https://doi.org/10.5194/essd-2025-20
Preprint. Discussion started: 12 February 2025

Open Access Earth System Science Data Discussions

A new upgraded high-precision gridded precipitation dataset considering spatiotemporal and physical correlations for mainland China

Jinlong Hu[1], Chiyuan Miao[1, *], Jiajia Su[1], Qi Zhang[1], Jiaojiao Gou[1, 2], Qiaohong Sun[3, 4]

1 State Key Laboratory of Earth Surface Processes and Resource Ecology, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China
2 Department of Geographic Science, Faculty of Arts and Sciences, Beijing Normal University, Zhuhai 519087, China
3 State Key Laboratory of Climate System Prediction and Risk Management/Key Laboratory of Meteorological Disaster, Ministry of Education/Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disasters, Nanjing University of Information Science and Technology, Nanjing 210044, China
4 School of Atmospheric Sciences, Nanjing University of Information Science and Technology, Nanjing 210044, China

Correspondence to: Chiyuan Miao (miaocy@bnu.edu.cn)

Abstract. Precipitation is a critical driver of the water cycle, profoundly influencing water resources, agricultural productivity, and natural disasters. However, existing gridded precipitation datasets exhibit markable deficiencies in capturing the spatiotemporal and physical correlations of precipitation, which limits their accuracy, particularly in regions with sparse meteorological stations. Therefore, this study proposes a completely new gridded precipitation generation scheme to address these issues. The long-term daily observation from 3,476 gauges and incorporated 11 related precipitation variables were utilized to characterize the correlations of precipitation. By employing an improved inverse distance weighting method combined with the machine learning-based light gradient boosting machine (LGBM) algorithm, a new high-precision, long-term, daily gridded precipitation dataset for mainland China (CHM_PRE V2) was developed, which aims to improve upon and surpass the CHM_PRE V1 dataset, developed in our previous work. Validation against 63,397 high-density gauges demonstrated that CHM_PRE V2 significantly outperforms existing datasets, achieving a mean absolute error of 1.48 mm/day and a Kling-Gupta efficiency of 0.88, representing improvements of 12.84% and 12.86%, respectively, compared to the previously optimal dataset. Regarding precipitation event detection, CHM_PRE V2 achieved a Heidke skill score of 0.68 and a false alarm ratio of 0.24, surpassing other datasets by 17.24% and 29.17%, respectively. Feature importance analysis revealed that spatiotemporal and physical correlations contributed 37.10%, 34.11%, and 28.78% to precipitation retrieval, underscoring the necessity of incorporating temporal and physical correlations. CHM_PRE V2 markedly enhances precipitation measurement accuracy, reduces overestimation of precipitation events, and provides a reliable foundation for hydrological modelling and climate assessments. This dataset features a resolution of 0.1°, spans from 1960 to 2023, and will be updated annually. Free access to the dataset can be found at https://doi.org/10.5281/zenodo.14632157 (Hu and Miao, 2025).

1 Introduction

Precipitation serves as the pivotal factor driving the water cycle, directly influencing the distribution and variability of water resources, agricultural productivity, ecosystem health, and the occurrence and progression of natural disasters (Ham et al., 2023; Sun et al., 2018; Zhang et al., 2017). At regional and global scales, gridded precipitation datasets provide detailed spatial resolution and temporal continuity, making them fundamental in hydrological and climate sciences and disaster forecasting (Qiu et al., 2024; Sun et al., 2021; Xiong et al., 2024). However, due to the high spatiotemporal variability of precipitation and the complexity of observation conditions, generating high-precision gridded precipitation data remains a formidable challenge (Jiang et al., 2023).

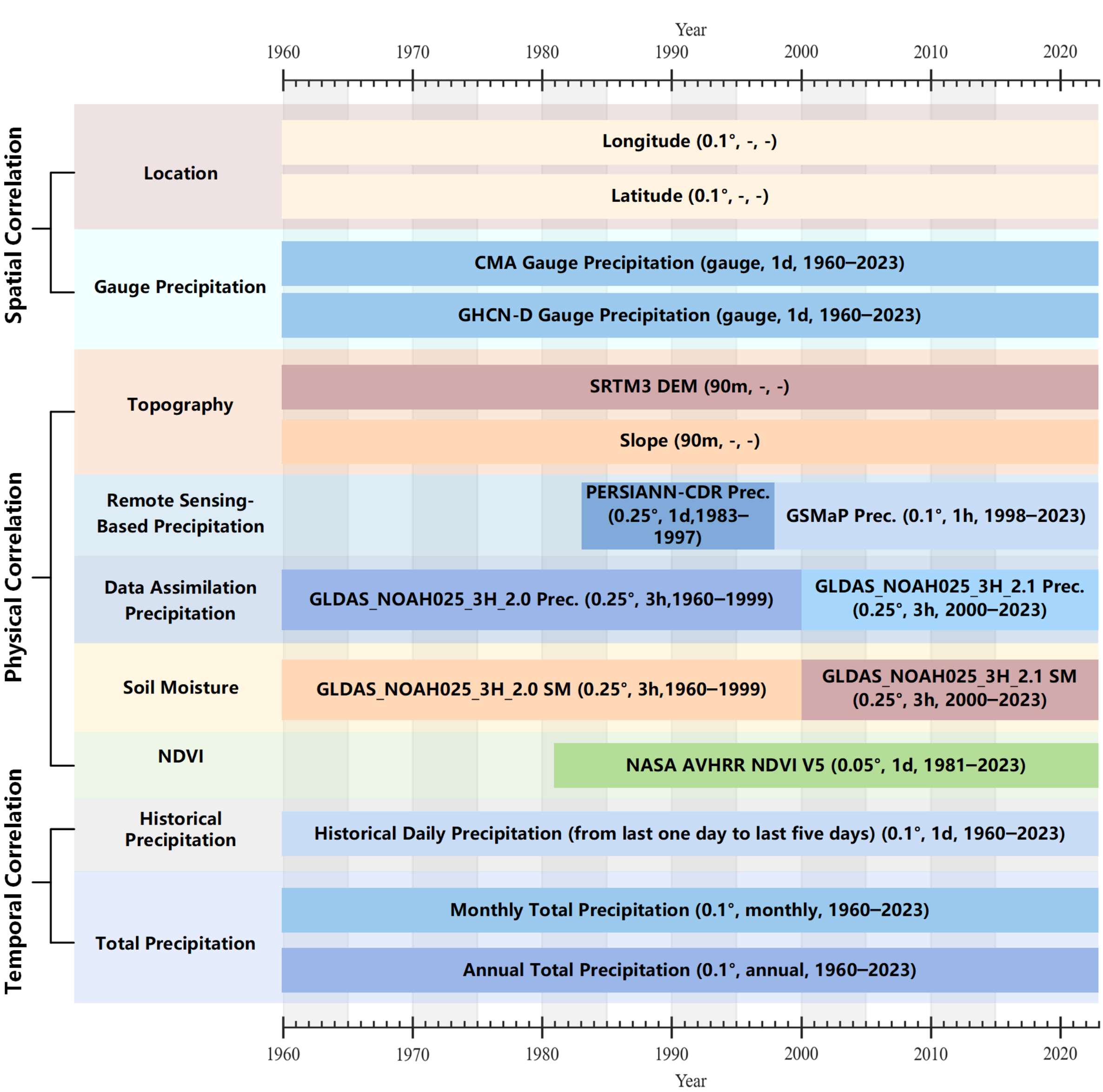
In China, various types of precipitation datasets have been extensively utilized in research, encompassing products derived from data assimilation techniques, remote sensing techniques, and gauge-based interpolation techniques. Precipitation data derived from data assimilation (Gelaro et al., 2017; Hersbach et al., 2020; Rodell et al., 2004) integrate meteorological models with observational data to provide highly consistent datasets. However, their accuracy is often constrained by the physical parameterization schemes of the models. Remote sensing-based precipitation datasets (Ashouri et al., 2015; Huffman et al., 2007, 2015; Kubota et al., 2020) offer global or regional precipitation distributions via satellite observations, ensuring extensive spatial coverage. Nonetheless, their precision is limited by data resolution and satellite orbital constraints, particularly in regions with complex terrain and high latitudes. Precipitation gauges, as the most direct and accurate tools for measuring precipitation, allow for gridded precipitation datasets generated through interpolation, effectively capturing the localized characteristics of precipitation with high accuracy (Harris et al., 2020; He et al., 2020; Qin et al., 2022; Shen et al., 2010; Wu and Gao, 2013; Xie et al., 2007). Our previous study also developed a gridded precipitation dataset for mainland China (a member of the China Hydro-Meteorology datasets, hereinafter called CHM_PRE V1) based on interpolation method, using data from 2,839 gauges. The CHM_PRE V1 demonstrates overall high accuracy across mainland China (Han et al., 2023), and has received widespread attention and extensive use, benefiting a large number of hydro meteorological related studies. However, interpolation-based precipitation datasets rely heavily on ground meteorological gauges, performing poorly in areas with sparse station distribution or missing data.

In summary, while existing precipitation datasets partially fulfil the requirements of various applications, they exhibit significant limitations. Precipitation exhibits spatiotemporal autocorrelation and physical correlation. The spatial correlation of precipitation indicates a strong interdependence between the precipitation of a given area and its surrounding regions (Chen et al., 2010, 2016; Fan et al., 2021; Huff and Shipp, 1969), forming the foundation of gauge-based precipitation interpolation. Moreover, current precipitation are also correlated with historical precipitation data and other relevant physical variables, such as elevation, cloud cover, and soil moisture, that is, the temporal and physical correlations of precipitation (Adler et al., 2008; Ham et al., 2023; Ravuri et al., 2021; Trucco et al., 2023). However, current precipitation datasets often account for either spatial correlations (especially those based on interpolation) or physical correlations (notably remote sensing and data assimilation datasets), but rarely both. This lack of comprehensive consideration for multiple correlations constrains the accuracy of these datasets, particularly in regions with sparse meteorological stations, such as western China (Jiang et al., 2023). Moreover, existing methods tend to generate excessive minor precipitation, leading to an overestimation of precipitation events, which will have considerable impacts on hydrologic modelling (Dong et al., 2020; Kang et al., 2024; Wei et al., 2022).

To address the aforementioned issues, based on the CHM_PRE V1 proposed by our previous work (Han et al., 2023), this study introduces a new high-precision, long-term daily gridded precipitation dataset for mainland China (a member of the China Hydro-Meteorology datasets, hereinafter called CHM_PRE V2). By integrating precipitation gauges, remote sensing observations, land surface assimilation outputs, and various precipitation-related factors, we employ advanced spatial interpolation and machine learning algorithms to model the spatiotemporal and physical correlations of precipitation data. As a result, we obtain a high-accuracy gridded dataset that covers the entire mainland China (18°N–54°N, 72°E–136°E) with a resolution of 0.1°. The dataset spans the period from 1960 to 2023 and will be updated annually. CHM_PRE V2 not only enhances the accuracy of precipitation measurements but also significantly reduces overestimations of precipitation events, The high-precision gridded precipitation dataset can reduce the uncertainty in hydrological modelling and analysis, providing a more reliable foundation for hydrologic and climatological studies.

