrievals for GPM
GSMaP	Global Satellite Mapping of Precipitation
STN	Station observations
EPE	Extreme precipitation event
POD	Probability of detection
FAR	False alarm ratio
MIS	Miss rate

RB	Relative bias
CC	Pearson correlation coefficient
RMSE	Root mean squared error
TE	Total error
HE	Hit error
ME	Missed error
FE	False error

References

1. Kong, F.; Shi, P.; Fang, J.; Lv, L.; Fang, J.; Guo, J. Progress and prospects of research on the spatiotemporal pattern changes and influencing factors of extreme precipitation under global change. *J. Catastrophol.* **2017**, *32*, 165–174.
2. Li, Z.; Liang, H.; Chen, S.; Li, X.; Li, Y.; Wei, C. Performance assessment of satellite-based precipitation products in the 2023 summer extreme precipitation events over North China. *Atmosphere* **2024**, *15*, 1315. [CrossRef]
3. Pei, Z.; Cui, X.; Jiang, R. 40-year statistics of warm-season extreme hourly precipitation over North China. *Weather Clim. Extrem.* **2025**, *48*, 100761. [CrossRef]
4. Deng, S.; Liu, H.; Feng, X.; Wang, Y. Interdecadal change in summer extreme rainfall over North China around the early 2000s: Drivers and mechanisms. *Atmos. Sci. Lett.* **2026**, *27*, e70050. [CrossRef]
5. Xu, H.; Liu, S.; Chen, J.; Wang, H.; Bi, N.; Yu, Y.; Wang, Y.; Wu, X. Changes in hydrological characteristics of the yellow river estuary under the influence of the Henan "7.20" rainstorm flood. *Mar. Geol. Front.* **2025**, *41*, 1–14.
6. You, Z.; Lan, L. A comparative study of decision-making key elements: The Haihe River Basin extreme flood in July 2023 and the Henan Zhengzhou torrential rain-caused disaster on July 20. *J. China Emerg. Manag. Sci.* **2026**, *1*, 44–56.
7. Mo, X.; Liang, X. Analysis of the characteristics and main influencing factors of numerical forecast errors of the extreme precipitation in North China in July 2023. *J. Arid Meteorol.* **2026**, *44*, 126–137.
8. Sun, Q.; Miao, C.; Duan, Q.; Ashouri, H.; Sorooshian, S.; Hsu, K.-L. A review of global precipitation data sets: Data sources, estimation, and intercomparisons. *Rev. Geophys.* **2018**, *56*, 79–107. [CrossRef]
9. Villarini, G.; Krajewski, W.F. Review of the different sources of uncertainty in single polarization radar-based estimates of rainfall. *Surv. Geophys.* **2010**, *31*, 107–129. [CrossRef]
10. Hou, A.Y.; Kakar, R.K.; Neeck, S.; Azarbarzin, A.A.; Kummerow, C.D.; Kojima, M.; Oki, R.; Nakamura, K.; Iguchi, T. The global precipitation measurement mission. *Bull. Am. Meteorol. Soc.* **2014**, *95*, 701–722. [CrossRef]
11. Kubota, T.; Aonashi, K.; Ushio, T.; Shige, S.; Takayabu, Y.N.; Kachi, M.; Arai, Y.; Tashima, T.; Masaki, T.; Kawamoto, N.; et al. Global satellite mapping of precipitation (GSMaP) products in the GPM era. In *Satellite Precipitation Measurement*; Levizzani, V., Kidd, C., Kirschbaum, D.B., Kummerow, C.D., Nakamura, K., Turk, F.J., Eds.; Springer International Publishing: Cham, Switzerland, 2020; Volume 1, pp. 355–373.
12. Huffman, G.J.; Bolvin, D.T.; Braithwaite, D.; Hsu, K.-L.; Joyce, R.J.; Kidd, C.; Nelkin, E.J.; Sorooshian, S.; Stocker, E.F.; Tan, J.; et al. Integrated multi-satellite retrievals for the global precipitation measurement (GPM) mission (IMERG). In *Satellite Precipitation Measurement*; Levizzani, V., Kidd, C., Kirschbaum, D.B., Kummerow, C.D., Nakamura, K., Turk, F.J., Eds.; Springer International Publishing: Cham, Switzerland, 2020; Volume 1, pp. 343–353.
13. Gao, R.; Li, L.; Wang, Y.; Li, W.; Yun, Z.; Gai, Y. Improvements and limitations of the latest version 8 of GSMaP compared with its former version 7 and IMERG v06 at multiple spatio-temporal scales in Mainland China. *Atmos. Res.* **2024**, *308*, 107517. [CrossRef]
14. Liu, Z. Comparison of integrated multisatellite retrievals for GPM (IMERG) and TRMM multisatellite precipitation analysis (TMPA) monthly precipitation products: Initial results. *J. Hydrometeorol.* **2016**, *17*, 777–790. [CrossRef]
15. Tang, G.; Ma, Y.; Long, D.; Zhong, L.; Hong, Y. Evaluation of GPM day-1 IMERG and TMPA version-7 legacy products over Mainland China at multiple spatiotemporal scales. *J. Hydrol.* **2016**, *533*, 152–167. [CrossRef]
16. Ren, Y.; Yong, B.; Lu, D.; Chen, H. Evaluation of the integrated multi-satellite retrievals (IMERG) for global precipitation measurement (GPM) mission over Mainland China at multiple scales. *J. Lake Sci.* **2019**, *31*, 560–572. [CrossRef]
17. Sun, G.; Wei, Y.; Wang, G.; Shi, R.; Mo, C. Applicability analysis of GPM IMERG satellite precipitation data in Guangxi. *Yangtze River* **2022**, *53*, 98–106.
18. Gao, Y.; Xu, H.; Liu, G. Evaluation of the GSMaP estimates on monitoring extreme precipitation events. *Remote Sens. Technol. Appl.* **2019**, *34*, 1121–1132.
19. Liu, R.; Jiang, S.; Ren, L.; Wei, L.; Wang, M.; Lu, Y. Evaluation of GPM IMERG precipitation product in capturing extreme precipitation events over Mainland China. *China Rural Water Hydropower* **2021**, 57–63.
20. Shang, M.; Ren, Y.; Song, H.; Yao, Y.; Bai, L.; Li, Y.; Ma, J. Evaluation of IMERG and GSMaP precipitation products on monitoring extreme precipitation events over Mainland China. *J. Geo-Inf. Sci.* **2023**, *25*, 1813–1826.

