A two-step merging strategy for incorporating multi-source precipitation products and gauge observations using machine learning classification and regression over China

Huajin Lei [1], Hongyu Zhao [2], Tianqi Ao [1]

[1] State Key Laboratory of Hydraulics and Mountain River Engineering, College of Water Resource and Hydropower, Sichuan University, Chengdu 610065, China;
[2] State Key Laboratory of Earth Surface Processes and Resource Ecology, Beijing Normal University, Beijing 100875, China;

Correspondence to: Tianqi Ao（aotianqi@scu.edu.cn）

Abstract. Although many multi-source precipitation products (MSPs) with high spatio-temporal resolution have been extensively used in water cycle research, they are still subject to considerable uncertainties due to the spatial variability of terrain. Effective detection of precipitation occurrence is the key to enhancing precipitation accuracy. This study presents a two-step merging strategy to incorporate MSPs (GSMaP, IMERG, TMPA, PERSIANN-CDR, CMORPH, CHIRPS, and ERA-Interim) and rain gauges to improve the precipitation capture capacity and precipitation intensity simultaneously during 2000-2017 over China. Multiple environment variables and the spatial autocorrelation between precipitation observations are selected as auxiliary variables in the merging process. Three machine learning (ML) classification and regression models, including gradient boosting decision tree (GBDT), extreme gradient boosting (XGBoost), and random forest (RF), are adopted and compared. The strategy first employs classification models to identify wet and dry days in warm and cold seasons, then combines regression models to predict precipitation amounts based on wet days. The results are also compared with those of traditional methods, including multiple linear regression (MLR), ML regression models, and gauge-based Kriging interpolation. A total of 1680 (70%) rain gauges are randomly chosen for model training and 692 (30%) for performance evaluation. The results show that: (1) The multi-sources merged precipitation products (MSMPs) perform better than original MSPs in detecting precipitation occurrence under different intensities, followed by Kriging. The average Heidke Skill Score (HSS) of MSPs, Kriging, and MSMPs is 0.30-0.69, 0.71, 0.79-0.8, respectively. (2) The proposed method significantly alleviates the bias and deviation of original MSPs in temporal and spatial. The MSMPs strongly correlate with gauge observations with the CC of 0.85. Moreover, the modified Kling-Gupta efficiency (KGE) improves by 17%-62% (MSMPs: 0.74-0.76) compared with MSPs (0.34-0.65). (3) The spatial autocorrelation factor (KP) is the most important variable in models, which contributes considerably to improving the model accuracy. (4) The proposed method outperforms MLR and ML regression models, and XGBoost algorithm is more recommended for large-scale data merging owing to its high computational efficiency. This study provides a robust and reliable method to improve the performance of precipitation data with full consideration of multi-source information. This method could be applied globally and produce large-scale precipitation products if rain gauges are available.

1   Introduction

As one of the critical parameters of the natural water cycle, precipitation helps us realistically understand the interaction

35   between hydrological and climate systems. Meanwhile, precipitation monitoring is essential for extreme hydroclimatic disaster

forecasting and water resources management (Yilmaz et al., 2005; Tao et al., 2016; Xu et al., 2018). Accurate precipitation

estimates are of practical importance for social economy and security, agriculture, meteorology, ecology, and other fields

(Awange et al., 2019). Traditional rain gauge measurements can provide reliable precipitation data. However, restricted by

gauge spatial distribution, it heavily relies on geographical conditions and only reflects the precipitation characteristics within

40   a limited radius around the instruments (Collischonn et al., 2008; Jia et al., 2011). In many developing countries, the

distribution of rain gauges is inadequate and irregular, with scarce or even unavailable gauges in mountainous and arid/semi-

arid areas (Jiang et al., 2012). This phenomenon is particularly prominent in Tibetan Plateau, where this study is covered and

where precipitation has significant spatiotemporal variability. Hence, precipitation observations are insufficient to meet the

needs of hydrometeorological applications in remote areas (Bhuiyan et al., 2017; Woldemeskel et al., 2013). In contrast,

45   satellite-based precipitation estimates and atmospheric reanalysis are promising alternative tools for describing large-scale and

long-term precipitation spatial distribution due to their high spatio-temporal resolution.

Up to the present, a series of advanced remote sensing techniques and numerical weather models have been employed to

retrieve various multi-source precipitation products (MSPs) (Huffman et al., 2007; Joyce et al., 2004). For example, the

Tropical Rainfall Measuring Mission (TRMM) algorithm combines detection information from multiple sensors (including

50   the microwave imager, infrared radiometer, and radar) to provide valuable precipitation information for tropical and subtropical

regions (Huffman et al., 2007). The Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) (Funk et al.,

2015) incorporates infrared cold cloud duration observations and satellite information to prepare a long time and high spatial

resolution (0.05°) dataset, which contributes to global climate change trend analysis. The Precipitation Estimation from

Remotely Sensed Information using Artificial Neural Networks (PERSIANN) applies a state-of-the-art algorithm to generate

55   global precipitation based on the geostationary longwave infrared imagery (Hsu et al., 1997). As an extension of TRMM, the

Integrated Multi-satellitE Retrievals for GPM (IMERG) algorithm enhances the estimation efficiency of solid and light

precipitation, which has finer temporal resolution and wider spatial coverage than TRMM (Huffman et al., 2019). In addition,

one of the most popular atmospheric reanalysis products, ERA-Interim provides a variety of climate variables including

precipitation of atmospheric, oceanic, and land. Nevertheless, these MSPs usually suffer from inherent uncertainty caused by

60   retrieval algorithms, complex terrain, and limitation sensors resampling frequency (Nerini et al., 2015; Arshad et al., 2021).

This uncertainty tends to be more severe at shorter time scales (such as sub-daily and daily) and varies among different

precipitation products (Lei et al., 2021). Therefore, how to alleviate the errors of MSPs is a key priority step to improve their

application efficiency (Jiang et al., 2012; Sharifi et al., 2016; Lu et al., 2019).

Hydrology and
Earth System
Sciences
Discussions

An important means to improve the accuracy of MSPs is to merge multi-source products and gauge-based precipitation

65　information. It can fully integrate the advantages of high precision of rain gauges and high temporal-spatial resolution of MSPs. In recent years, numerous methods have been proposed and applied to derive more satisfactory precipitation. For instance, Chao et al. (2018) combined CMORPH precipitation with daily gauge-based precipitation in the Ziwuhe Basin of China by using geographically weighted regression. Xie and Xiong (2011) removed the bias of CMORPH over China based on gauge analysis through probability density function and optimal interpolation (OI). Shen et al. (2014) and Wu et al. (2018) also

70　employed similar approaches to reduce the systematic bias and random errors in products. In addition, Yang et al. (2017) proposed an inverse-root-mean-square-error weighting approach to combine PERSIANN-Cloud Classification System (PERSIANN-CCS) estimates with gauge observations over Chile. Massari et al. (2019) developed a Bayesian inversion approach to merge the TRMM with the soil moisture-based rainfall product SM2RAIN-CCI in Itay. The simple model averaging, one-outlier removed schemes (Ma et al., 2017), geographical ratio analysis (GRA), and geographical difference

75　analysis (GDA) technologies (Arshad et al., 2021) are also widely used and proved to be effective in improving MSPs accuracy. However, the aforementioned approaches are strongly based on solid mathematical assumptions and are unlikely to consistently achieve the desired results in any region and time scale in reality (Wu et al., 2020). Most importantly, many environmental variables associated with precipitation process are not taken into account.

　　　The rapid development of machine learning (ML) technology provides us with a new way to enhance MSPs' performance.

80　Compared to traditional mathematical approaches, ML can deal with complex nonlinear relationships without constructing explicit statistical models. Moreover, the strength of ML comes from its ability to solve different types of problems, from classification to regression and prediction, as well as its efficiency in learning and generalizing big data (He et al., 2016). Those features lead ML particularly suitable for precipitation/no-precipitation recognition and long-time series data merging. Tao et al. (2015) presented a deep neural network (DNN) scheme by integrating infrared imagery to reduce the bias and the false

85　alarm of PERSIANN-CCS over the central United States. Kumar et al. (2019) employed a nonlinear support vector machine (SVR) model to combine TRMM with soil moisture. Wu et al. (2020) exploited a deep fusion model by coupling the convolutional neural network (CNN) with a long-short-term memory network (LSTM) to merge TRMM and rain gauge data from 2001 to 2005 over China. Recently, Baez-Villanueva et al. (2020) presented an RF-based procedure to combine multiple satellite precipitation products (SPPs) with gauge-based data and geographical features, particularly Euclidean distances to

90　gauge locations in Chile. Hong et al. (2021) considered the correlation weights for MSPs and combined them with environmental variables using artificial neural networks over the Tibetan Plateau. Zhang et al. (2021) proposed a novel double machine learning approach to merge SPPs (near-real-time IMERG, GSMaP, and PERSIANN with ASCAT) with gauge observations and proved reliable over China. However, Zhou et al. (2020) demonstrated that gauge-corrected IMERG and GSMaP significantly outperform near-real-time ones over China. The reliability of input precipitation could determine the

95   quality of merging results to a certain extent. Similarly, a higher density rain gauge network is expected to provide sufficient

training and validation samples to acquire a more robust prediction model (Tang et al., 2018; Herrera et al., 2019). Furthermore,

most studies mainly considered limited environmental information and spatial correlation (such as Euclidean distance,

coordinates, inverse distance weighted (IDW), etc.) while neglecting the spatial autocorrelation between gauge observations

in merging processes. Moreover, the majority of previous researches directly improved the magnitude of precipitation without

100   paying particular attention to the hit and false ratio of products. Correctly judging whether precipitation occurs is the key to

enhancing precipitation performance fundamentally. To the best of our knowledge, the application and comparison of GBDT,

XGBoost, and RF models (both classification and regression) in precipitation merging have not been well explored yet.

This study proposes a two-step merging strategy to simultaneously enhance the precipitation discrimination ability and

precipitation intensity over China by incorporating MSPs and relatively high-density rain gauges based on ML algorithms.

105   The objectives of this study include three-folds: (1) assessing and comparing the effectiveness of GBDT, XGBoost, and RF in

precipitation merging by considering multiple covariates and spatial autocorrelation between gauges; (2) quantitatively

analyzing the contribution of each covariable and MSP in merging processes; (3) further comparing the performance of the

proposed strategy with the ML regression, multiple linear regression (MLR) and Kriging interpolation methods. The two-step

merging strategy is expected to provide a straightforward and effective method for preparing high precision precipitation data

110   over China.