2 Data

The CHM_PRE V2 dataset was constructed based on extensive precipitation observations and incorporates a wealth of additional related data to characterize the spatiotemporal and physical correlations of precipitation. This approach aims to maximize the accuracy of precipitation data, particularly in regions with sparse observational coverage. **Figure 1** illustrates details of the various datasets employed in precipitation retrieval, including the names of the datasets, their original spatial and temporal resolutions, and the time spans covered. A total of 16 datasets across 11 categories were utilized, all of these datasets can be categorized into three groups: spatially correlated datasets (encompassing four datasets), physically correlated datasets (encompassing nine datasets), and temporally correlated datasets (encompassing three datasets). In addition, the CHM_PRE V2 dataset is designed to represent precipitation characteristics across mainland China, excluding Taiwan, Hong Kong, Macau, and other Chinese islands. Next, we will provide a detailed introduction to the data used for retrieval.



90

Figure 1. The data used for precipitation retrieval.

2.1 Spatiotemporal correlated data

The spatial correlation of precipitation was characterized using location information (latitude and longitude) along with precipitation gauge data. The daily precipitation gauge data sourced from China Meteorological Administration (CMA, 95 http://data.cma.cn) spans the entirety of mainland China, encompassing records from 2,816 stations between 1960 and 2023. Daily precipitation is defined as the cumulative precipitation recorded between 20:00 on one day and 20:00 on the following day (local time in Beijing), with all data subjected to rigorous quality control (Zhang et al., 2020). To mitigate the limit of boundary effects (Ahrens, 2006), the Global Historical Climatology Network-Daily Version 3 (GHCND) dataset was employed to obtain precipitation gauges near mainland China. The GHCND is a reliable and globally comprehensive climate dataset, and maintained by the National Climatic Data Center (NCDC) of the National Oceanic and Atmospheric Administration (NOAA) (Durre et al., 2008, 2010; Menne et al., 2012). The dataset was sourced from NOAA (https://www.ncei.noaa.gov/products/land-based-station/global-historical-climatology-network-daily) on September 11, 2024. To ensure data quality, only stations with more than 70% effective days (over 255 days) in a year were included in the retrieval. **Figure 2**(a) illustrates the locations of observation stations, and **Figure 2**(b) shows the annual availability of CMA and GHCND stations. It is evident that the number of available CMA stations increased from 1,992 in 1960 to 2,767 in 2023. In contrast, the number of accessible GHCND stations in the region declined from 674 in 1960 to 264 in 2023.

Regarding the temporal correlation of precipitation, we analysed from two perspectives: overall characteristics and short-term precipitation characteristics. The total precipitation of the current month and year were used to represent the overall precipitation characteristics. The daily precipitation values from the past five days were considered as the short-term precipitation characteristics, with each day's precipitation from the past first to the past fifth day serving as five distinct variables. For example, the variable named "1st-day prior Prec." means the precipitation of the from the 1st day ago compared to the current date, and "5th-day prior Prec." means the precipitation of the 5th day ago.

2.2 Physically correlated data

The Shuttle Radar Topography Mission (SRTM) Digital Elevation Model (DEM) dataset was utilized to characterize the influence of elevation on precipitation and to generate slope data. In this study, we used the SRTM DEM V4 acquired from Consortium for Spatial Information, Consultative Group for International Agricultural Research (CGIAR-CSI, https://srtm.csi.cgiar.org/) on August 8, 2024 with a spatial resolution of 3 arc-seconds (approximately 90 meters near the equator). The SRTM DEM V4 generated based on National Aeronautics and Space Administration (NASA) SRTM DEM V1, and has undergone post-processing of the NASA data to "fill in" the no data voids, such as water bodies (lakes and rivers), areas with snow cover and in mountainous regions (e.g., the Himalayas), resulting in seamless elevation for the globe.

The Global Satellite Mapping of Precipitation (GSMaP) V8 (Kubota et al., 2020) and the Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks (PERSIANN-CDR) (Ashouri et al., 2015) datasets were used as the remote sensing-based precipitation data. GSMaP data spans from 1998 to the present, featuring a spatial resolution of 0.1° and a temporal resolution of one hour. We acquired the GSMaP data from Japan Aerospace Exploration Agency (JAXA, https://sharaku.eorc.jaxa.jp) on September 9, 2024, and used the data from 1998 to 2023. PERSIANN-CDR data spans from 1983 to the present, and the data from 1983 to 1997 was used for the retrieval.

The precipitation and soil moisture from Global Land Data Assimilation System Noah Land Surface Model (GLDAS NOAH) (Rodell et al., 2004) were also used for the retrieval. The data spans from 1960 to 1999 and the data spans from 2000 to 2023 were acquired from the GLDAS Noah L4 V2.0 and GLDAS Noah L4 V2.1 datasets. The NOAA Climate Data Record (CDR) of AVHRR Normalized Difference Vegetation Index (NDVI) (Vermote and NOAA CDR Program, 2019) was utilized to depict the vegetation characteristics, and the data from 1981 to 2023 was used.

2.3 Other datasets

To verify the reliability of the proposed CHM_PRE V2, we compared it with five existing gridded precipitation datasets. These datasets include GSMaP, PERSIANN-CDR, and GLDAS precipitation datasets, as mentioned above. Additionally, CHM_PRE V1 (Han et al., 2023), previously developed by our team, and the Integrated Multi-satellitE Retrievals for GPM (IMERG) Final L3 V7 precipitation dataset (Huffman et al., 2023) were also included in the comparison. Details of the original spatiotemporal resolution and accessible time span of these datasets are provided in **Table S1** in the supplementary materials. All datasets were resampled to daily values at a resolution of 0.1°. To ensure a fair comparison, the analysis focused on the period from 2001 to 2022, during which all datasets were available.

To further validate the reliability of precipitation data, we obtained daily precipitation observations from 72,901 high-density automatic rain gauge stations across mainland China (hereafter we refer to it as CMD-HD), provided by the National Meteorological Information Center of CMA (Li et al., 2018). The data spans the period from 2013 to 2019, and we got 63,397 available stations after quality control and annual integrity control. **Figure 2**(c) illustrates the number of CHM-HD stations within each 0.1° grid cell. The dataset demonstrates high station density throughout the eastern region, while maintaining basic coverage in the northwest and Tibetan Plateau areas. This extensive distribution ensures the validation results based on this dataset are highly reliable. Additionally, to examine the dataset's performance across various regions, we adopted the climatic regionalization scheme proposed by Ren et al.(1985), dividing China into seven distinct regions shown in **Figure 2**(d): North East China (NEC), North China (NC), South and Central China (SCC), Inner Mongolia (IM), North West China (NWC), South West China (SWC) and Qinghai-Tibet Plateau (QT).

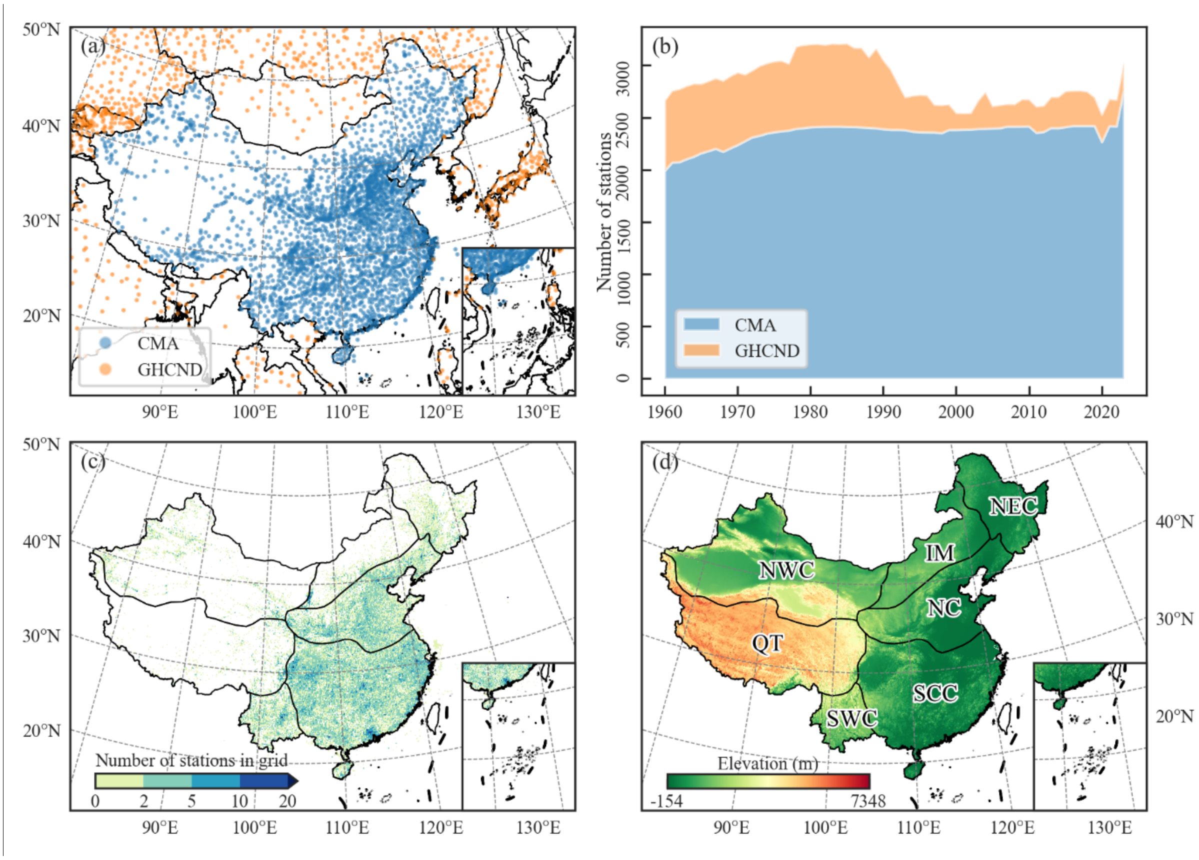


Figure 2. (a) locations of CMA and GHCND stations used for precipitation retrieval; (b) the numbers of annual availability of precipitation stations; (c) locations of CMA-HD stations used for validation; (d) climatic regions.

3 Methodology

155 The generation of CHM_PRE V2 can be divided into three stages: data preprocessing, precipitation interpolation based on spatial correlations, and precipitation retrieval based on spatiotemporal and physical correlations. **Figure 3** depicts the detailed steps involved in creating the dataset, which we will now introduce step by step.