21. Yang, X.; Zeng, S.; Lin, Z. Evaluation of monitoring ability of GPM satellite precipitation products for extreme precipitation over Sichuan province. *Remote Sens. Technol. Appl.* **2023**, *38*, 1496–1508.
22. Ding, M.; Yong, B.; Yang, Z. Extreme precipitation monitoring capability of the multi-satellite jointly retrieval precipitation products of global precipitation measurement (GPM) mission. *Natl. Remote Sens. Bull.* **2022**, *26*, 657–671. [CrossRef]
23. Tian, Y.; Lv, X.; Guo, H.; Li, J.; Meng, X.; Guo, C.; Zhu, L.; De Maeyer, P. Evaluation of GPM IMERG product over the Yellow River Basin using an improved error-component procedure. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 8918–8937. [CrossRef]
24. Guo, C.; Ning, N.; Guo, H.; Tian, Y.; Bao, A.; De Maeyer, P. Does ERA5-land effectively capture extreme precipitation in the Yellow River Basin? *Atmosphere* **2024**, *15*, 1254. [CrossRef]
25. Tian, Y.; Guo, H.; Guo, C.; Li, B.; Cao, Y.; Jiang, L.; Bao, A.; De Maeyer, P. Event-based evaluation of IMERG for extreme precipitation in China: Insights from an error decomposition approach. *J. Hydrol. Reg. Stud.* **2025**, *61*, 102657. [CrossRef]
26. Li, Z.; Hu, Z.; Wei, L.; Xiao, M. Changes in extreme precipitation events in North China from 1951 to 2013. *Remote Sens. Technol. Appl.* **2016**, *31*, 773–783.
27. Li, X. Changes in Extreme Precipitation in North China During 1960–2011. Master's Thesis, Northwest Normal University, Lanzhou, China, 2013.
28. Chen, S.; Chen, T.; Fu, J.; Chen, Y. Applicability assessment of satellite-retrieved precipitation for monitoring extreme precipitation in North China. *Chin. J. Atmos. Sci.* **2025**, *49*, 873–891.
29. Chen, S.; Chen, T.; Cao, Y.; Yang, S.; Xie, J. Applicability assessment of IMERG and GSMaP hourly precipitation products over Mainland China. *Torrential Rain Disasters* **2026**, *45*, 263–276.
30. Ng, C.-P.; Zhang, Q.; Li, W. Changes in hourly extreme precipitation over Eastern China from 1970 to 2019 dominated by synoptic-scale precipitation. *Geophys. Res. Lett.* **2021**, *48*, e2020GL090620. [CrossRef]
31. National Meteorological Information Center, China Meteorological Administration. China Surface Meteorological Station Hourly Observation Data. Available online: https://data.cma.cn/ (accessed on 21 July 2025).
32. Zhao, Q.; Wang, H.; Liu, X.; Ding, G.; Huang, L.; Gan, H. Applicability evaluation of GPM satellite precipitation data in upper Huaihe River Basin. *Yangtze River* **2025**, *56*, 110–118.
33. Qu, X.; Fu, Y.; Yuan, X.; Xie, X.; Bai, M. Applicability analysis of GPM-IMERG daily precipitation data in inner Mongolia. *Torrential Rain Disasters* **2020**, *39*, 293–299.
34. Ren, P.; Li, J.; Feng, P.; Guo, Y.; Ma, Q. Evaluation of multiple satellite precipitation products and their use in hydrological modelling over the Luanhe River Basin, China. *Water* **2018**, *10*, 677. [CrossRef]