2   Study area and Materials

2.1 Study area

China between 73°-135°E and 15°-53°N is selected as the study area, which is located in eastern Asia and west of the

Pacific Ocean with a land area of 9.6 million km^2 (Fig. 1). The elevation of China gradually increases from southeast to

115   northwest, resulting in the complexity of topography and the diversity of vegetation. The climate is characterized by cold and

dry in the northwest while warm and humid in the southeast. Due to the complex topography, annual precipitation over China

has high spatial variability, varying between 50 mm and 2000 mm from west to east. Meanwhile, the distribution of

precipitation throughout the year is also extremely uneven. Much more precipitation (70% - 80%) occurs during the warm

season (May to October) than during the cold season (November to April), which is the primary factor for this study to conduct
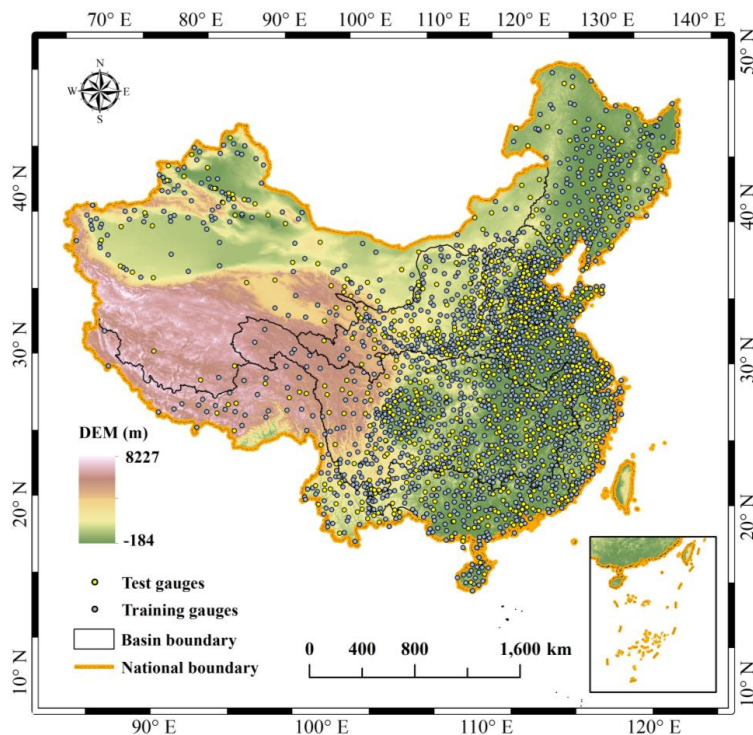
120   model training according to different seasons.

Figure 1: The topography of China and distribution of rain gauges.

2.2 Materials

2.2.1 Rain gauge observations

125   A relatively dense network of 2372 rain gauges over mainland China from 2000 to 2017 is collected in this study, provided by China Meteorological Administration (CMA). The daily precipitation data have been conducted strictly quality control by CMA and can be used after simple processing, such as removing missing values and converting units. It should be noted that there is a temporal mismatch (12h) between daily gauge-based precipitation (Beijing Time from 20:00 to 20:00, UTC + 8:00) and MSPs (UTC, from 00:00 to 24:00). Considering that not all products have a sub-daily scale temporal resolution, we

130   recalculate daily observations using sub-daily precipitation (i.e., 8:00 to 20:00 and 20:00 to 8:00) to keep consistent with MSPs. Gauges are mainly distributed in eastern but sparsely located in western China, especially in the hinterland of Qinghai-Tibet Plateau (TP) (as shown in Fig. 1). The gauge density used in this study is higher than in some previous studies (Wu et al., 2020; Yin et al., 2020; Zhang et al., 2021). The average control area for a single gauge is approximately 4000 km^2 (9.6×10^6 $km^2/2372$). Nevertheless, it is far from meeting the requirement of the World Climate Organization that the control area should be about

135   600 km^2 for plain and even smaller for arid/semi-arid regions (WMO, 1965).

2.2.2 MSPs

Seven MSPs, including a reanalysis product and six satellite precipitation products retrieved from multiple sensors, are

5

selected to integrate. Limited by the availability of MSPs, the period of this study is from June 2000 to December 2017 (hereafter: 2000-2017). Specific information about MSPs is summarized in Table 1.

140   TRMM was launched on 27 November 1997 by the National Aeronautics and Space Administration (NASA) in collaboration with the Japan Aerospace Exploration Agency (JAXA). The TRMM algorithm incorporates the multi-source information from TRMM Microwave Imager (TMI), Visible and Infrared Radiometer (VIRS), and radar instruments. The TRMM Multisatellite Precipitation Analysis (TMPA) 3B42 can be classified into two categories: Real-Time (3B42RT) and gauged-corrected (3B42) products. The latter outperforms the former because it combines the Global Precipitation Climatology

145   Centre (GPCC) gauge observations. The latest version 7 TMPA 3B42 product is therefore chosen in this study. In addition, IMERG provides three types of products, including Early, Late, and Final Run products, which are retrieved around 4h, 12h, and 4month, respectively, after satellite monitoring. The latest version 6 IMERG Final run product has been corrected according to GPCC and selected in this study. Moreover, the version 7 Global Satellite Mapping of Precipitation (GSMaP) GSMaP_Gauge applied in this study incorporates Climate Prediction Center (CPC) gauge data analysis (Kubota et al., 2007),

150   which is more accurate than other GSMaP products such as GSMaP near-real-time (NRT).

PERSIANN-Climate Data Record (PERCDR) has a long record from 1983 to the present. The PERSIANN algorithm is mainly based on Gridded Satellite (GridSat-BI) IR data and National Centers for Environmental Prediction (NCEP) Stage IV radar data (Ashouri et al., 2015), which does not fuse microwave information. The reliability of PERCDR is improved by using GPCC for calibration. CHIRPS v.2 product is also used in this study. It has higher spatial resolution than other MSPs,

155   integrating satellite imagery, global climatology, and gauge observations. In addition, Climate Prediction Center Morphing Technique (CMORPH) version 1 dataset (Joyce et al., 2004) covers three categories' products: CMORPH RAW, CMORPH bias-corrected (CRT), and CMORPH gauge blended datasets (BLD). CMORPH CRT is selected in this study due to its superior quality.

ERA-Interim (herein ERAI) is a global atmospheric reanalysis data generated by the European Centre for Medium-Range

160   Weather Forecasts (ECMWF). It is produced using cycle 31r2 of ECMWF's Integrated Forecast System (IFS), including a 4-dimensional variational analysis (4D-var) with a 12-hour analysis window. The 4D-Var assimilation technique is the core component of the ERA-Interim forecast system, which updates the parameters that determine bias correction required for most satellite radiation observations (Dee et al., 2011). ERAI replaces ERA-40 with numerous improvements such as stratospheric circulation quality and the time consistency of the reanalysis field.

165   The information sources employed in MSPs show significant differences, especially whether microwave signals are incorporated or not (Table 1). Moreover, various algorithms are adopted to retrieve precipitation in different MSPs. For instance, the Goddard Profiling Algorithm 2014 is used for IMERG, the histogram matching method is employed for TMPA, and the morphing technique is applied for CMORPH (Table 1). Each algorithm and signal source has its cons and pros. It is necessary

to combine them to maximize their advantages. Although several products already combine gauge observation data to reduce

170   bias, only a few gauges within China are used. Given the relatively high gauge density used in this study, this has little impact

on the independence of gauges and the reliability of results (Shen et al., 2013).

Table 1. The information about seven MSPs used in this study

MSPs	Temporal-spatial resolution	Spatial coverage	Input sources	Retrieval algorithm
GSMaP	1h, 0.1°	60°S-60°N	PMW, IR and Gauge	Kalman filtering technique
IMERG	0.5h, 0.1°	60°S-60°N	PMW, IR and Gauge	Goddard Profiling Algorithm
TMPA	1h, 0.25°	50°S-50°N	PMW, IR and Gauge	Histogram matching method
PERCDR	3h, 0.25°	60°S-60°N	IR and Gauge	adaptive ANN
CHIRPS	daily, 0.05°	50°S-50°N	IR, Gauge, and reanalysis	Kalman filter model
CMORPH	3h, 0.25°	60°S-60°N	PMW, IR and Gauge	Morphing technique
ERAI	3h,0.25°	Global	Reanalysis and Gauge	IFS Cy31r2 4D-Var

2.2.3 Environment variables

The environment variables used in this study include DEM, longitude, latitude, wind speed, relative humidity, soil

175   moisture, cloud cover, and air temperature.

DEM is downloaded from the Shuttle Radar Topographic Mission (SRTM) with a resolution of 90 m. Wind speed, relative

humidity, soil moisture, and air temperature are obtained from NASA Global Land Data Assimilation System Noah Land

Surface Model (GLDAS_NOAH), with 3 h and 0.25° resolutions (Redell., 2004). Cloud cover is collected from ERA-interim

developed by the ECMWF, with the resolution of 3 hourly and 0.25°. Although Normalized Differential Vegetation Index

180   (NDVI) is often used as a critical auxiliary variable to predict precipitation, it is susceptible to soil type and human activities.

NDVI is more suitable for applying on the monthly or annual scale due to its temporal resolution (Ghorbanpour et al., 2021;

Shen et al., 2021; Tan et al., 2021). Inversely, the response of air temperature and soil moisture to daily precipitation is better

than NDVI, especially in the desert and bare land (Bhuiyan et al., 2018). In addition, the interactions between cloud properties

and precipitation are equally important (Sharifi et al., 2019).

185   **3   Methodology**

3.1 Data preprocessing

To maintain the temporal and spatial consistency of data used in this study, all MSPs and environment variables at a sub-

daily scale are aggregated to daily data. The spatial resolution of DEM (90m) and CHIRPS (0.05°) are upscaled to 0.1°, the

TMPA, PERCDR, CMORPH, ERAI, cloud cover, and GLDAS_NOAH are downscaled to 0.1° using the bilinear interpolation

190   method. In this study, the gauges are divided into two groups, 70% of rain gauges are spatially and randomly selected as

training and calibrating samples and the remaining 30% as validation samples. Due to the irregular distribution of rain gauges

over China, random sampling is carried out for each river basin to ensure the spatial representativeness of the validation gauges.