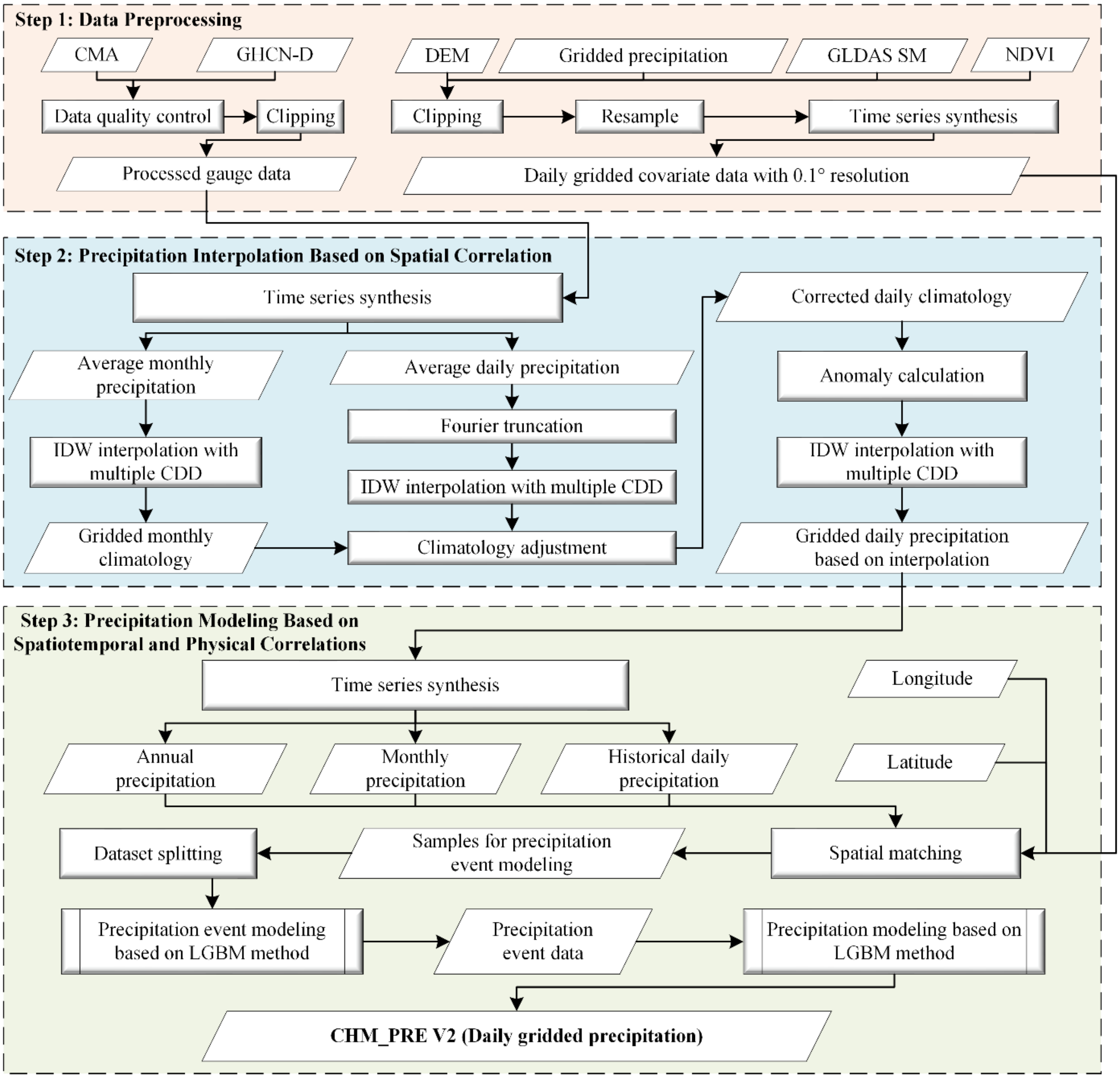

Step 1: Data Preprocessing
CMA
GHCN-D
DEM
Gridded precipitation
GLDAS SM
NDVI
Data quality control
Clipping
Clipping
Resample
Time series synthesis
Processed gauge data
Daily gridded covariate data with 0.1° resolution
Step 2: Precipitation Interpolation Based on Spatial Correlation
Time series synthesis
Corrected daily climatology
Average monthly precipitation
Average daily precipitation
Anomaly calculation
Fourier truncation
IDW interpolation with multiple CDD
IDW interpolation with multiple CDD
IDW interpolation with multiple CDD
Gridded monthly climatology
Climatology adjustment
Gridded daily precipitation based on interpolation
Step 3: Precipitation Modeling Based on Spatiotemporal and Physical Correlations
Time series synthesis
Longitude
Annual precipitation
Monthly precipitation
Historical daily precipitation
Latitude
Dataset splitting
Samples for precipitation event modeling
Spatial matching
Precipitation event modeling based on LGBM method
Precipitation event data
Precipitation modeling based on LGBM method
CHM_PRE V2 (Daily gridded precipitation)

Figure 3. The flowchart for dataset generation.

3.1 Data Preprocessing

Data preprocessing consists of two main components: gauge data preprocessing and gridded data preprocessing. Initially, we performed quality control on the CMA and GHCND gauge data and excluded stations outside the region of interest. The latitude and longitude range of primary interest to us in mainland China spans approximately 18°N to 54°N and 72°E to 136°E. However, to mitigate boundary effects, we have extended the area of interest outward by roughly 3°, defining it as 15°N to 57°N and 70°E to 140°E. The remaining stations were merged to serve as the gauge dataset for retrieval. Similarly, for various gridded datasets, data outside the region of interest were removed, and all data were resampled to a spatial resolution of 0.1°. Finally, the gridded data were converted to a daily time scale, resulting in the final gridded covariate data for retrieval.

3.2 Precipitation interpolation based on spatial correlation

Spatial correlation is the most significant characteristic of precipitation data, and the most common approach to constructing gridded precipitation datasets involves interpolation based on gauge data (Harris et al., 2020; He et al., 2020). Consequently, in this section, we also utilize gauge-based interpolation to characterize the spatial correlation of precipitation. The inverse distance weighting (IDW) method is widely used for interpolation due to its simplicity and computational efficiency. As a global interpolation method, IDW considers only the distance factor, applying inverse distance weighting to all stations for interpolation. However, the spatial correlation of many geographical features is often non-uniform. For example, many features may exhibit strong spatial correlation within a specific distance range, which rapidly diminishes beyond that range. To address this, Shepard (1968) introduced the concept of correlation decay distance (CDD) into interpolation and proposed the adaptive distance weighting (ADW) method. CDD measures how the spatial correlation between stations decreases with increasing distance and ensures that the search radius is set to an appropriate value, rather than using a fixed value for all situations (Dunn et al., 2020). Numerous datasets employ this method to interpolate gauge data to grids (Caesar et al., 2006; Dunn et al., 2020; Harris et al., 2020; Zhang et al., 2024a). Based on this, Han et al. (2023) incorporated CDD into the IDW method, and calculated the CDD values suitable for interpolating precipitation over mainland China. Given a target grid cell *G* surrounded by *n* known stations ***P*** $\{P_1, P_2, \ldots, P_n\}$, where the precipitation value at station P_i is z_i, the precipitation value at grid cell *G* is calculated as:

$$z(G) = \frac{\sum_{i=1}^{n} d(G, P_i)^{-p} z_i}{\sum_{i=1}^{n} d(G, P_i)^{-p}} \tag{1}$$

where $d(G, P_i)$ represents the distance between grid cell *G* and gauge station P_i, and *p* is the distance weighting exponent. In this study, *p* is set to 2, representing the Euclidean distance.

The selection of the station set ***P*** for interpolation markedly impacts the interpolated results. In this study, we adopt the improved IDW method and use the CDD values calculated by Han et al. (2023) for interpolating precipitation over mainland China (CDD1=244.7 km, CDD2=1336 km). When more than three stations are available within the CDD1range, CDD1 is used as the interpolation CDD; otherwise, CDD2 is applied. Meanwhile, if more than ten stations are available within the interpolation CDD range, only the ten closest stations to the grid cell are used, to mitigate overestimation of precipitation events in the densely populated station areas of eastern China.

Furthermore, previous research has demonstrated that interpolated precipitation anomalies (Harris et al., 2020; He et al., 2020) generally yield higher accuracy compared to direct precipitation interpolation. Thus, we adopt the interpolation scheme based on climatology anomaly rather than interpolating the raw precipitation values. To achieve this, daily and monthly precipitation climatology data were first generated. First, we calculated the average daily precipitation for 1971–

2000 to derive preliminary station-level daily climatology data by using the daily precipitation gauges from the previous step. The daily climatology series at each station was then processed by the Fourier truncation, retaining only the first six harmonic components to suppress high-frequency noise (Xie et al., 2007). Subsequently, the station-level daily climatology data were interpolated using the improved IDW, producing preliminary gridded daily climatology data. To enhance the reliability of the daily climatology data, we followed the same procedure to generate gridded monthly precipitation climatology data. These monthly data were used to correct the gridded daily climatology, yielding the adjusted gridded daily climatology data (Han et al., 2023). The precipitation anomalies were defined as the difference between the actual station precipitation and the adjusted gridded daily climatology data. Finally, the station-level daily precipitation anomalies were interpolated using the improved IDW method. The gridded daily precipitation data based on interpolation were final obtained by summing the interpolated anomalies with the adjusted gridded daily climatology data.

3.3 Precipitation retrieval based on spatiotemporal and physical correlations

Except spatial correlation, precipitation also exhibits physical and temporal correlations, which are often overlooked. In this study, we utilized the gridded precipitation data derived from gauge-based interpolation in Section 3.2, along with location information (longitude and latitude), to characterize precipitation's spatial correlations. Topographic data (elevation and slope), remote sensing-based precipitation data, data assimilation precipitation, soil moisture, and normalized difference vegetation index (NDVI) were employed to depict the precipitation's physical correlations. Meanwhile, historical daily precipitation and overall precipitation characteristics were used to capture temporal correlations. The details of the retrieval data can be found in **Figure 1**.

To obtain precipitation considered the three correlations comprehensively, we need to model the relationship between precipitation and these variables about the three correlations. This involves developing a response model with precipitation as the dependent variable and the relevant variables as independent variables. While linear regression models are the most commonly used response models, they are limited by their inability to capture nonlinear relationships and their relatively weak fitting capacity (Breiman, 2001; Chen and Guestrin, 2016; Yang et al., 2021). Machine learning-based models, in contrast, offer significant improvements in fitting performance and are more effective in representing nonlinear relationships (Guo et al., 2024; Hu et al., 2023). Among numerous machine learning-based models, Light gradient boosting machine (LGBM), developed by Microsoft (Ke et al., 2017), is renowned for its high precision and high generalizability. Consequently, we employed the LGBM method to integrate all these variables for precipitation retrieval, effectively accounting for the spatiotemporal and physical correlations of precipitation. Fundamentally, it employs a series of decision tree models for iterative training, progressively minimizing errors (or residuals) to ultimately generate predictions through a weighted summation. Unlike traditional gradient-boosted decision tree (GBDT) methods, LGBM utilizes a histogram-based technique for data binning, rather than processing each individual data record. This method iterates, calculates gains, and splits data accordingly (Zhang and Gong, 2020). Gradient-based one-side sampling (GOSS) is employed to sample the dataset, assigning greater weights to data points with larger gradients during gain computation. Under equivalent sampling rates, this method often outperforms random sampling (Candido et al., 2021). Owing to these features, LGBM demonstrates exceptional accuracy and generalization, making it widely applicable to various tasks such as classification, regression, and ranking (Bian et al., 2023; Hu et al., 2023; Jiang et al., 2024; Zhang et al., 2024b).

In the precipitation retrieval process, we employed a two-stage strategy: precipitation event retrieval and precipitation value retrieval. For the convenience of updating and maintaining data every year in the future, we constructed separate models for each year. Specifically, all variables required for retrieval in a given year were consolidated and split into training and validation sets in an 8:2 ratio. The training set was utilized to develop a precipitation event classification model based on LGBM method, while the validation set was used for hyperparameter optimization. Then, the established classification model was applied to all samples to determine whether each sample is a precipitation event. Samples identified as precipitation events were used for training the precipitation reversal model, while non-precipitation samples were excluded from the retrieval process. This approach effectively removed the majority of non-precipitation samples, simplifying the capture of precipitation characteristics and enhancing the accuracy of the reversal model. Additionally, this strategy notably improved the discrimination of precipitation events and mitigated the overestimation of precipitation events commonly associated with traditional interpolation-based methods. Upon completing the retrieval process, the trained models were used to generate final gridded daily precipitation for the entire mainland China from 1960 to 2023.