35. Li, R.; Wang, K.; Qi, D. Validating the integrated multisatellite retrievals for global precipitation measurement in terms of diurnal variability with hourly gauge observations collected at 50,000 stations in China. *J. Geophys. Res. Atmos.* **2018**, *123*, 10423–10442. [CrossRef]
36. Li, R.; Wang, K.; Qi, D. Event-based evaluation of the GPM multisatellite merged precipitation product from 2014 to 2018 over China: Methods and results. *J. Geophys. Res. Atmos.* **2021**, *126*, e2020JD033692. [CrossRef]
37. Su, J.; Lü, H.; Crow, W.T.; Zhu, Y.; Cui, Y. The effect of spatiotemporal resolution degradation on the accuracy of IMERG products over the Huai River Basin. *J. Hydrometeorol.* **2020**, *21*, 1073–1088. [CrossRef]
38. Dunkerley, D. Identifying individual rain events from pluviograph records: A review with analysis of data from an Australian dryland site. *Hydrol. Process.* **2008**, *22*, 5024–5036. [CrossRef]
39. Gan, W.; Guo, H.; Cao, Y.; Wang, W.; Yao, N.; Wang, Y.; Maeyer, P.D. Spatiotemporal evolution and intensification of extreme precipitation events in Mainland China from 1961 to 2022. *Remote Sens.* **2025**, *17*, 2037. [CrossRef]
40. Lv, X.; Guo, H.; Tian, Y.; Meng, X.; Bao, A.; De Maeyer, P. Evaluation of GSMaP version 8 precipitation products on an hourly timescale over Mainland China. *Remote Sens.* **2024**, *16*, 210. [CrossRef]
41. Zhou, Z.; Guo, B.; Xing, W.; Zhou, J.; Xu, F.; Xu, Y. Comprehensive evaluation of latest GPM-era IMERG and GSMaP precipitation products over Mainland China. *Atmos. Res.* **2020**, *246*, 105132. [CrossRef]
42. Tan, X.; Yong, B.; Ren, L. Error features of the hourly GSMaP multi-satellite precipitation estimates over nine major basins of China. *Hydrol. Res.* **2018**, *49*, 761–779. [CrossRef]
43. Li, R.; Guilloteau, C.; Kirstetter, P.-E.; Foufoula-Georgiou, E. How well does the IMERG satellite precipitation product capture the timing of precipitation events? *J. Hydrol.* **2023**, *620*, 129563. [CrossRef]
44. Li, Y.; Han, B.; Chen, L.; Zhou, R.; Li, R. Refine the uncertainty of GPM IMERG precipitation product accounting for the inherent error from rain gauges estimations. *Earth Space Sci.* **2026**, *13*, e2025EA004745. [CrossRef]
45. Tokay, A.; Roche, R.J.; Bashor, P.G. An experimental study of spatial variability of rainfall. *J. Hydrometeorol.* **2014**, *15*, 801–812. [CrossRef]

46. Tang, G.; Behrangi, A.; Long, D.; Li, C.; Hong, Y. Accounting for spatiotemporal errors of gauges: A critical step to evaluate gridded precipitation products. *J. Hydrol.* **2018**, *559*, 294–306. [CrossRef]
47. Miao, C.; Gou, J.; Hu, J.; Duan, Q. Impacts of different satellite-based precipitation signature errors on hydrological modeling performance across China. *Earth's Future* **2024**, *12*, e2024EF004954. [CrossRef]

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.