Inspired by previous researches (Baez-Villanueva et al., 2020; Zhang et al., 2020), we consider a covariate describing spatial autocorrelation between rain gauges in this study. The semivariogram based on Ordinary Kriging is adopted to calculate

195 spatial autocorrelation factor, i.e., Kriging_based prediction (KP). Compared with other predict models, such as Inverse distance interpolation (IDW), Kriging_based semivariogram considers not only the spatial relationship between predicted and neighboring known points but considers the statistical autocorrelation between known points. The Ordinary Kriging assumes the model:

$$z^*(x_0) = \sum_{i=1}^{n} \lambda_i z(x_i),$$ (1)

200 Where $z^*(x_0)$ is the predicted value of the unknown x_0 point. $z(x_i)$ and λ_i are the known value of neighboring rain gauge x_i and its weight. Unbiasedness and minimum estimation variance are the conditions for choosing weights. The weight depends on the distance between the known points, the predicted position, and the overall spatial arrangement based on the known points. Spatial autocorrelation must be quantified before spatial arrangement can be applied in weights. The calculation processes of KP are as follows:

205 (1) Calculate the distance and semivariogram between known points;

$$\gamma(h) = \frac{1}{2}[z(x_i) - z(x_j)],$$ (2)

Where $\gamma(h)$ is the semivariogram of x_i and x_j, h is the distance, z is the value of known of points.

(2) A theoretical model is used to fit semivariogram and distances. The nugget, sill, and range can be obtained according to the fitted semivariogram. The commonly used semivariogram models are spherical, exponential, Gaussian, and Linear

210 models. The spherical model is selected in this study:

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ C_0 + C\left(\frac{3}{2} \cdot \frac{b}{a} - \frac{1}{2} \cdot \frac{b^3}{a^3}\right) & 0 < h \le a \\ C_0 + C & h > 0 \end{cases},$$ (3)

Where $\gamma(h)$ is semivariogram, h is the distance, C_0, C, and a is the nugget, sill, and range, respectively.

(3) Calculate the semivariogram between the unknown point and known points, and form a matrix to solve the weights:

$$\begin{bmatrix} \gamma(h_{11}) & \cdots & \gamma(h_{1n}) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \gamma(h_{n1}) & \cdots & \gamma(h_{nn}) & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \mu \end{bmatrix} = \begin{bmatrix} \gamma(h_{10}) \\ \vdots \\ \lambda_n \\ \gamma(h_{n0}) \end{bmatrix},$$ (4)

215 Where μ is Lagrange parameter.

(4) Predict the value of the unknown point using eq. (1) according to the weights obtained from eq. (4).

3.2 A two-step merging strategy

The two-step merging strategy for multiple MSPs is illustrated in Fig.2. The ML models containing random forest (RF), gradient boosting decision tree (GBDT), and extreme gradient boosting (XGBoost) are chosen to incorporate seven MSPs

220   (GSMaP, IMERG, TMPA, PERCDR, CMORPH, CHIRPS, and ERAI) and rain gauges. Although the RF method has been

extensively employed in most previous studies, few studies compared it with GBDT and XGBoost models in precipitation

merging. The environment variables, including soil moisture, cloud cover, relative humidity, air temperature, DEM, longitude,

latitude, and spatial autocorrelation (KP) are selected as auxiliary variables (i.e., covariate) of the merging strategy. The values

of multiple covariables and MSPs extracted according to gauge locations are taken as independent variables, while gauge

225   observations are taken as the dependent variable. Meanwhile, according to the annual distribution characteristics of

precipitation, we group all input datasets into two seasons: warm season (May and October) and cold season (November to

April), and models are trained and evaluated in each season.

The two-step merging strategy explored in this study can be generally described in two stages (Fig. 2). Firstly, the gauge

observations are classified into wet and dry days to be used as the benchmark for classification models. Considering the

230   measurement accuracy of China's standard rain gauges is 0.1mm, the daily precipitation less than 0.1 is set as a dry day, greater

than or equal to 0.1 is a wet day. The RF, GBDT, and XGBoost classification models are adopted to improve the precipitation

identification ability of MSPs based on the classified gauge observations. Then, the regression models are employed to predict

the precipitation amounts of wet days. The final multi-source merged precipitation products (MSMPs) combine dry days

(precipitation value is 0) identified by classification models with precipitation intensity of wet days corrected by regression

235   models. The period of model training, evaluation, and precipitation interpolation are from 2000 to 2017 at the daily scale. A

total of 1680 gauges (Fig.1) are used for training and calibration models; 692 gauges are used for validating models'

performance. To highlight the superiority of the two-step merging strategy, we compare it with single ML regression, multiple

linear regression (MLR), and gauge-based Kriging interpolation methods. Meanwhile, the best-performing algorithm is

selected by intercomparing the three ML models in the two-step merging strategy. The detailed merging algorithms are

240   introduced in 3.2.1-3.2.4.
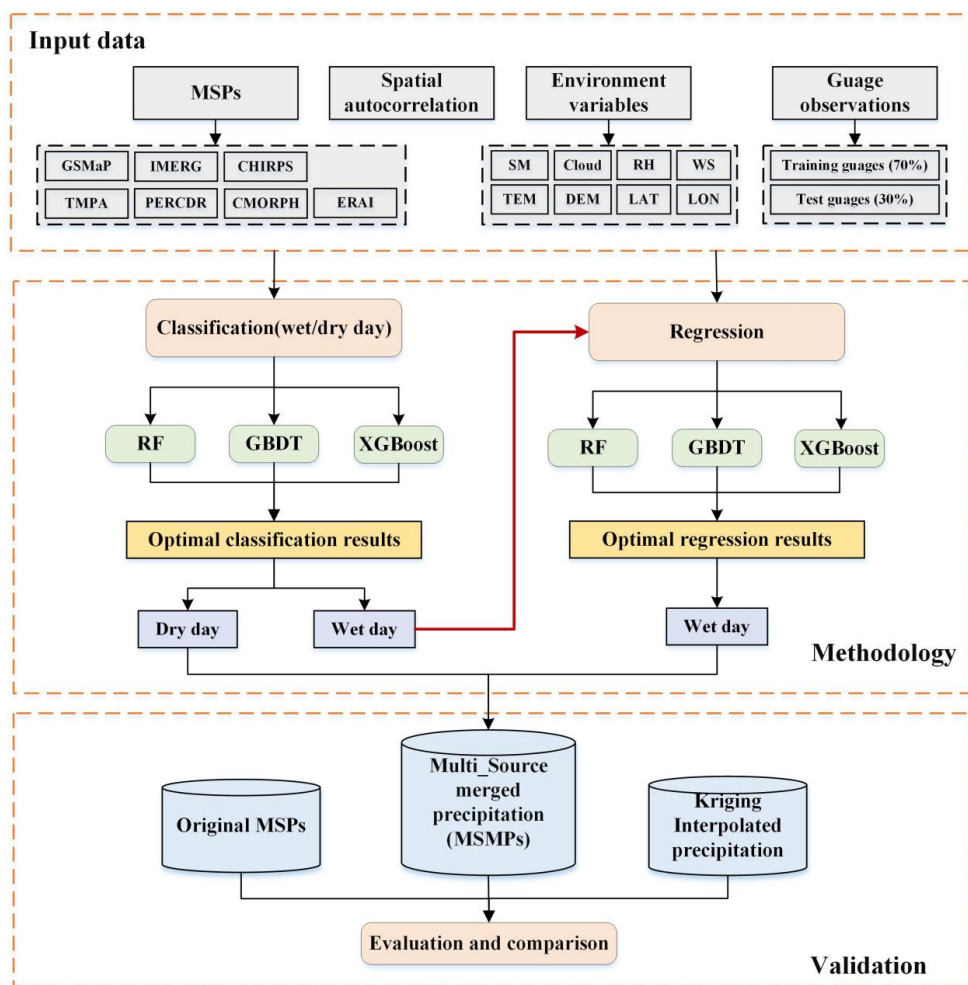
Hydrology and
Earth System
Sciences
Discussions

Figure 2. The flowchart of merging strategy of this study (LAT is latitude, LON is longitude, RH is relative humidity, SM is soil moisture, TEM is temperature, and WS is wind speed).

3.2.1 RF

245   The RF model was proposed by Breiman (2001) and is widely applied to deal with regression, classification, and other tasks (Rodriguez-Galiano et al., 2012; Nguyen et al., 2021). The general structure of RF is shown in Fig. 3. RF is an ensemble learning algorithm composed of multiple decision trees and generally outperforms a single tree. For regression problems, the model returns predictions by averaging all individual decision trees. For classification problems, each tree in the forest is judged and classified separately, and the output of RF is the class of a majority vote on classification trees (Ho, 1998).

250   The Bootstrap Aggregation (i.e., Bagging) technique is applied by the RF training algorithm for tree learners, which is designed to improve the accuracy and stability of ML algorithms in classification and regression processes. The Bagging algorithm utilizes the out-of-bag (OOB) error to measure the prediction error of RF. It creates two independent datasets. One

dataset, the Bootstrap sample (approximately two-thirds of all samples), is selected as "in-the-bag" data through sampling and

replacement, while the remaining out-of-bag dataset (one-third) that is not selected during the sampling process is used to

255   calculate the model's OOB error (Breiman, 2001). The advantages of RF can be mainly summarized in four points: (1)

processing high-dimensional data (a mass of features) without dimensionality reduction and feature selection; (2) measuring

the importance of features and how they interact with each other; (3) avoiding overfitting and easy to implement; (4) balancing

errors for asymmetric datasets, which is critical in the cold season when wet and dry days are unevenly distributed. In addition,

several important parameters in RF are the number of decision trees (n_estimators), the maximum depth of each decision tree

260   (max_depth), and the minimum number of samples required to split an internal node (min_samples_split). The optimal

parameters vary according to the characteristics of training samples, which are therefore different in warm and cold seasons.
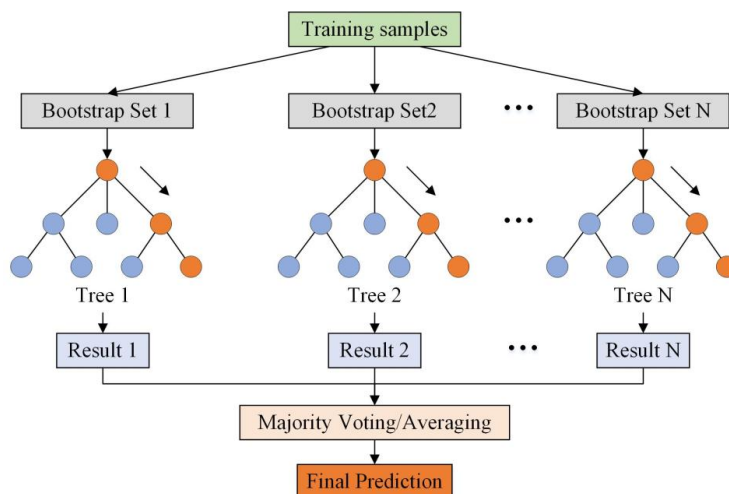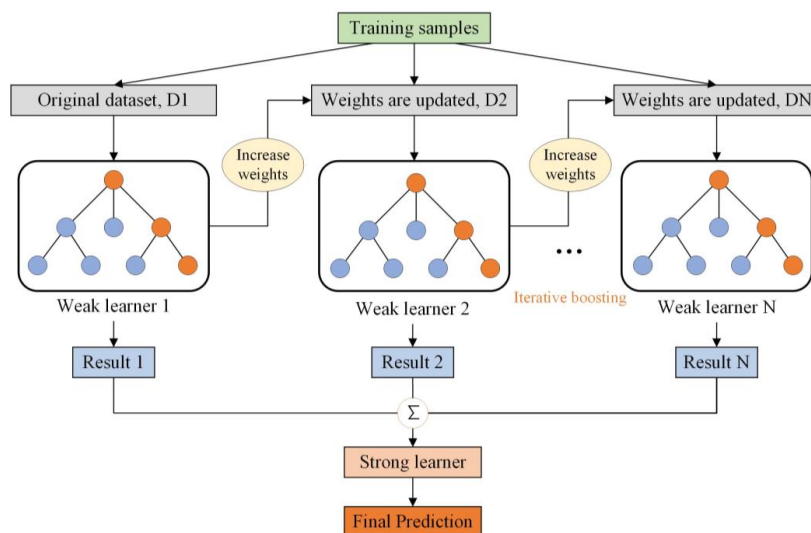

Figure 3. The overview structure of RF.