3.4 Validation

We compared the CHM_PRE V2 precipitation dataset with five existing gridded precipitation datasets to verify its high precision and reliability. To ensure comparability, the comparison focused on the period from 2001 to 2022 for which all data have time coverage. A total of 63,397 available CMA-HD station observations were utilized to validate the accuracy of precipitation data. To align with the 0.1° gridded precipitation data, station observations were mapped onto a 0.1° grid, and the average precipitation of all stations within each grid cell was regarded as the true precipitation value for that gird cell. Metrics such as absolute error (AE), Kling-Gupta efficiency (KGE, the values range is (-∞, 1], with 1 being the optimal), and relative standard deviation (RSD, the values range is (0, +∞), with 1 being the optimal) were employed to evaluate precipitation accuracy:

$$AE = abs(y - \hat{y}) \tag{2}$$

$$KGE = 1 - \sqrt{(R(y,\hat{y}) - 1)^2 + (RSD(y,\hat{y}) - 1)^2 + (Bias(y,\hat{y}) - 1)^2} \tag{3}$$

$$RSD = \frac{\sigma_{\hat{y}}/\mu_{\hat{y}}}{\sigma_y/\mu_y} \tag{4}$$

where y and $\hat{y}$ represent the observed precipitation values and the gridded precipitation values, respectively; μ denotes the mean value, σ signifies the standard deviation; R denotes the correlation coefficient, and Bias represents the variability ratio, each defined as follows:

$$R(y,\hat{y}) = \frac{\frac{1}{N}\sum_{i=1}^{N}\big(y_i - \mu(y)\big) * \big(\hat{y}_i - \mu(\hat{y})\big)}{\sigma_y \sigma_{\hat{y}}} \tag{5}$$

$$Bias = \frac{\mu_{\hat{y}}}{\mu_y} \tag{6}$$

Precipitation errors can be categorized into systematic errors, random errors, and precipitation event detection errors (Tian et al., 2009; Wei et al., 2022). Beyond precipitation amount (systematic errors and random errors), the occurrence of precipitation events also markedly impacts hydrological modelling (Dong et al., 2020). However, commonly used precipitation accuracy metrics such as KGE and RSD only account for systematic and random errors, neglecting the precipitation event detection errors. Thus, we adopted the Heidke skill score (HSS, the values range is (0, 1], with 1 being the optimal), false alarm ratio (FAR, the values range is [0, 1], with 0 being the optimal), and Accuracy score (the values range is (0, 1], with 1 being the optimal) to assess the accuracy of precipitation event detection (AghaKouchak and Mehran, 2013; Dong et al., 2020):

$$HSS = \frac{2(TP \times TN - FP \times FN)}{(TP + FN)(FN + TN) + (TP + FP)(FP + TN)} \tag{7}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$FAR = \frac{FP}{TP + FP}$$

where TP, FP, TN, and FN represent the precipitation events' matching relationship between gauged precipitation and precipitation products, with their meanings outlined in **Table S2** in the supplementary materials. The threshold for whether it is a precipitation event is more than 0.1mm of precipitation per day. Notably, to ensure the comparability of accuracy, instances where any precipitation products lack values were excluded during the accuracy calculations.

We assess the contribution of each variable to precipitation retrieval by analysing feature importance during the LGBM modelling process. LGBM, a gradient-boosted decision tree method, employs a greedy algorithm to select the optimal features for splitting based on specific criteria such as information gain or mean squared error during the construction of each decision tree. Features that are used more frequently to split nodes in the decision tree have a greater impact on the model, signifying higher importance (Breiman, 2001; Ke et al., 2017). Thus, we determine a variable's importance by the total number of times it is used for node splitting. To facilitate comparison among variables, we express feature importance as relative importance in percentage form.

4 Results and discussion

4.1 Precipitation amount and spatial patterns

Figure 4 illustrates the spatial distribution patterns of the multi-year average annual total precipitation for these datasets. It can be seen that all datasets exhibit similar precipitation distribution patterns, with annual totals generally decreasing from southeastern to northwestern China. Notably, CHM_PRE V2, CHM_PRE V1, GSMaP, and IMERG datasets effectively capture the high precipitation characteristics of the southern Tibetan Plateau, whereas PERSIANN-CDR and GLDAS datasets tend to underestimate precipitation in this region. Moreover, compared to satellite remote sensing-based datasets like GSMaP and IMERG, CHM_PRE V2 and CHM_PRE V1, which are based on extensive gauged observations, provide finer spatial pattern in precipitation distribution, particularly in regions with high variability, such as southeastern China.

Figure 5(a-b) depicts the temporal characteristics of precipitation across mainland China. The various datasets show highly consistent patterns in monthly average precipitation (**Figure 5**(a)) and multi-year monthly average precipitation (**Figure 5**(b)) across all grid cells. Precipitation is higher in spring and summer (March to August), peaking in July, and lower in autumn and winter (September to February).

To assess the contribution of different factors to precipitation retrieval we applied the established precipitation retrieval model to derive the feature importance of various variables. Based on this, we calculated the relative contribution of each variable (**Figure 5**(c)). Overall, spatial correlation contributed the most to precipitation retrieval (37.10%), followed by temporal correlation (34.11%) and physical correlation (28.78%). Specifically, interpolated precipitation is the most significant contributor, accounting for 24.44%—over three times the contribution of the second most important variable. This can be attributed to the strong alignment between interpolated precipitation, derived from gauged data, and actual precipitation. Latitude and longitude ranked third and fourth, with relative contributions of 6.61% and 6.06%, respectively, highlighting the pronounced regional differentiation in China's precipitation, consistent with the southeast-to-northwest decreasing trend shown in **Figure 4**. Among temporal variables, monthly total precipitation, precipitation from the previous day, and annual total precipitation had relative contributions of 7.81%, 5.07%, and 4.83%, respectively. These findings align with the seasonal variability of precipitation driven by China's monsoon climate, as evidenced in **Figure 5**(a-b). The importance of historical precipitation variables gradually diminished from one day to five days prior, reflecting the characteristic of precipitation time dependence gradually weakening over time. Within physical variables, remote sensing-based precipitation (5.86%), GLDAS soil moisture (5.58%), and NDVI (4.69%) emerged as the most influential. This may be attributed to the supplementary information provided by remote sensing and data assimilation technologies in the areas of lack of observation, especially western China (Jiang et al., 2023; Lyu et al., 2021).

In conclusion, the feature importance results demonstrate that the inclusion of temporal and physical correlations significantly enhances traditional gauge-based interpolated precipitation methods that consider only spatial correlations, thereby advancing the accuracy of precipitation retrieval.

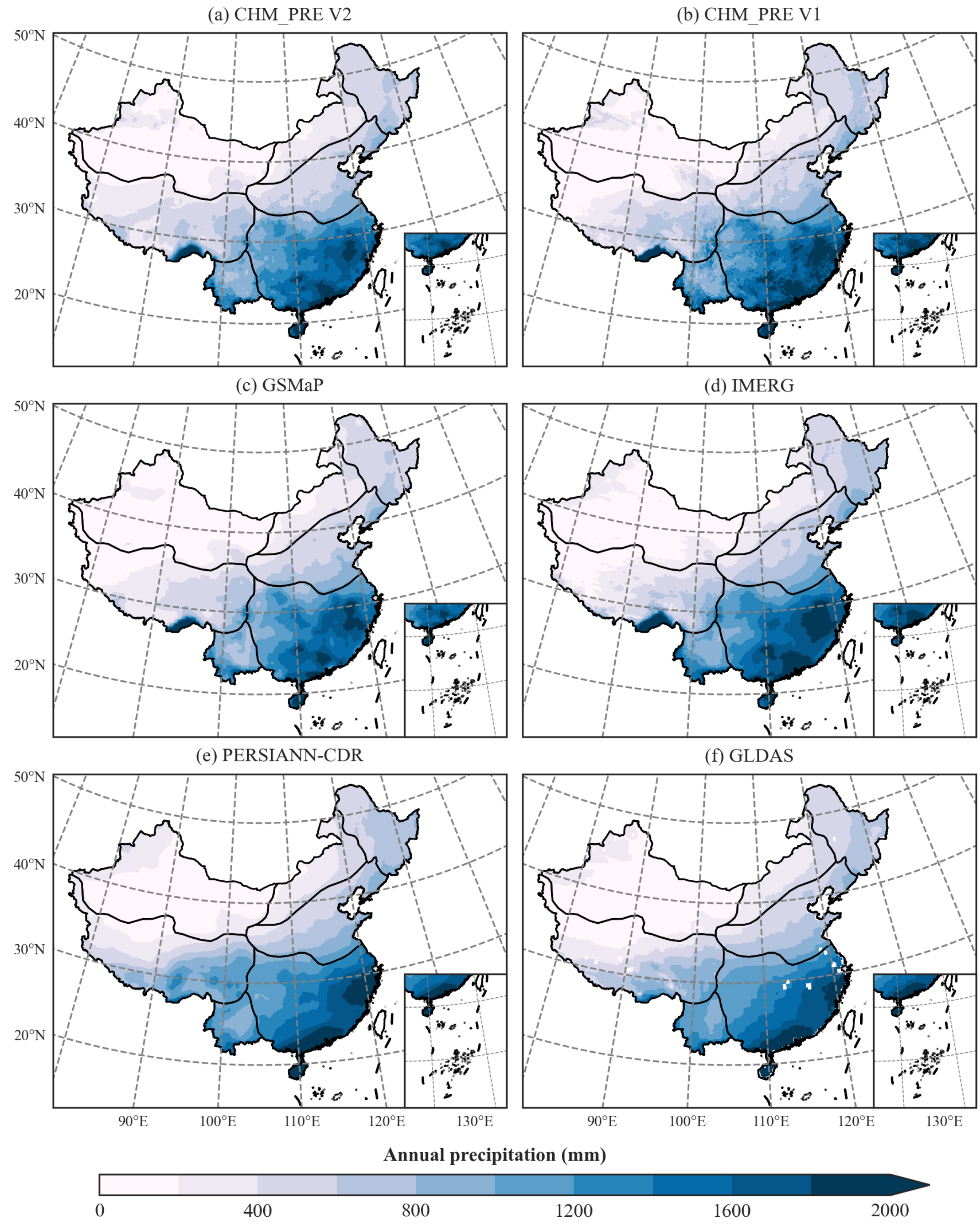


Figure 4. Spatial distribution patterns of multi-year average annual total precipitation from 2001 to 2020.