3.2.2 GBDT

265   The GBDT is an iterative decision tree model created by Breiman (1997) and subsequently developed by Friedman (2002),

which is also called the multiple additive regression tree (MART) (shown in Fig. 4). The additive algorithm is utilized for

classification or regression to reduce residuals generated in the training process continuously. GBDT uses the forward

distribution algorithm and selects the classification and regression tree (CART) learner as a weak base learner. GBDT generates

numerous weak learners through multiple iterations, and each learner is trained based on the residual of the previous learner.

270   It finally integrates the multiple weak learners into a single strong learner by weighting the summation of each tree.

The main difference between RF and GBDT is that RF can be trained in parallel to reduce variances, while GBDT reduces

the biases by fitting the residual of former trees. Due to the strong connection between weak learners, GBDT is difficult to be

paralleled. Generally speaking, GBDT has superior generalization ability and robustness, which is less affected by training

samples size and can deal with various data flexibly, including outliers and irrelevant features. Moreover, the prediction

275 accuracy of GBDT is high in the case of relatively little parameter adjustment time. The main parameters of GBDT include

the number of boosting stages to perform (n_estimators), the learning rate shrinks the contribution of each tree by learning_rate

(learning rate,) and the maximum depth of trees (max_depth). The n_estimators and learning rate are highly correlated with

the performance of the model.

280 **Figure4. The overview structure of GBDT.**

3.2.3 XGBoost

The XGBoost model was proposed by Chen and Guestrin (2016) based on the structure of GBDT. XGBoost also combines

multiple weak learners into a strong one, and the base learner in XGBoost can be either CART or linear classifier. XGBoost

possesses the strength of GBDT and has several additional improvements: First, GBDT only uses the first-order derivative

285 information in optimization, while XGBoost performs second-order Taylor expansion on the cost function to obtain the first-

order and second-order derivatives, thus acquiring more accurate loss functions. Second, XGBoost introduces a regularization

term into the cost function to effectively control the complexity of the model. From the perspective of bias-variance tradeoff,

it reduces the variance of the model, making the learned model more straightforward and preventing over-fitting. Third,

XGBoost allows users to define custom optimization goals and evaluation criteria, increasing its flexibility. Moreover,

290 XGBoost implements parallel processing when selecting the best split node for enumeration, significantly improving the

computational efficiency compared with Gradient Boosting Machine (GBM). The key parameters of XGBoost are n_estimators,

learning rate, max_depth, and scale_pos_weight. The default value of scale_pos_weight is 1, indicating the positive and

negative samples are in equilibrium. This is not applicable for precipitation classification in the cold season. More attention

should be paid to scale_pos_weight when model training. In addition, due to the large sample size used in this study

295 (approximately 14 million pieces of data) and limited computing resources, we use the trial-and-error method to conduct model

Hydrology and
Earth System
Sciences
Discussions

parameter optimization.

3.2.4 MLR

The MLR is the first type of regression algorithm used extensively in many fields, assuming a stable linear relationship between a dependent variable and multiple independent variables. Compared with nonlinear relationships, the MLR is easier

300   to fit and each explanatory variable's statistical property is more intuitive. MLR is usually fitted using the ordinary least square method to minimize the sum of squares of residuals predicted by the model and observed by the sample. The overall model for MLR is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_i X_i, \quad i = 1, \dots, n, \tag{5}$$

Where n is the number of explanatory variables, Y is the dependent variable predicted by $X_1, X_2 \dots, X_n$. β_0 is the intercept, and

305    $\beta_1, \beta_2 \dots, \beta_i$ are regression coefficients.

3.3 Performance evaluation and comparison

In this study, the performance and accuracy of all products are evaluated using randomly selected independent daily rain gauges from 2000 to 2017. The evaluation metrics mainly involve categorical and statistical metrics. The categorical metrics focus on analyzing the ability of products to capture precipitation events, including the probability of detection (POD), false

310   alarm ratio (FAR), critical success index (CSI), Precision (*precision*), frequency bias (FB), Heidke Skill Score (HSS), and classification accuracy (*Accuracy*). The POD, also called hit bias, represents the probability of precipitation events correctly detected. FAR and *precision* describes the ratio of falsely and correctly detected events among total detected precipitation events, respectively. The sum of FAR and *precision* is 1. The CSI incorporates POD and FAR, which demonstrates the overall ability of precipitation detection. The FB is the ratio of POD and FAR. It shows the balanced ability of products in detecting

315   precipitation events. FB < 1 indicates that precipitation events are underestimated, FB > 1 indicates overestimated. The FB equals 1 meaning that the number of missed events is equal to false alarmed events. HSS compares the predicted performance with random chance. The negative HSS shows random chance is better than the model predicted. The range of HSS is -∞ to 1, the perfect value is 1.

$$POD = \frac{H}{H+M}, \tag{6}$$

320    $$FAR = \frac{F}{H+F}, \tag{7}$$

$$Precision = \frac{H}{H+F}, \tag{8}$$

$$CSI = \frac{H}{H+M+F}, \tag{9}$$

$$FB = \frac{POD}{Precision} = \frac{H+F}{H+M} \ , \tag{10}$$

$$HSS = \frac{2(HN-FM)}{(H+M)\cdot(M+N)+(H+F)\cdot(F+N)} \ , \tag{11}$$

The *Accuracy* shows the proportion of total days that are correctly classified as wet and dry days. One point that needs to be emphasized is that this study takes *Accuracy* as the evaluation metric to describe the accuracy of ML classification models (RF, GBDT, and XGBoost) in training processes, thereby determining the optimal parameters of the model.

$$Accuracy = \frac{H+N}{H+M+F+N} \times 100\%, \tag{12}$$

Where H is the total number of precipitation events simultaneously observed and predicted, M is the total number of precipitation events observed but not predicted, F is the total number of precipitation events predicted but not detected, N is the total number of no-precipitation events. The optimal value of POD, *precision*, CSI, *Accuracy*, and FB is 1, while FAR is 0.

The statistical metrics include Pearson correlation coefficient (CC), root mean square error (RMSE), the modified Kling-Gupta efficiency (KGE) and its components (β and γ). The CC measures the magnitude of the correlation between the model predicted and observed values. The RMSE accesses the error between predicted and observed values. The KGE combining the CC, bias (β), and variability ratio (γ) reflects the overall goodness of fit between model predicted and observed. $\beta > 1$ indicates precipitation amount is overestimated and vice versa. The formulas for these metrics are expressed as follows:

$$CC = \frac{\sum_{i=1}^{n}(P_{oi}-\overline{P_o})(P_{mi}-\overline{P_m})}{\sqrt{\sum_{i=1}^{n}(P_{oi}-\overline{P_o})^2 \cdot (P_{mi}-\overline{P_m})^2}}, \tag{13}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_{mi}-P_{oi})^2}, \tag{14}$$

$$KGE = 1 - \sqrt{(CC-1)^2+(\beta-1)^2+(\gamma-1)^2} \ , \tag{15}$$

$$\beta = \frac{\mu_m}{\mu_o} \ , \tag{16}$$

$$\gamma = \frac{SD_m/\mu_m}{SD_o/\mu_o} \ , \tag{17}$$

Where P_o and P_m are the value of gauge observed and predicted precipitation, respectively. N is the total number of samples. μ_m and μ_o are the mean value of gauge observed and predicted precipitation. SD_o and SD_m are the standard deviation of gauge observed and predicted precipitation, respectively. The optimal value for CC, KGE, β, γ is 1, while for MAE, and RMSE is 0.

4   Result

4.1 Performance assessment for classification results

The classification accuracy (*Accuracy*) of different ML models for wet/dry days is shown in Table 2. The general

performances are considerable. The *Accuracy* for three models is higher than 91% in the whole period, which is 91.9%, 91.8%,

and 91.8% for RF, GBDT, and XGBoost, respectively. The *Accuracy* in the cold season is better than that in the warm season.

350 There is no significant difference among the three classification algorithms, the main reason is that the input variables used in

this study are sufficient in variety and quantity, but their efficiency and ease of use should be different.

Table 2. The classification accuracy (*Accuracy*) of wet/dry day during the warm season and cold season

	RF	GBDT	XGBoost
Cold season	93.7	93.6	93.6
Warm season	90.0	89.9	90.0
Whole period	91.9	91.8	91.8

To better understand the performance of different technologies, the multi-source merged precipitation products (MSMPs),

gauge-based Kriging interpolated (Kriging), and original precipitation products (MSPs) are evaluated based on independent

355 gauge observations. The six categorical metrics (POD, FAR, CSI, *precision*, FB, and HSS) are shown in Fig.5, and the average

values of all gauges are expressed in Table 3. The overall accuracy of three MSMPs (GBDT, XGBoost, and RF) significantly

outperforms other products. The best values of all metrics (except for POD) are generated in MSMPs. ERAI and Kriging have

the highest POD with a value of 0.93 (Fig. 5a), followed by GSMaP (0.91). However, the POD of GBDT, XGBoost, and RF

are 0.84, 0.86, and 0.85, respectively. The FAR (Fig. 5b) of MSMPs ranges from 0.12 to 0.13, decreased by 55- 75% and 52%

360 compared with the original MSPs (0.29-0.52) and Kriging (0.27). In addition, the RF and XGBoost obtain the highest CSI

with the value of 0.76, much higher than original MSPs (0.3-0.67) and the best performing Kriging (0.69) (Fig. 5c). In terms

of *precision* (Fig. 5d), MSMPs achieve a significant improvement, the *precision* increases from 0.47-0.71 (MSPs) to 0.87-0.88

(MSMPs). For FB (Fig. 5e), MSPs and Kriging deviate from 1, among which PERCDR has the worst value (1.83) and TMPA

has the best value (0.92). Although ERAI achieves a high POD, its FB is 1.74, indicating ERAI has seriously overestimated

365 wet days and misclassified many precipitation events. Fortunately, MSMPs strike a good balance between hit and false alarmed

rates. The FB of MSMPs is closer to 1, which is 0.96 for GBDT, 1 for XGBoost, and 0.98 for RF. In terms of HSS (Fig. 5f),

except for Kriging (0.71) and GSMaP (0.69), the HSS of MSPs is lower than 0.5 (0.3-0.49). In contrast, the MSMPs (0.79-0.8)

improve by 14% to 163%.

Table 3. The average value of categorical metrices of multiple products compared with gauge observations during whole period.

Metrics	CHIRPS	CMORPH	PERCDR	GSMaP	TMPA	IMERG	ERAI	Kriging	GBDT	XGBoost	RF
POD	0.36	0.70	0.75	0.91	0.59	0.78	**0.93**	**0.93**	0.84	0.86	0.85
FAR	0.36	0.37	0.52	0.29	0.32	0.41	0.45	0.27	**0.12**	0.13	0.13
CSI	0.30	0.48	0.39	0.67	0.46	0.50	0.53	0.69	0.75	**0.76**	**0.76**
precision	0.64	0.63	0.48	0.71	0.68	0.59	0.55	0.73	**0.88**	0.87	0.87
FB	0.61	1.20	1.83	1.31	0.92	1.38	1.74	1.34	0.96	**1.00**	0.98

| HSS | 0.30 | 0.48 | 0.31 | 0.69 | 0.48 | 0.49 | 0.48 | 0.71 | 0.79 | **0.8** | **0.8** |

370   Note: the values in bold are the best performing of each metric.

The general performance of all products (including MSPs, Kriging, and MSMPs) in the warm season is better than that in the cold season (Fig. 5). However, the MSMPs' performance difference between warm and cold seasons is smaller than that of MSPs, demonstrating that the ability of MSMPs is more balanced throughout the year. Moreover, the metrics' variation of original MSPs is considerable in the cold season, particularly FAR and *precision*. The boxplot of FAR (Fig. 5b) and *precision*

375   (Fig. 5d) for CHIRPS, CMORPH, PERCDR, and TMPA have wider ranges, which represents these values are unevenly spatially distributed. In contrast, MSMPs have more concentrated ranges of boxplots in most metrics. These results emphasize the necessity of giving priority to precipitation state recognition in the merging process, which can largely improve the precipitation capture efficiency of MSPs.

Hydrology and
Earth System
Sciences
Discussions

Open Access

EGU
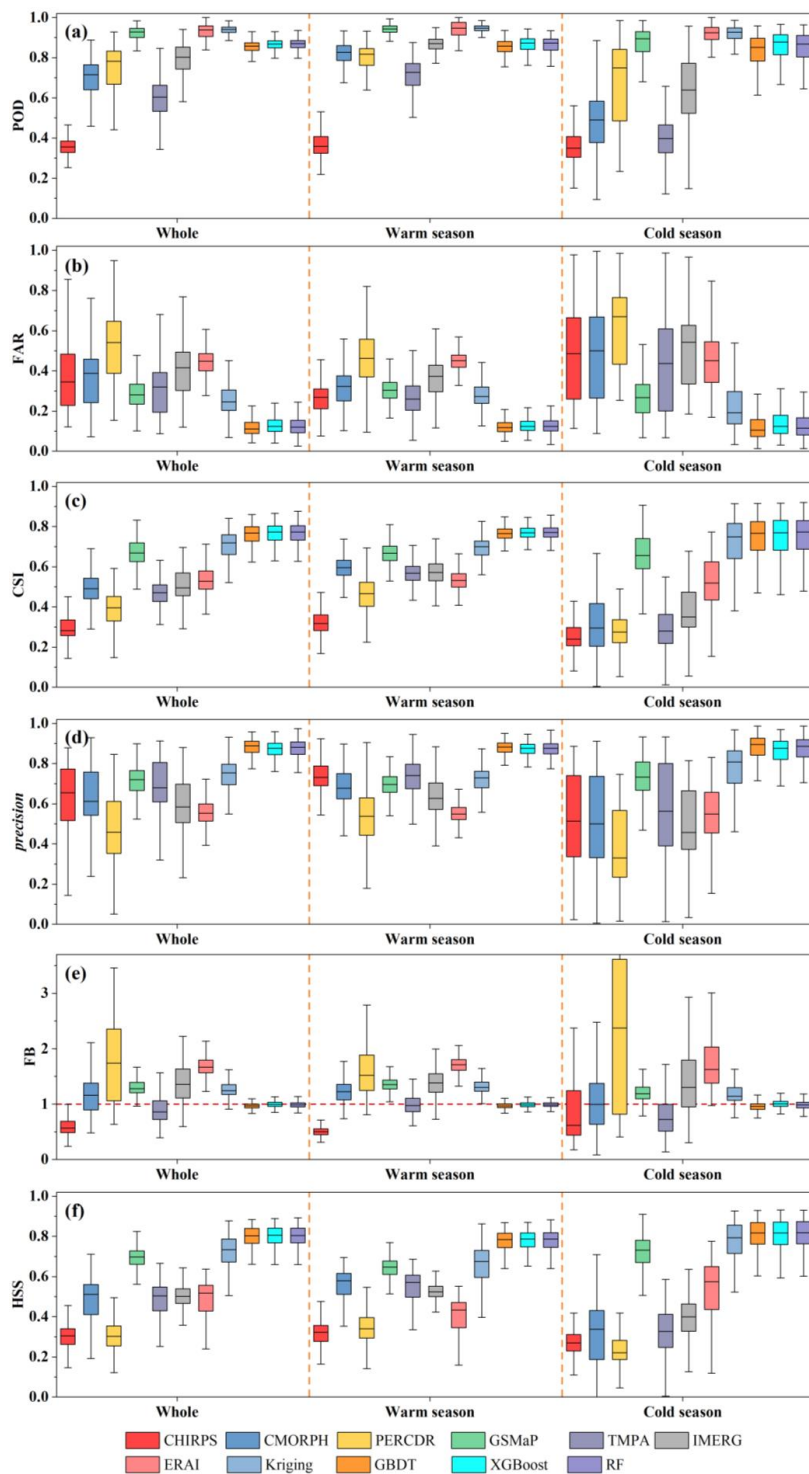
Figure 5. Boxplots of six categorical metrics (POD, FAR, CSI, *precision*, FB, and HSS) for eleven products, including seven MSPs, one gauge-based interpolated data, and three ML-based merged data.
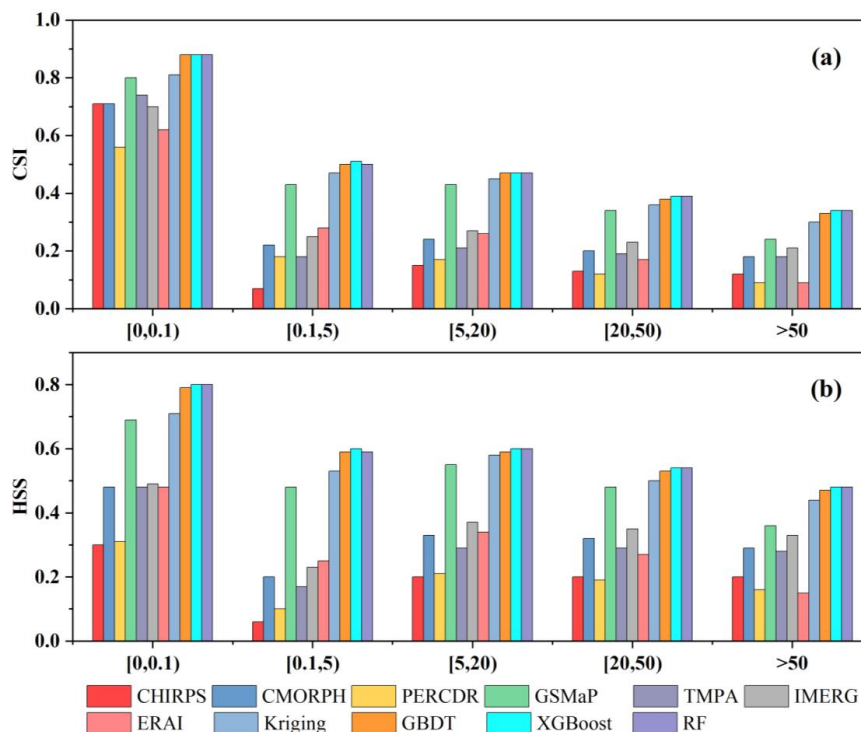
380

Figure 6. The CSI and HSS of eleven products under various daily precipitation threshold.

Fig. 6 shows the average value of CSI and HSS for eleven products under different precipitation intensities, including

385　no-precipitation (<0.1 mm/d), light precipitation ([0.1, 5)), moderate precipitation ([5, 20)), heavy precipitation ([20, 50)), and

violent precipitation (>50 mm/d). Overall, MSMPs have the best performance regardless of precipitation intensities, followed

by Kriging and GSMaP. This signifies that ML classification techniques improve the detection capability of all precipitation

thresholds, not only for light and moderate precipitation events. The performance of all products for no-precipitation is

considerably better than other precipitation intensities. The CSI and HSS of MSMPs are 0.88 and 0.79-0.8, respectively, in no-

390　precipitation. Meanwhile, Most MSPs (except GSMaP) have poor ability to capture light and moderate precipitation (0.1-20

mm/d). Their CSI is below 0.3 and HSS is below 0.4, while the HSS of MSMPs varies between 0.59 to 0.6. The accuracy

begins to decrease when precipitation intensity is above 20mm/d (i.e., heavy and violent precipitation). For violent precipitation

(> 50mm/d), the accuracy reduction of MSMPs and Kriging is relatively tiny compared with original MSPs. RF and XGBoost

have the highest CSI (0.34) and HSS (0.48). Generally, XGBoost and RF are slightly superior to GBDT when dividing

395　precipitation thresholds. In addition, although Kriging exhibits better performance than original MSPs, its accuracy is strongly

dependent on gauge density. This only gauge-based interpolation method would have limited in complex mountainous areas

with few gauges.