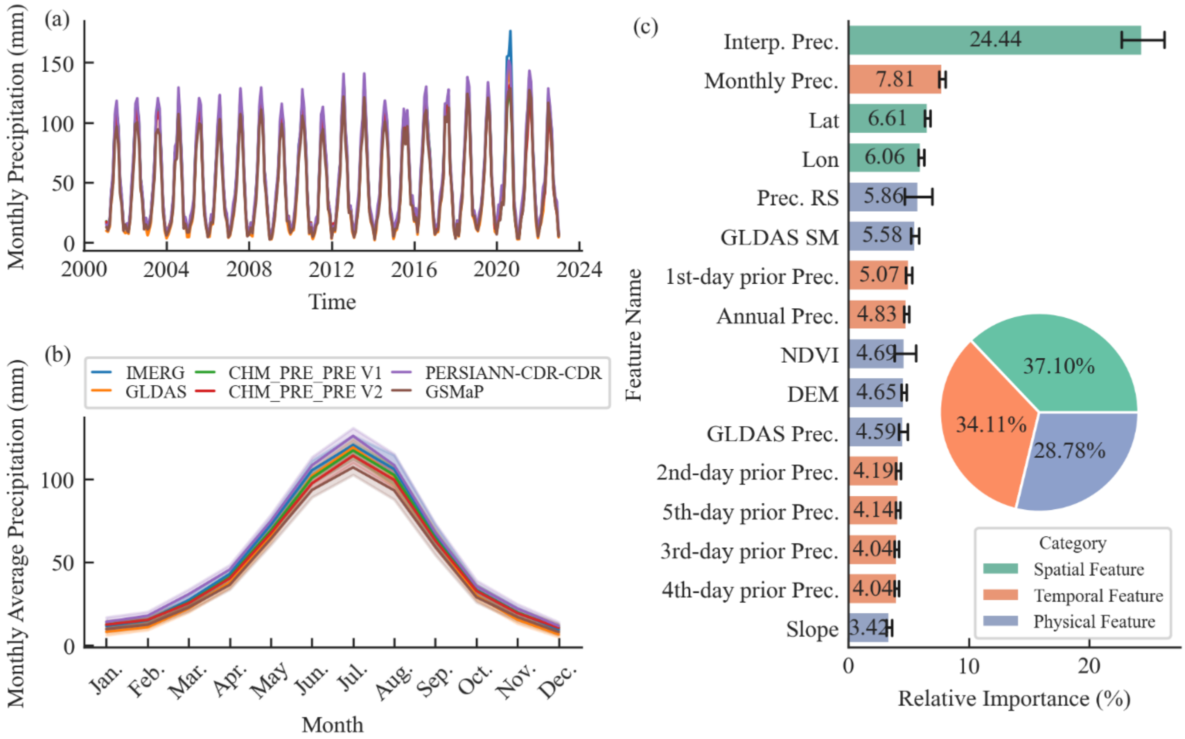


Figure 5. (a) time series of monthly precipitation; (b) multi-year mean monthly precipitation from 2001 to 2020; (c) feature importance of precipitation retrieval. In the figure, precipitation is abbreviated as "Prec.," interpolation-based precipitation is denoted as "Interp. Prec.," while remote sensing and soil moisture are represented by "RS" and "SM," respectively; "1st-day prior Prec." to "5th-day prior Prec." means the precipitation from the 1st day ago to 5th day ago.

4.2 Accuracy validation of precipitation value

Figure 6 illustrates the overall accuracy of these datasets based on CMA-HD. Precipitation datasets derived from gauge-based interpolation (CHM_PRE V1 and CHM_PRE V2) demonstrates significantly higher accuracy compared to those based on remote sensing (GSMaP, IMERG, and PERSIANN-CDR) and data assimilation (GLDAS), as evidenced by lower absolute error, higher KGE) and RSD (**Figure 6**(a-c)). CHM_PRE V2 achieved an overall MAE, KGE, and RSD of 1.48 mm/day, 0.79, and 0.88, respectively, outperforming other datasets by 12.84%, 12.86%, and 4.76% (**Table S3** in the supplementary material). Furthermore, the accuracy of precipitation datasets was analysed across different climatic regions. Given the superior performance of CHM_PRE V2 and CHM_PRE V1, the comparison focused exclusively on these two datasets. **Figure 6**(d-e) presents their absolute error, KGE, and RSD across different climatic regions. The results reveal a marked improvement in CHM_PRE V2's accuracy over CHM_PRE V1, with MAE increasing by 6.18% to 14.58% and KGE improving by 7.63% to 14.94% across various regions (**Figure 6**(d-e) and **Table S4** in the supplementary material).

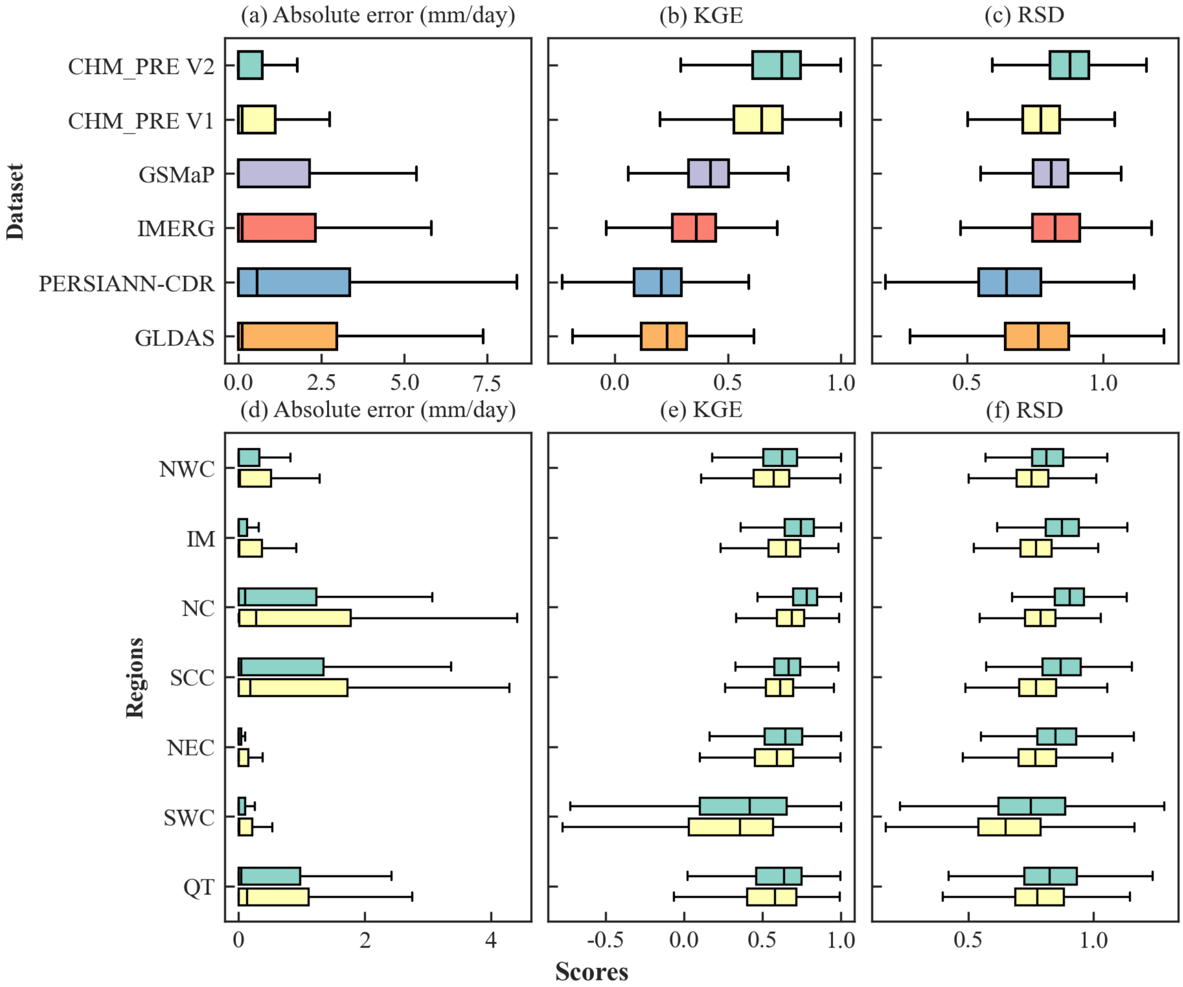
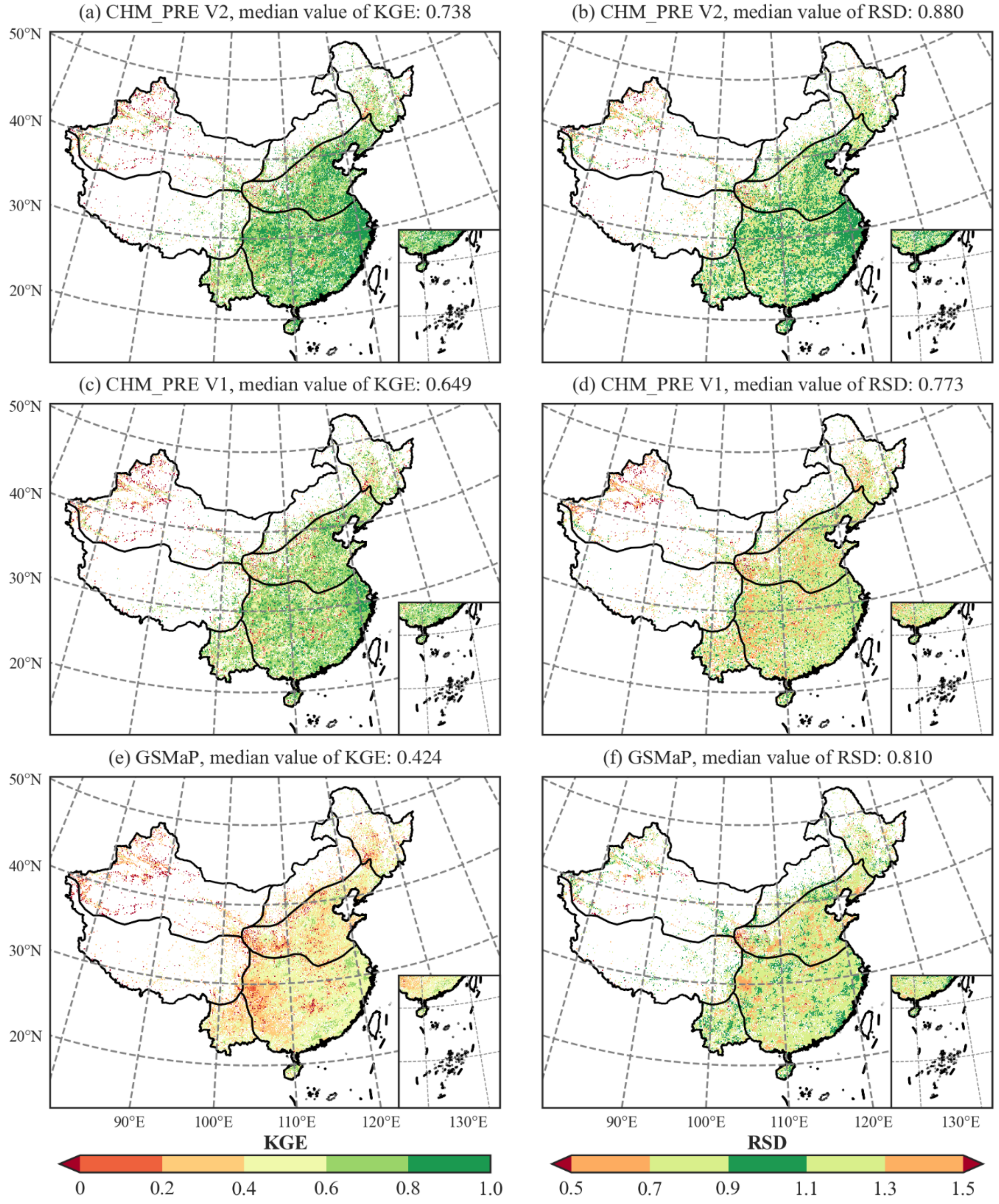


Figure 6. Accuracy of different precipitation datasets on the testing dataset CMA-HD. The green and yellow boxes in subfigures (d-f) represent CHM_PRE V2 and CHM_PRE V1, respectively.