4.2 Performance assessment for regression results

The final MSMPs are predicted by regression models (GBDT, XGBoost, and RF) based on classification results. The

400   regression results include GBDT-GBDT, XGBoost-XGBoost, and RF-RF, respectively named GBDT2, XGB2, and RF2 for

simplicity. We also employ five statistical metrics (CC, RMSE, KGE and its components (β and γ)) to compare original MSPs
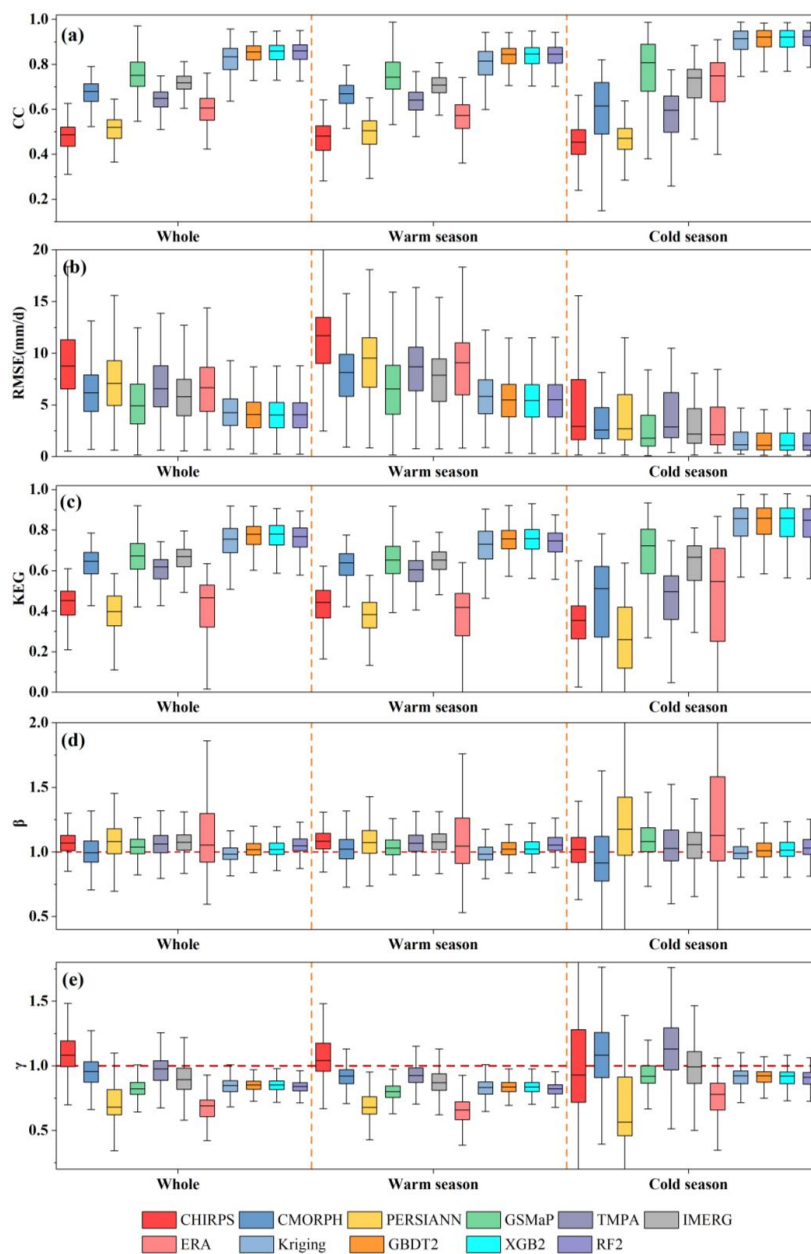

Figure 7. Boxplots of six statistical metrics (CC, RMSE, KGE, β, and γ) for eleven products, including seven MSPs, one gauge-based interpolated data, and three ML-based merged data.

405 and Kriging with GBDT2, XGB2, and RF2 based on daily observations. According to comparison results (Fig. 7, Table 4), the

MSMPs outperform original MSPs to varying degrees, followed by Kriging. The MSMPs have a strong correlation with gauge

observations in the warm season (CC: 0.83-0.84), cold season (CC: 0.9), and the whole period (CC: 0.85) (Fig. 7a), which is

significantly better than MSPs (warm:0.45-0.76; cold: 0.45-0.72; whole: 0.47-0.76). For RMSE (Fig. 7b), the values in the

warm season are higher than that in the cold season. This is because precipitation is mainly concentrated in the warm season,

410 and higher precipitation amounts often lead to larger RMSE. The RMSE for MSMPs decreases by 25% - 53% compared with

original MSPs (5.19 - 8.85mm/d). Among MSMPs, XGB2 exhibits the smallest RMSE with a value of 4.16 mm/d. In addition,

KGE and its components β and γ of MSMPs also present outstanding performances (Fig. 7d, e, f). The KGE is 0.74-0.76 for

MSMPs, while it is 0.68 for Kriging and 0.36-0.65 for MSPs in the whole period. Furthermore, the KGE of MSPs is as low as

0.23-0.59 in the cold season. The β shows that all MSPs and Kriging overestimate precipitation amounts. This overestimation

415 is more prominent in the cold season, with values ranging between 6%-49%. In contrast, MSMPs show substantial

improvement and GBDT2 obtains the best skill in all seasons (whole: 3%, warm: 3%, cold: 4%) compared with XGB2 and

RF2. Although Kriging and CMORPH have better performance than RF2 in the whole period and warm season, they suffer

from a large magnitude of overestimation (Kriging:10%, CMORPH (13%)) in the cold season. In terms of γ, the variability

ratio of CHIRPS, CMORPH, TMPA, and IMERG are more consistent with 1 than MSMPs in the whole period and the cold

420 season. However, they show more discreteness, particularly for CHIRPS. In comparison, the distribution of MSMPs values is

more compact. The results indicate that MSMPs can merge the complementary advantages of original data and reduce errors

to a large extent, especially in the cold season.

Table 4. The average values of statistic metrices of multiple products compared with gauge observations during whole period (The unit of RMSE is mm/d).

Metrics	CHIRPS	CMORPH	PERCDR	GSMaP	TMPA	IMERG	ERAI	Kriging	GBDT2	XGB2	RF2
CC	0.47	0.66	0.51	0.76	0.63	0.71	0.59	0.81	**0.85**	**0.85**	**0.85**
RMSE	8.85	6.29	7.22	5.19	6.84	5.94	6.71	4.49	4.18	**4.16**	4.16
KGE	0.41	0.58	0.34	0.65	0.57	0.64	0.36	0.68	0.75	**0.76**	0.74
β	1.09	1.05	1.14	1.07	1.09	1.09	1.18	1.04	**1.03**	1.04	1.07
γ	1.10	0.95	0.71	0.83	**0.97**	0.9	0.67	0.83	0.85	0.85	0.83

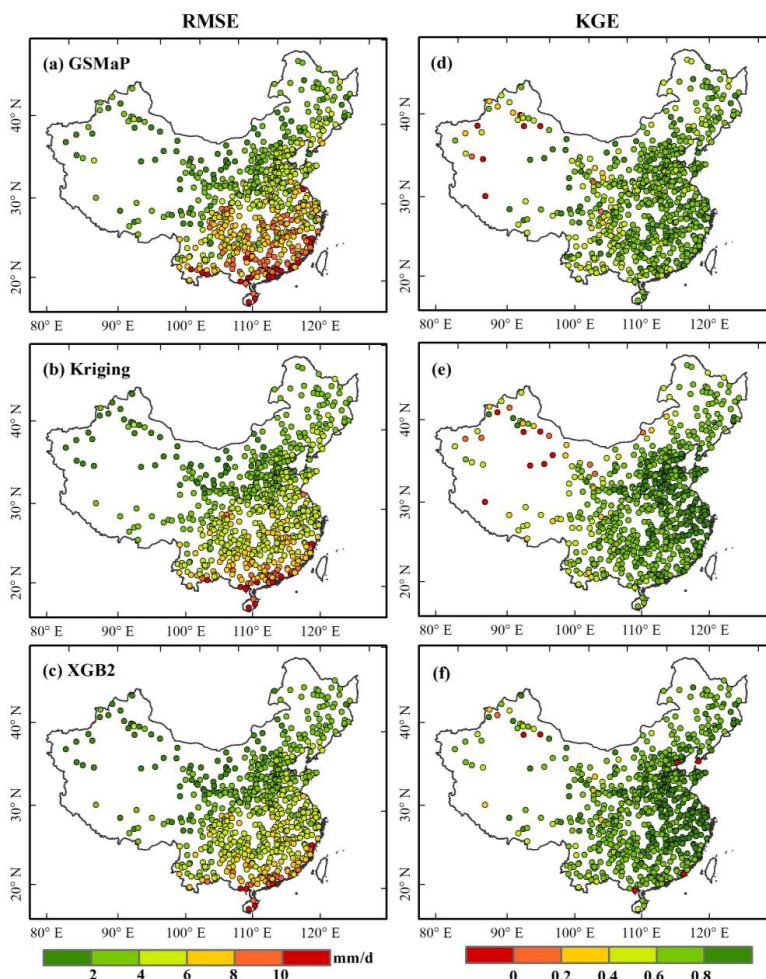425 Note: the values in bold are the best performing of each metric.

Figure 8. Spatial distribution of RMSE (a-c) and KGE (d-e) for GSMaP, Kriging, and XGB2 in the whole period from 2000-2017 using independent rain gauges over mainland China.

Fig.8 illustrates the spatial distribution of RMSE and KGE for GSMaP, Kriging, and XGB2 in the whole period. The reason for showing only these three products is that they have better performance among original products and MSMPs. The spatial comparison among them is more representative and brevity. The RMSE gradually increases from north to south, which is consistent with the precipitation change pattern (Fig. 8a). The XGB2's RMSE in south China has better performance than Kriging's, followed by GSMaP. For XGB2, approximately 49% of the gauges have RMSE less than 4 mm/d. The percentage of gauges with RMSE higher than 8 mm/d is 14% for GSMaP, 6% for Kriging, and 4% for XGB2. In addition, the spatial distribution of KGE shows that the low values are mainly gathered in the northwest (Fig. 8d-f). About 38% of the gauges with KGE higher than 0.8 for XGB2, while only 14% for GSMaP and 30% for Kriging. The XGB2 improves KGE performance over the northwest region, indicating the two-step merging approach could mitigate the spatial variability of products and is

21

less susceptible to topography.

4.3 Variable importance of ML models

440   The variable importance can quantitatively explain their contribution to improving model accuracy and recognize crucial input variables. The permutation feature importance is utilized to calculate variable importance values of models. The basic idea of this method is to randomly shuffle the order of specific variable while keeping other variables unchanged and compute the accuracy difference (the evaluation metric is *Accuracy* for the classification model, mean squared error for the regression model) with the original model. As Fig. 9 shown, the importance of variables for GBDT, XGBoost, and RF and their ranks are

445 different, which is related to the inherent structure of each model. This phenomenon also exists between classification and regression models. Nonetheless, KP is always the most important variable in every model, proving that the Kriging_based predictor considering the spatial autocorrelation between rain gauges is pretty helpful to improve model efficiency. For all models, the top three variables in importance are KP, GSMaP, and IMERG. The CMORPH, PERCDR, ERAI, and temperature is considered next significant. The importance of ERAI and temperature in regression models is more obvious than that in

450 classification models. Additionally, TMPA, longitude, latitude, DEM, cloud cover, and relative humidity exhibit relatively low influence on precipitation merging. The impacts of CHIRPS, soil moisture, and wind speed on prediction results are negligible. However, this does not mean that these predictors are not important for precipitation in whole regions. The slight importance of the latter variables may be affected by data quality and the correlation degree with precipitation. For example, CHIRPS is the worst performance product among original MSPs. Overall, it is necessary to employ multiple covariables in classification

455 and regression models since complex precipitation processes cannot be thoroughly described by a single variable.
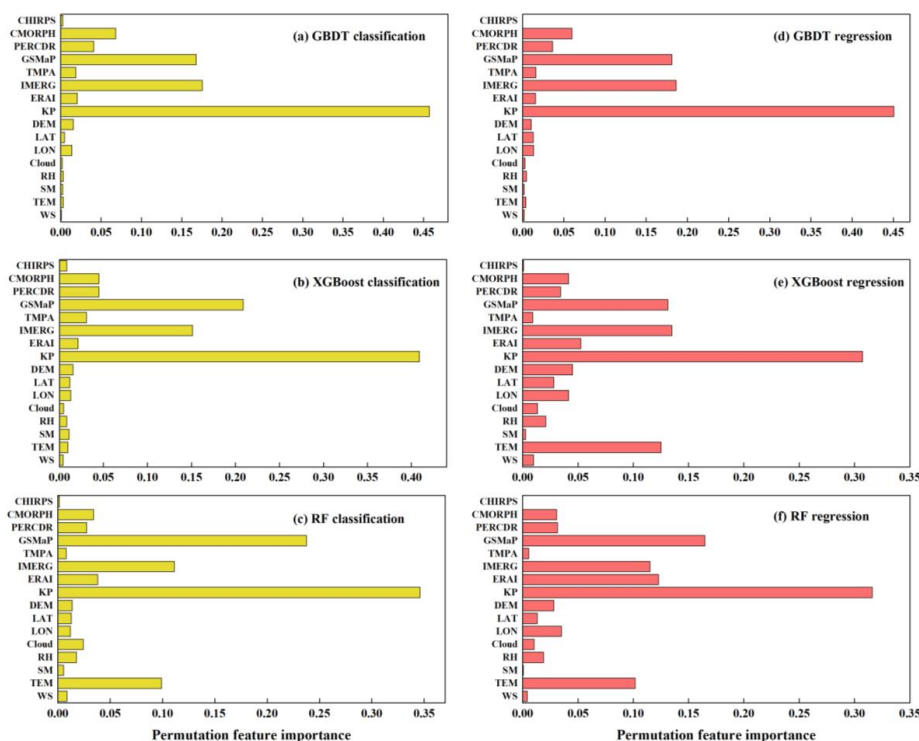
Figure 9. Permutation feature importance of three (GBDT, XGBoost, and RF) classification models (a-c) and three regression model (d-f) in the warm season (LAT is latitude, LON is longitude, RH is relative humidity, SM is soil moisture, TEM is temperature, and WS is wind speed).

5   Discussion

5.1 Comparison of prediction accuracy of various merging approaches

From the aspect of merging processes, different models and training samples could affect the accuracy of the integrated dataset. Therefore, three additional merging scenarios are considered for quantitative comparison with the proposed strategy to highlight the impact of samples' seasonal division and algorithm selection on fusion results. Fig. 10 gives a brief overview of four scenarios and their corresponding merged precipitation products. Scenario1 is the method adopted in this study; scenario 2 separately trains and predicts precipitation in each season only based on four regression models (GBDT, XGBoost, RF, and MLR), the corresponding results are GBDT_R, XGB_R, RF_R, and MLR; scenario3 applies both three classification and regression models (GBDT, XGBoost, and RF) during the entire period, not considering seasonal precipitation, the results are GBDT_E, XGB_E, and RF_E; while scenario4 solely employs four regression models during the entire period, the results are GBDT_ER, XGB_ER, RF_ER, and MLR_ER.
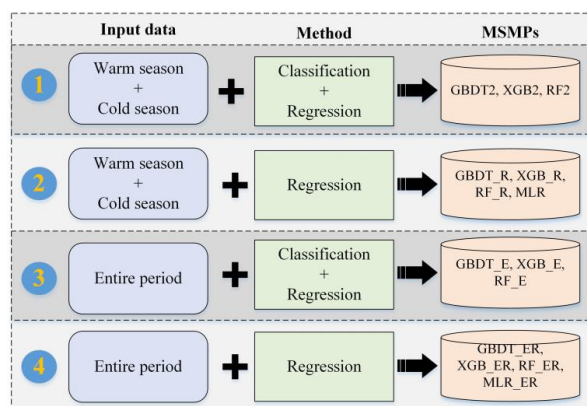
Figure 10. Four scenarios with different sample periods and different models

Fig. 11 shows the evaluation results (CC, CSI, KGE, FB, and HSS) of four scenarios between 14 MSMPs and independent

475   gauge observations. The performance of scenario1 is apparently better than other scenarios. For scenarios2, although the

statistical metrics (CC and KGE) are only slightly worse than scenario1, the categorical metrics (CSI, FB, and HSS) are

considerably weakened. In the whole period (Fig. 11a), the HSS is between 0.66-0.69 for scenario2, much lower than 0.79-0.8

for scenario1. Moreover, the FB of scenario2 is larger than 1.38 (Fig. 11a), indicating that a number of precipitation events

have been overestimated. A similar phenomenon also occurs in warm and cold seasons (Fig. 11b, c). Meanwhile, the MLR

480   performs worse than the three ML models. The results of scenario2 demonstrates that only relying on regression models to

merge precipitation can describe precipitation intensity but not capture precipitation occurrence well. In term of scenario3, the

overall performance is superior to scenario2 but inferior to scenario1. Compared with scenario1, the accuracy deterioration of

scenario3 is more obvious in the cold season. The CSI (Fig. 11c) for scenario1 and scenario3 range from 0.73-0.74 and 0.68-

0.70, respectively. Scenario3 suggests that seasonal precipitation fusion based on precipitation characteristics could effectively

485   balance the performance differences within a year. Scenario4 shows the worst performance regardless of season, with poor

CSI, FB, and HSS. Especially for the MLR-ER dataset, its accuracy is even worse than GSMaP and Kriging. This is because

MLR is difficult to describe the complex relationship between precipitation and other variables. In general, the four scenarios

can be ranked by prediction accuracy from best to worst: Scenario1 > Scenario3 > Scenario2 > Scenario4. The approach (i.e.,

Scenario1) employed in this study is proved to be more reliable and robust than other traditional strategies.
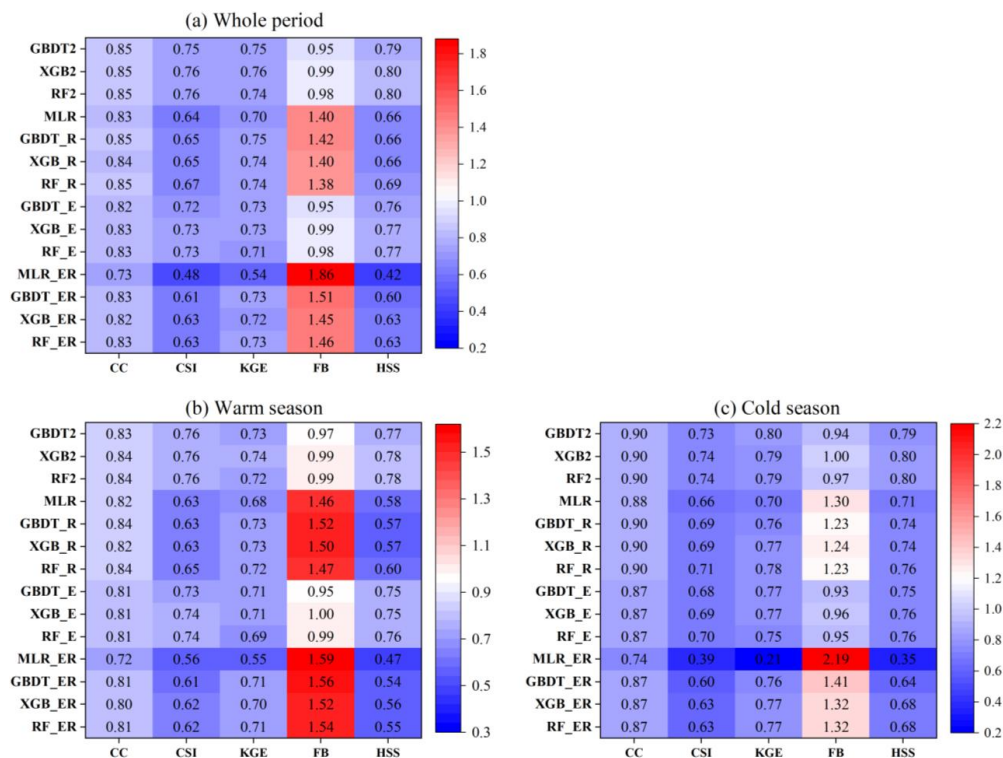
490

Figure 11. Five evaluation metrics (CC, CSI, KGE, FB, and HSS) for different products under four scenarios during the whole period, warm season, and cold season.