Further comparison at the grid scale of the three precipitation datasets with relatively highest accuracy (CHM_PRE V2, CHM_PRE V1, and GSMaP) was conducted. **Figure 7** illustrates the spatial distribution of KGE and RSD for the three datasets. CHM_PRE V2 demonstrates a significant improvement in KGE compared to CHM_PRE V1 and GSMaP, with many grid cells in the NWC and IM regions showing an increase from below 0.2 to above 0.4, and numerous grid cells in the SCC and NC regions rising from the 0.6–0.8 range to above 0.8. The median KGE value of CHM_PRE V2 across all grid cells reaches 0.738, representing an approximate 13.87% improvement over CHM_PRE V1. Regarding RSD, GSMaP's accuracy slightly outperforms CHM_PRE V1; however, CHM_PRE V2 exhibits a distinct advantage over the other datasets, with a median RSD value of 0.880, reflecting an 8.64% enhancement compared to the other datasets.

https://doi.org/10.5194/essd-2025-20
Preprint. Discussion started: 12 February 2025


(a) CHM_PRE V2, median value of KGE: 0.738
(b) CHM_PRE V2, median value of RSD: 0.880
(c) CHM_PRE V1, median value of KGE: 0.649
(d) CHM_PRE V1, median value of RSD: 0.773
(e) GSMaP, median value of KGE: 0.424
(f) GSMaP, median value of RSD: 0.810
50°N
40°N
30°N
20°N
90°E
100°E
110°E
120°E
130°E
KGE
0
0.2
0.4
0.6
0.8
1.0
RSD
0.5
0.7
0.9
1.1
1.3
1.5

350 **Figure 7. Accuracy of different precipitation datasets at each grid cell on the testing data CMA-HD. (a), (c) and (e) show the KGE of each grid for CHM_PRE V2, PRE V1, and GSMaP, respectively; (b), (d) and (f) show the RSD of each grid for CHM_PRE V2, PRE V1, and GSMaP, respectively.**

4.3 Accuracy validation of precipitation event detection capability

Figure 8(a-c) illustrate the HSS, Accuracy score, and FAR metrics, evaluated using CMA-HD across different datasets. CHM_PRE V2 demonstrates a significantly superior ability to capture precipitation events across all three metrics compared to other precipitation datasets. Specifically, CHM_PRE V2 achieves an overall HSS of 0.68, an Accuracy score of 0.85, and a FAR of 0.24, surpassing other datasets by approximately 17.24%, 7.59%, and 29.17%, respectively (**Table S5** in the supplementary materials). Notably, a lower FAR value indicates better performance, with 0 being optimal, which distinguishes it from the other two metrics. Similarly, we analysed the precision of CHM_PRE V2 and CHM_PRE V1 in capturing precipitation events across different climatic regions. **Figure 8** (d-f) and **Table S6** in the supplementary materials reveal that CHM_PRE V2 consistently outperforms CHM_PRE V1 across all regions. The overall HSS values for CHM_PRE V2 in different regions reach 0.52–0.68, representing an improvement of approximately 10.16% to 22.98% over CHM_PRE V1. Further analysis of the FAR and probability of detection (POD) metrics shows that CHM_PRE V2 achieves improvements in FAR by 15.73% to 70.79% compared to CHM_PRE V1 across different climatic regions. However, the POD values for CHM_PRE V2 decrease by approximately 6.79% to 11.25% compared to CHM_PRE V1. This indicates that the improved accuracy of CHM_PRE V2 in capturing precipitation events is primarily due to a reduction in overestimation, attributable to the two-stage retrieval approach described in Section 3.3.

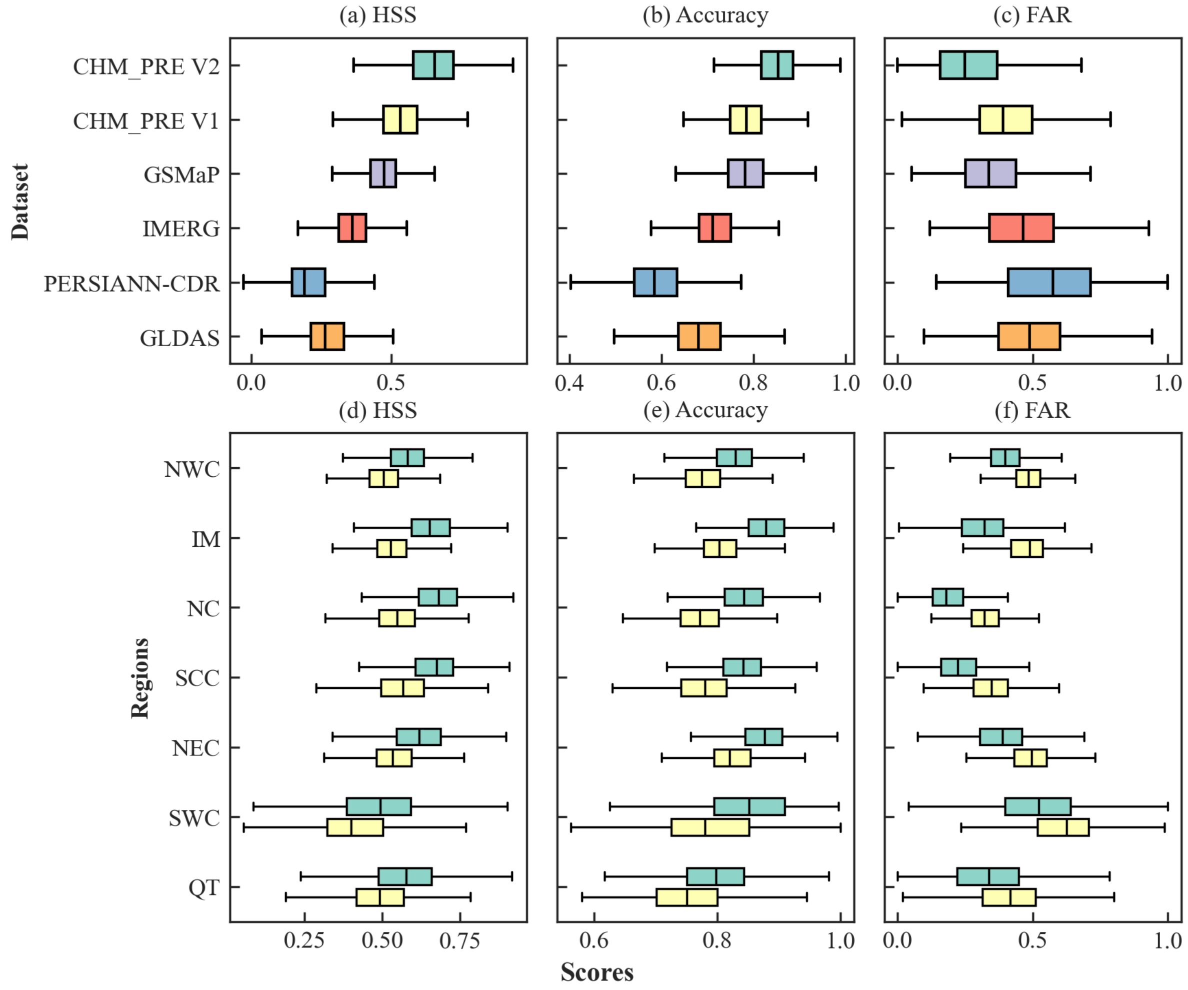
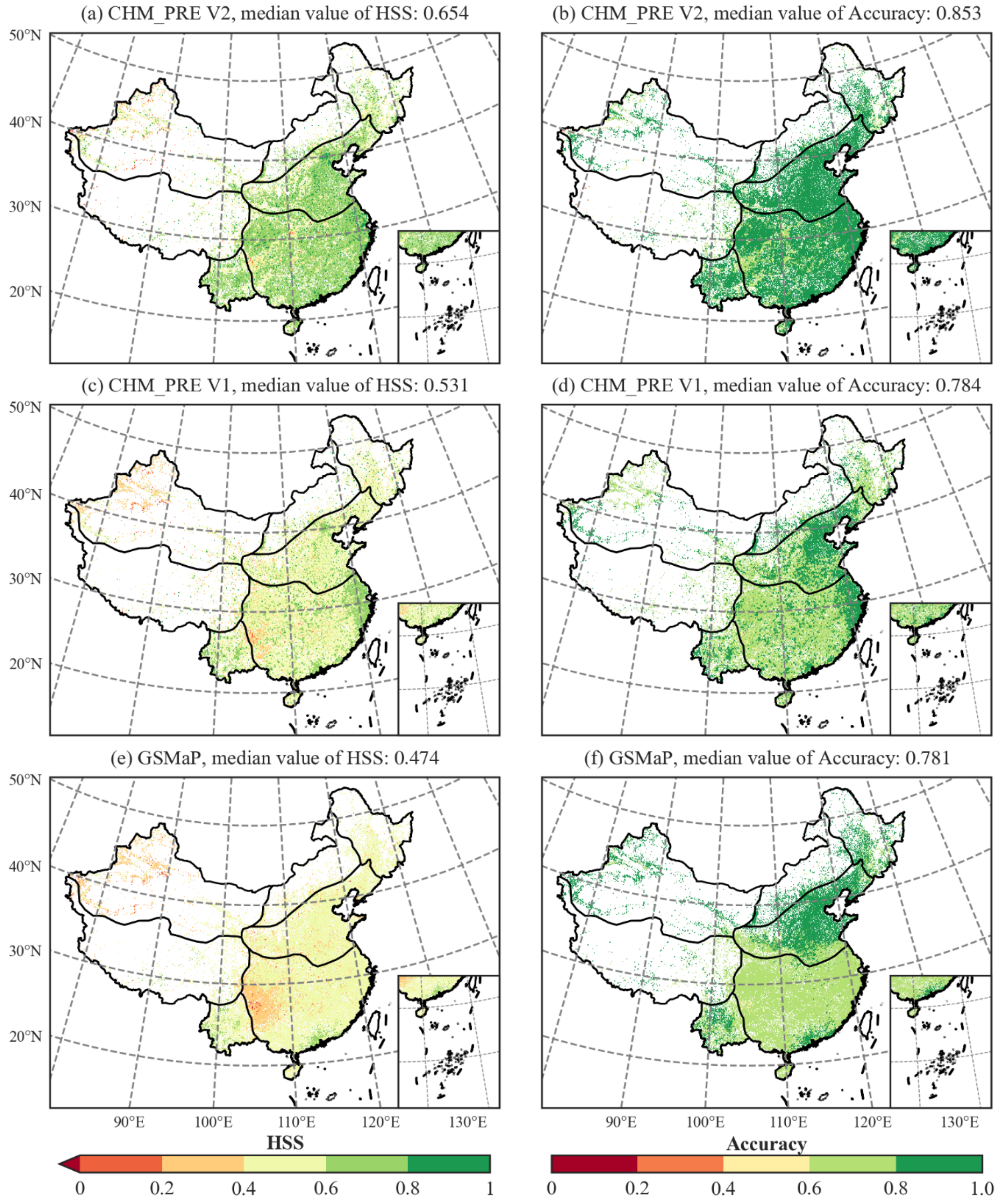


Figure 8. Accuracy precipitation events for different precipitation datasets on the testing dataset CMA-HD. The green and yellow boxes in subfigures (d-f) represent CHM_PRE V2 and CHM_PRE V1, respectively.

We further analyse the accuracy of precipitation events from the CHM_PRE V2, CHM_PRE V1, and GSMaP datasets across different grids. **Figure 9** illustrates the spatial distribution of the HSS and Accuracy scores for the three datasets. The KGE for CHM_PRE V2 shows a significant improvement over both CHM_PRE V1 and GSMaP, with the HSS values for many grid cells rising from 0.2–0.6 to 0.6–0.8. The total HSS across all grid cells reaches 0.654, representing a 23.16% improvement compared to other datasets. Regarding the Accuracy score, it is evident that GSMaP outperforms CHM_PRE V1 in regions such as NWC, NEC, and IM, while CHM_PRE V1 surpasses GSMaP in regions like SCC and NC. In contrast, CHM_PRE V2, which combines the advantages of interpolation-based and remote sensing-based precipitation data, outperforms all other datasets across all regions.


(a) CHM_PRE V2, median value of HSS: 0.654
(b) CHM_PRE V2, median value of Accuracy: 0.853
(c) CHM_PRE V1, median value of HSS: 0.531
(d) CHM_PRE V1, median value of Accuracy: 0.784
(e) GSMaP, median value of HSS: 0.474
(f) GSMaP, median value of Accuracy: 0.781
50°N
40°N
30°N
20°N
90°E
100°E
110°E
120°E
130°E
HSS
0
0.2
0.4
0.6
0.8
1
Accuracy
0
0.2
0.4
0.6
0.8
1.0

380 **Figure 9. Accuracy of precipitation events for different precipitation datasets at each grid cell on the testing data CMA-HD. (a), (c) and (e) show the HSS of each grid for CHM_PRE V2, PRE V1, and GSMaP, respectively; (b), (d) and (f) show the Accuracy score of each grid for CHM_PRE V2, PRE V1, and GSMaP, respectively.**

5 Data availability

The CHM_PRE V2 dataset provides daily precipitation data with a resolution of 0.1°, covering the entire mainland China (18°N–54°N, 72°E–136°E). This dataset covers the period of 1960–2023, and will be continuously updated annually. The daily precipitation data is provided in NetCDF format, and for the convenience of users, we also offer annual and monthly total precipitation data in both NetCDF and GeoTIFF formats. All of these data can be freely accessed at https://doi.org/10.5281/zenodo.14632157 (Hu and Miao, 2025).

6 Conclusions

This study utilizes precipitation data from long-term precipitation observations from a total of 3,746 stations and 11 related precipitation variables. By employing the improved inverse distance weighting method and the machine learning-based light gradient boosting machine (LGBM) algorithm, this study generated a new, high-precision daily gridded precipitation dataset for mainland China (CHM_PRE V2) from 1960 to 2023, with a spatial resolution of 0.1°, which accounts for both spatiotemporal and physical correlations. The CHM_PRE V2 was compared with five existing gridded precipitation datasets and validated for accuracy using precipitation data from over 63k automated rain gauge stations. The results demonstrate that CHM_PRE V2 aligns closely with the overall spatiotemporal distribution patterns of existing gridded precipitation datasets, but significantly outperforms them in terms of precipitation values and event accuracy. The overall mean absolute error and Kling-Gupta efficiency of CHM_PRE V2 reaches 1.48 mm/day and 0.88, respectively, surpassing other datasets by 12.84% and 12.86%. In terms of precipitation event capture, CHM_PRE V2 achieves an overall Heidke skill score, Accuracy Score, and false alarm ratio of 0.68, 0.85, and 0.24, respectively, outperforming other datasets by 17.24%, 7.59%, and 29.17%. Particularly in the precipitation-heavy regions of north China and central-south China, the false alarm ratio improvement reaches 53.33% and 68.42%, significantly reducing the overestimation of precipitation events. These findings prove that CHM_PRE V2 is a high-precision precipitation dataset, offering substantial support for various studies in hydrology, climatology, and climate change research.

Author contributions

JH and CM contributed to designing the research; JH implemented the research and wrote original draft; CM supervised the research; all co-authors revised the manuscript and contributed to the writing.

Competing interests

The contact author has declared that none of the authors has any competing interests.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (No. U24A20572), the National Key Research and Development Program of China (No. 2024YFF0809301) and the Fundamental Research Funds for the Central Universities. We would like to express our gratitude for the China Meteorological Administration (CMA, http://data.cma.cn) for providing the observed climate data of China. We also extend our thanks to NOAA, NASA, CGIAR-CSI, and JAXA for their valuable contribution of publicly available research data that supported the completion of this work.

References

Adler, R. F., Gu, G., Wang, J.-J., Huffman, G. J., Curtis, S., and Bolvin, D.: Relationships between global precipitation and surface temperature on interannual and longer timescales (1979–2006), Journal of Geophysical Research: Atmospheres, 113, https://doi.org/10.1029/2008JD010536, 2008.

AghaKouchak, A. and Mehran, A.: Extended contingency table: Performance metrics for satellite observations and climate model simulations, Water Resources Research, 49, 7144–7149, https://doi.org/10.1002/wrcr.20498, 2013.

Ahrens, B.: Distance in spatial interpolation of daily rain gauge data, Hydrology and Earth System Sciences, 10, 197–208, https://doi.org/10.5194/hess-10-197-2006, 2006.

Ashouri, H., Hsu, K.-L., Sorooshian, S., Braithwaite, D. K., Knapp, K. R., Cecil, L. D., Nelson, B. R., and Prat, O. P.: PERSIANN-CDR: Daily Precipitation Climate Data Record from Multisatellite Observations for Hydrological and Climate Studies, Bulletin of the American Meteorological Society, 96, 69–83, https://doi.org/10.1175/BAMS-D-13-00068.1, 2015.

Bian, L., Qin, X., Zhang, C., Guo, P., and Wu, H.: Application, interpretability and prediction of machine learning method combined with LSTM and LightGBM-a case study for runoff simulation in an arid area, Journal of Hydrology, 625, 130091, https://doi.org/10.1016/j.jhydrol.2023.130091, 2023.

Breiman, L.: Random Forests, Machine Learning, 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Caesar, J., Alexander, L., and Vose, R.: Large-scale changes in observed daily maximum and minimum temperatures: Creation and analysis of a new gridded data set, Journal of Geophysical Research: Atmospheres, 111, https://doi.org/10.1029/2005JD006280, 2006.

Candido, C., Blanco, A. C., Medina, J., Gubatanga, E., Santos, A., Ana, R. S., and Reyes, R. B.: Improving the consistency of multi-temporal land cover mapping of Laguna lake watershed using light gradient boosting machine (LightGBM) approach, change detection analysis, and Markov chain, Remote Sensing Applications: Society and Environment, 23, 100565, https://doi.org/10.1016/j.rsase.2021.100565, 2021.

Chen, D., Ou, T., Gong, L., Xu, C.-Y., Li, W., Ho, C.-H., and Qian, W.: Spatial interpolation of daily precipitation in China: 1951–2005, Advances in Atmospheric Sciences, 27, 1221–1232, https://doi.org/10.1007/s00376-010-9151-y, 2010.

Chen, D., Tian, Y., Yao, T., and Ou, T.: Satellite measurements reveal strong anisotropy in spatial coherence of climate variations over the Tibet Plateau, Scientific Reports, 6, 30304, https://doi.org/10.1038/srep30304, 2016.

Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, in: Kdd'16: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, New York, 785–794, https://doi.org/10.1145/2939672.2939785, 2016.

Dong, J., Crow, W. T., and Reichle, R.: Improving Rain/No-Rain Detection Skill by Merging Precipitation Estimates from Different Sources, Journal of Hydrometeorology, 21, 2419–2429, https://doi.org/10.1175/JHM-D-20-0097.1, 2020.

Dunn, R. J. H., Alexander, L. V., Donat, M. G., Zhang, X., Bador, M., Herold, N., Lippmann, T., Allan, R., Aguilar, E., Barry, A. A., Brunet, M., Caesar, J., Chagnaud, G., Cheng, V., Cinco, T., Durre, I., de Guzman, R., Htay, T. M., Wan Ibadullah, W. M., Bin Ibrahim, M. K. I., Khoshkam, M., Kruger, A., Kubota, H., Leng, T. W., Lim, G., Li-Sha, L., Marengo, J., Mbatha, S., McGree, S., Menne, M., de los Milagros Skansi, M., Ngwenya, S., Nkrumah, F., Oonariya, C., Pabon-Caicedo, J. D., Panthou, G., Pham, C., Rahimzadeh, F., Ramos, A., Salgado, E., Salinger, J., Sané, Y., Sopaheluwakan, A., Srivastava, A., Sun, Y., Timbal, B., Trachow, N., Trewin, B., van der Schrier, G., Vazquez-Aguirre, J., Vasquez, R., Villarroel, C., Vincent, L., Vischel, T., Vose, R., and Bin Hj Yussof, M. N.: Development of an Updated Global Land In Situ-Based Data Set of Temperature and Precipitation Extremes: HadEX3, Journal of Geophysical Research: Atmospheres, 125, e2019JD032263, https://doi.org/10.1029/2019JD032263, 2020.

Durre, I., Menne, M. J., and Vose, R. S.: Strategies for Evaluating Quality Assurance Procedures, Journal of Applied Meteorology and Climatology, 47, 1785–1791, https://doi.org/10.1175/2007JAMC1706.1, 2008.

Durre, I., Menne, M. J., Gleason, B. E., Houston, T. G., and Vose, R. S.: Comprehensive Automated Quality Assurance of Daily Surface Observations, Journal of Applied Meteorology and Climatology, 49, 1615–1633, https://doi.org/10.1175/2010JAMC2375.1, 2010.

Fan, C., Yin, S., and Chen, D.: Spatial correlations of daily precipitation over mainland China, International Journal of Climatology, 41, 6350–6365, https://doi.org/10.1002/joc.7199, 2021.

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov, A., Bosilovich, M. G., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty, A., Silva, A. M. da, Gu, W., Kim, G.-K., Koster, R., Lucchesi, R., Merkova, D., Nielsen, J. E., Partyka, G., Pawson, S., Putman, W., Rienecker, M., Schubert, S. D., Sienkiewicz, M., and Zhao, B.: The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2), Journal of Climate, 30, 5419–5454, https://doi.org/10.1175/JCLI-D-16-0758.1, 2017.

Guo, F., Ren, Y., Zhou, Y., Sun, S., Cui, M., and Khim, J.: Machine learning vs. statistical model for prediction modeling and experimental validation: Application in groundwater permeable reactive barrier width design, Journal of Hazardous Materials, 469, 133825, https://doi.org/10.1016/j.jhazmat.2024.133825, 2024.

Ham, Y.-G., Kim, J.-H., Min, S.-K., Kim, D., Li, T., Timmermann, A., and Stuecker, M. F.: Anthropogenic fingerprints in daily precipitation revealed by deep learning, Nature, 622, 301–307, https://doi.org/10.1038/s41586-023-06474-x, 2023.

Han, J., Miao, C., Gou, J., Zheng, H., Zhang, Q., and Guo, X.: A new daily gridded precipitation dataset for the Chinese mainland based on gauge observations, Earth System Science Data, 15, 3147–3161, https://doi.org/10.5194/essd-15-3147-2023, 2023.

Harris, I., Osborn, T. J., Jones, P., and Lister, D.: Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset, Scientific Data, 7, 109, https://doi.org/10.1038/s41597-020-0453-3, 2020.

He, J., Yang, K., Tang, W., Lu, H., Qin, J., Chen, Y., and Li, X.: The first high-resolution meteorological forcing dataset for land process studies over China, Scientific Data, 7, 25, https://doi.org/10.1038/s41597-020-0369-y, 2020.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Hu, J. and Miao, C.: CHM_PRE V2: A new upgraded high-precision gridded precipitation dataset considering spatiotemporal and physical correlations over China (V2.0), https://doi.org/10.5281/zenodo.14632157, 2025.

Hu, J., Miao, C., Zhang, X., and Kong, D.: Retrieval of suspended sediment concentrations using remote sensing and machine learning methods: A case study of the lower Yellow River, Journal of Hydrology, 627, 130369, https://doi.org/10.1016/j.jhydrol.2023.130369, 2023.

Huff, F. A. and Shipp, W. L.: Spatial Correlations of Storm, Monthly and Seasonal Precipitation, Journal of Applied Meteorology and Climatology, 8, 542–550, 1969.

Huffman, G. J., Bolvin, D. T., Nelkin, E. J., Wolff, D. B., Adler, R. F., Gu, G., Hong, Y., Bowman, K. P., and Stocker, E. F.: The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-Global, Multiyear, Combined-Sensor Precipitation Estimates at Fine Scales, Journal of Hydrometeorology, 8, 38–55, https://doi.org/10.1175/JHM560.1, 2007.

Huffman, G. J., Bolvin, D. T., Braithwaite, D., Hsu, K., Joyce, R., Xie, P., and Yoo, S.-H.: NASA global precipitation measurement (GPM) integrated multi-satellite retrievals for GPM (IMERG), Algorithm theoretical basis document (ATBD) version, 4, 2020–05, 2015.

Huffman, G. J., Stocker, E. F., Bolvin, E. J., and Nelkin, J. T.: GPM IMERG Final Precipitation L3 1 day 0.1 degree x 0.1 degree V07, Edited by Andrey Savtchenko, Greenbelt, MD, Goddard Earth Sciences Data and Information Services Center (GES DISC), https://doi.org/10.5067/GPM/IMERGDF/DAY/07, 2023.

Jiang, S., Tarasova, L., Yu, G., and Zscheischler, J.: Compounding effects in flood drivers challenge estimates of extreme river floods, Science Advances, 10, eadl4005, https://doi.org/10.1126/sciadv.adl4005, 2024.

Jiang, Y., Yang, K., Qi, Y., Zhou, X., He, J., Lu, H., Li, X., Chen, Y., Li, X., Zhou, B., Mamtimin, A., Shao, C., Ma, X., Tian, J., and Zhou, J.: TPHiPr: a long-term (1979–2020) high-accuracy precipitation dataset (1 / 30°, daily) for the

Third Pole region based on high-resolution atmospheric modeling and dense observations, Earth System Science Data, 15, 621–638, https://doi.org/10.5194/essd-15-621-2023, 2023.

Kang, X., Dong, J., Crow, W. T., Wei, L., and Zhang, H.: The Conditional Bias of Extreme Precipitation in Multi-Source Merged Data Sets, Geophysical Research Letters, 51, e2024GL111378, https://doi.org/10.1029/2024GL111378, 2024.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y.: LightGBM: A Highly Efficient Gradient Boosting Decision Tree, in: Advances in Neural Information Processing Systems, 2017.

Kubota, T., Aonashi, K., Ushio, T., Shige, S., Takayabu, Y. N., Kachi, M., Arai, Y., Tashima, T., Masaki, T., Kawamoto, N., Mega, T., Yamamoto, M. K., Hamada, A., Yamaji, M., Liu, G., and Oki, R.: Global Satellite Mapping of Precipitation (GSMaP) Products in the GPM Era, in: Satellite Precipitation Measurement, vol. 67, edited by: Levizzani, V., Kidd, C., Kirschbaum, D. B., Kummerow, C. D., Nakamura, K., and Turk, F. J., Springer, Cham, 355–373, https://doi.org/10.1007/978-3-030-24568-9_20, 2020.

Li, R., Wang, K., and 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, Journal of Geophysical Research: Atmospheres, 123, 10423–10442, https://doi.org/10.1029/2018JD028991, 2018.

Lyu, F., Tang, G., Behrangi, A., Wang, T., Tan, X., Ma, Z., and Xiong, W.: Precipitation Merging Based on the Triple Collocation Method Across Mainland China, IEEE Transactions on Geoscience and Remote Sensing, 59, 3161–3176, https://doi.org/10.1109/TGRS.2020.3008033, 2021.

Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E., and Houston, T. G.: An Overview of the Global Historical Climatology Network-Daily Database, Journal of Atmospheric and Oceanic Technology, 29, 897–910, https://doi.org/10.1175/JTECH-D-11-00103.1, 2012.

Qin, R., Zhao, Z., Xu, J., Ye, J.-S., Li, F.-M., and Zhang, F.: HRLT: a high-resolution (1d, 1km) and long-term (1961–2019) gridded dataset for surface temperature and precipitation across China, Earth System Science Data, 14, 4793–4810, https://doi.org/10.5194/essd-14-4793-2022, 2022.

Qiu, H., Zhou, T., Chen, X., Wu, B., and Jiang, J.: Understanding the Diversity of CMIP6 Models in the Projection of Precipitation Over Tibetan Plateau, Geophysical Research Letters, 51, e2023GL106553, https://doi.org/10.1029/2023GL106553, 2024.

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., Prudden, R., Mandhane, A., Clark, A., Brock, A., Simonyan, K., Hadsell, R., Robinson, N., Clancy, E., Arribas, A., and Mohamed, S.: Skilful precipitation nowcasting using deep generative models of radar, Nature, 597, 672–677, https://doi.org/10.1038/s41586-021-03854-z, 2021.

Ren, M., Zhang, Y., and Sheng, B.: An Outline of China's Physical Geography, Foreign Languages Press, Beijing, 1985.

Rodell, M., Houser, P. R., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C.-J., Arsenault, K., Cosgrove, B., Radakovich, J., Bosilovich, M., Entin, J. K., Walker, J. P., Lohmann, D., and Toll, D.: The Global Land Data Assimilation System, Bulletin of the American Meteorological Society, 85, 381–394, https://doi.org/10.1175/BAMS-85-3-381, 2004.

Shen, Y., Feng, M., Zhang, H., and Gao, F.: Interpolation Methods of China Daily Precipitation Data, Journal of Applied Meteorological Science (In Chinese), 21, 279–286, 2010.

Shepard, D.: A two-dimensional interpolation function for irregularly-spaced data, in: Proceedings of the 1968 23rd ACM national conference, Citation Key: shepard1968two, 517–524, 1968.

Sun, Q., Miao, C., Duan, Q., Ashouri, H., Sorooshian, S., and Hsu, K.: A Review of Global Precipitation Data Sets: Data Sources, Estimation, and Intercomparisons, Reviews of Geophysics, 56, 79–107, https://doi.org/10.1002/2017RG000574, 2018.

Sun, Q., Zhang, X., Zwiers, F., Westra, S., and Alexander, L. V.: A Global, Continental, and Regional Analysis of Changes in Extreme Precipitation, Journal of Climate, 34, 243–258, https://doi.org/10.1175/JCLI-D-19-0892.1, 2021.

Tian, Y., Peters-Lidard, C. D., Eylander, J. B., Joyce, R. J., Huffman, G. J., Adler, R. F., Hsu, K., Turk, F. J., Garcia, M., and Zeng, J.: Component analysis of errors in satellite-based precipitation estimates, Journal of Geophysical Research: Atmospheres, 114, D24101, https://doi.org/10.1029/2009JD011949, 2009.

Trucco, A., Barla, A., Bozzano, R., Pensieri, S., Verri, A., and Solarna, D.: Introducing Temporal Correlation in Rainfall and Wind Prediction From Underwater Noise, IEEE Journal of Oceanic Engineering, 48, 349–364, https://doi.org/10.1109/JOE.2022.3223406, 2023.

Vermote, E. and NOAA CDR Program: NOAA Climate Data Record (CDR) of AVHRR Normalized Difference Vegetation Index (NDVI) (5), https://doi.org/10.7289/V5ZG6QH9, 2019.

Wei, G., Lü, H., Crow, W. T., Zhu, Y., Su, J., and Ren, L.: Comprehensive Evaluation and Error-Component Analysis of Four Satellite-Based Precipitation Estimates against Gauged Rainfall over Mainland China, Advances in Meteorology, 2022, 9070970, https://doi.org/10.1155/2022/9070970, 2022.

Wu, J. and Gao, X.: A gridded daily observation dataset over China region and comparison with the other datasets, Chinese Journal of Geophysics (in Chinese), 56, 1102–1111, https://doi.org/10.6038/cjg20130406, 2013.

Xie, P., Chen, M., Yang, S., Yatagai, A., Hayasaka, T., Fukushima, Y., and Liu, C.: A Gauge-Based Analysis of Daily Precipitation over East Asia, Journal of Hydrometeorology, 8, 607–626, https://doi.org/10.1175/JHM583.1, 2007.

Xiong, J., Guo, S., Abhishek, Yin, J., Xu, C., Wang, J., and Guo, J.: Variation and attribution of probable maximum precipitation of China using a high-resolution dataset in a changing climate, Hydrology and Earth System Sciences, 28, 1873–1895, https://doi.org/10.5194/hess-28-1873-2024, 2024.

Yang, Y., Huang, T. T., Shi, Y. Z., Wendroth, O., and Liu, B. Y.: Comparing the Performance of an Autoregressive State-Space Approach to the Linear Regression and Artificial Neural Network for Streamflow Estimation, Journal of Environmental Informatics, 37, 36–48, https://doi.org/10.3808/jei.202000440, 2021.

Zhang, D. and Gong, Y.: The comparison of LightGBM and XGBoost coupling factor analysis and prediagnosis of acute liver failure, IEEE Access, 8, 220990–221003, https://doi.org/10.1109/ACCESS.2020.3042848, 2020.

Zhang, Q., Miao, C., Su, J., Gou, J., Hu, J., Zhao, X., and Xu, Y.: A New High-Resolution Multi-Drought Indices Dataset for Mainland China, Earth System Science Data Discussions, 1–29, https://doi.org/10.5194/essd-2024-270, 2024a.

Zhang, X., Zwiers, F. W., Li, G., Wan, H., and Cannon, A. J.: Complexity in estimating past and future extreme short-duration rainfall, Nature Geoscience, 10, 255–259, https://doi.org/10.1038/ngeo2911, 2017.

Zhang, Y., Ren, Y., Ren, G., and Wang, G.: Precipitation Trends Over Mainland China From 1961–2016 After Removal of Measurement Biases, Journal of Geophysical Research: Atmospheres, 125, e2019JD031728, https://doi.org/10.1029/2019JD031728, 2020.

Zhang, Y., Feng, X., Zhou, C., Sun, C., Leng, X., and Fu, B.: Aridity threshold of ecological restoration mitigated atmospheric drought via land–atmosphere coupling in drylands, Commun Earth Environ, 5, 1–11, https://doi.org/10.1038/s43247-024-01555-9, 2024b.