495 **5.2 Models efficiency**

The GBDT, XGBoost, and RF models show similar improvements in the two-step merging strategy. Nevertheless, different models have their inherent advantages and disadvantages. There is an apparent disproportion between positive and negative samples (wet and dry days) when training the classification model, which directly impacts the model's classification accuracy. In this study, the proportion of positive and negative samples in the cold season is approximately 1: 3.2. In terms of

500 this imbalance problem, RF and XGBoost algorithms have built-in parameters to adjust. However, GBDT requires additional oversampling methods such as the Synthetic Minority Over-sampling Technique (SMOTE) method to solve, which increases the complexity of model training. Meanwhile, it can be inferred from the results of Table 3, Fig. 5, and Fig. 11 that the FB of XGBoost outperforms RF in all seasons, indicating XGBoost has better equilibrium ability for disproportional samples. In addition, Fig. 12 displays the computational costs of training for three models under different sample sizes. The result exhibits

505 regression models are more time-consuming than classification models, and the training time of GBDT and RF is much higher than XGBoost. The RF algorithm shows considerable computational costs in the regression process, which is mainly related to the model structure and parallel training. XGBoost parallels the feature granularity rather than the tree granularity. The most

time-consuming part of decision tree learning is sorting feature values to determine the optimal split node. XGBoost ranks the

values before training and then saves them into a block structure, which is repeatedly used in subsequent iterations. In this way,

510   the training time could be vastly reduced (Chen et al., 2016b; Wang et al., 2019). Therefore, considering the complexity,

accuracy, and computational costs of the model, XGBoost is an optimal choice for predicting daily precipitation over China.
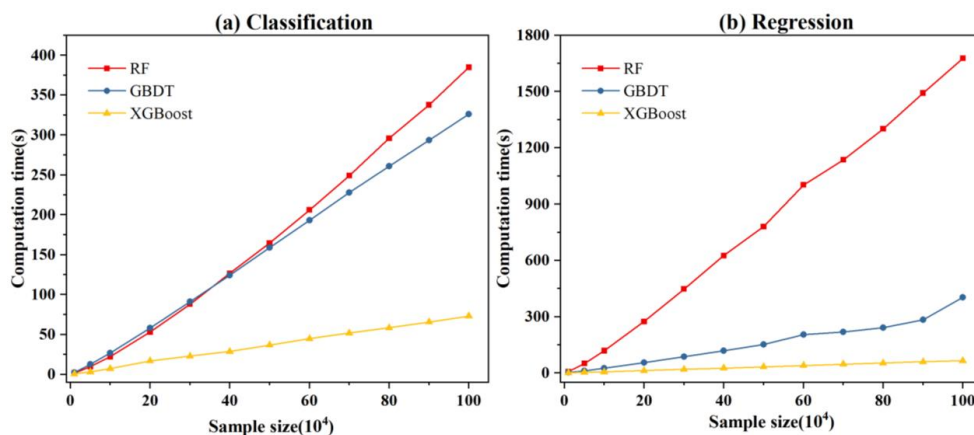
Figure 12. Comparison of computation time of three ML models (both classification (a) and regression (b)) under different sample sizes.

515   **5.3 Novelty and limitation**

The shining points of this study can be summarized from several aspects. Firstly, the input precipitation products from

multiple sources (including rain gauges, satellite, and reanalysis) have their strength and weakness in describing precipitation

features, which can integrate valuable information from different sensors and algorithms more comprehensively. Compared

with previous study over China, such as Wu et al. (2020), Zhang et al. (2020), Yin et al. (2021), Xiao et al. (2021) and Chen et

520   al. (2021), the MSPs used in this study are diverse and more accurate in their respective product series. Meanwhile, relatively

dense rain gauges are utilized for model training and validation, making the fusion results more reliable and persuasive. Second,

various covariates that might influence precipitation are taken into account, particularly spatial autocorrelation between gauges,

which further improves the performance of products. Using multiple covariates to merge precipitation could reduce the

uncertainty caused by solely relying on a single variable to some extent (Xiao et al., 2021). Last but most importantly, in

525   addition to reducing the bias and variation of precipitation intensity, the proposed method pays special attention to optimizing

precipitation detection ability. The ML method can solve the nonlinear problem between precipitation and variables and is less

affected by terrain and climate. Few studies have compared the application of GBDT, XGBoost classification and regression

models with RF in precipitation fusion. This research obtains great improvements built on the excellent investigations that

previous studies have been done. Compared with Zhang et al.'s (2020) research, the best *Accuracy*, CC, and KGE of their

530   study was 89.9%, 0.78, and 0.71 over China, while for this study is 91.7%, 0.85, and 0.77. In addition, the classification

accuracy is also better than the ANN model (86.5%) applied by Xiao et al. (2020) over Hanjiang River Basin and the RF model

employed by Pham (77.5%) at Shih-Men Reservoir catchment in northern Taiwan. Apart from the merging framework, the

superiority of this study is mainly attributed to the quality of input products, the richness of covariates, and the sufficiency of

training samples.

535   Although this study has achieved outstanding merging results, some limitations still need to be further improved in future

studies. This study directly uses the precipitation data over entire China for model training. The input datasets could be divided

into more groups according to different terrain or altitude zones in the subsequent work, and snow properties, such as snow

water equivalent and snow cover, could be combined in high latitude areas. Meanwhile, the warm and cold seasons can be

separated according to the characteristics of precipitation occurrence and end within a year instead of specific months. In

540   addition, although the accuracy of different precipitation intensities has been improved, quantile regression based on ML

should be considered in the follow-up to enhance the performance of extreme precipitation. Another pity is that the input data

used in this study are not near-real-time products. The merging results are suitable for studying hydrometeorological changes

in long time series, but there are still limitations for flood prediction in the middle or short term. Finally, a trial-and-error

procedure is used to optimize model parameters due to the limitation of computing resources. K-Fold grid search cross-

545   validation algorithm may be chosen to obtain a better combination of hyperparameters if computing resources are available.

6   Conclusion

This study proposes a two-step merging strategy including GBDT, XGBoost, and RF classification and regression

algorithms to merge multi-sources precipitation products, multiple environment variables, and rain gauges in each season from

2000 to 2017 over China. The performance of three merged products (MSMPs) is validated based on 692 randomly selected

550   independent gauges and compared with original MSP, Kriging, and other traditional merging scenarios (e.g., ML regression

and MLR). Several statistical and categorical metrics are employed to quantitatively describe the precipitation detection

capability and precipitation uncertainties. The main findings of this study can be concluded as follows:

(1)   The precipitation capture ability of the MSMPs is better than the original MSPs regardless of the precipitation

intensities, followed by Kriging. The CSI for original MSPs and Kriging is 0.30-0.67 and 0.69, while MSMPs are

555   increased to 0.75-0.76. The *precision* improves by 23-81% (0.87-0.88) compared with MSPs (0.48-0.71).

(2)   The MSMPs suffer from fewer errors and variations in spatial and temporal. They have a higher correlation with

gauges with the CC value up to 0.85, larger than MSPs (0.47-0.76) and Kriging (0.81). The average KGE increases

from 0.34-0.65 (MSPs) to 0.74-0.76 (MSMPs). Considerable improvements are also exhibited in RMSE and β.

(3)   Kriging-based predictor (KP) is the most important covariable in precipitation merging, followed by GSMaP, IMERG,

560 ERAI, and temperature, showing that it is essential to incorporate spatial autocorrelation in the merging strategy. In

addition, the degree of importance for covariables in models may also be related to their inherent accuracy.

(4) Compared with the merging results based on traditional MLR and solely on ML regression models, the method

proposed in this study has superior performance. It can balance and improve the quality of precipitation in all aspects.

Meanwhile, the merging results conducted in each season outperform those in the whole period.

565 (5) Among the three ML models used in this study, the comprehensive ability of RF and XGBoost is slightly better than

GBDT. Considering the computation efficiency, it is recommended to use XGBoost to prepare precipitation merging

products.

The two-step merging strategy proposed in this study achieves satisfactory performance over China, bringing new

information for ecological and hydrological researches. The method based on XGBoost has high efficiency, thus having great

570 referential significance for other areas with similar topography and extensive computation needing.

Author contribution: Huajin Lei designed the methodology and collected datasets, Huajin Lei and Hongyu Zhao implemented

580 algorithms' code, Huajin Lei analyzed the results and wort the original manuscript. Tianqi Ao supervision. All the authors

revised and improved the manuscript.

585 **References**

Arshad, A., Zhang, W., Zhang, Z., Wang, S., Shalamzari, M. J.: Reconstructing high-resolution gridded precipitation data using

an improved downscaling approach over the high altitude mountain regions of upper Indus basin (UIB), Sci. Total Environ.,

784, 147140, https://doi.org/10.1016/j.scitotenv.2021.147140, 2021.

Ashouri, H., Hsu, K.L., Sorooshian, S., Braithwaite, D.K., Knapp, K.R., Cecil, L.D., Nelson, B.R., Prat, O.P.: PERSIANN-

590   CDR: Daily Precipitation Climate Data Record from Multisatellite Observations for Hydrological and Climate Studies, Bull.

Amer. Meteor. Soc., 96, 69-83, http://dx.doi.org/10.1175/BAMS-D-13-00068.1, 2015.

Awange, J. L., Hu, K. X., Khaki, M.: The newly merged satellite remotely sensed, gauge and reanalysis-based multi-source

weighted-ensemble precipitation: evaluation over Australia and Africa (1981–2016), Sci. Total Environ., 670, 448-465,

https://doi.org/10.1016/j.scitotenv.2019.03.148, 2019

595   Baez-Villanueva, O.M., Zambrano-Bigiarini, M., Beck, H.E., McNamara, I., Ribbe, L., Nauditt, A., Birkel, C., Verbist, K.,

Giraldo-Osorio, J.D., Xuan Thinh, N.: RF-MEP: A novel Random Forest method for merging gridded precipitation products

and ground-based measurements, Remote Sens. Environ., 239, 111606, https://doi.org/10.1016/j.rse.2019.111606, 2020

Bhuiyan, M., Anagnostou, E. N., Kirstetter, P. E.: A nonparametric statistical technique for modeling overland TMI(2A12)

rainfall retrieval error, IEEE Geosci. Remote S., PP (11), 1-5, https://doi.org/10.1109/LGRS.2017.2728658, 2017.

600   Bhuiyan, M., Nikolopoulos, E. I., Anagnostou, E. N., P Quintana-Seguí, Barella-Ortiz, A.: A nonparametric statistical

technique for combining global precipitation datasets: development and hydrological evaluation over the Iberian Peninsula.

Hydrol. Earth Syst. Sci., 22(2), 1371-1389, https://doi.org/10.5194/hess-22-1371-2018, 2018.

Breiman, L.: Arcing the edge, Statistics Department, University of California at Berkeley, Tech. Rep. 486, 1997.

Breiman, L.: Random forests. Mach. Learn. 45, 5–32, 2001.

605   Chao, L., Zhang, K., Li, Z., Zhu, Y., Wang, J., Yu, Z.: Geographically weighted regression based methods for merging satellite

and gauge precipitation, J. Hydrol., 558, 275-289, https://doi.org/10.1016/j.jhydrol.2018.01.042, 2018.

Chen, C., Hu, B., Li, Y.: Easy-to-use spatial Random Forest-based downscaling-calibration method for producing high

resolution and accurate precipitation data, Hydrol. Earth Syst. Sci., https://doi.org/10.5194/hess-2021-332, 2021.

Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system, in: Proceedings of the ACM SIGKDD International

610   Conference on Knowledge Discovery and Data Mining. 785-794, 2016.

Collischonn, B., Collischonn, W., Carlos, E., Morelli, T.: Daily hydrological modeling in the Amazon basin using TRMM

rainfall estimates, J. Hydrol., 360, 207–216, https://doi.org/10.1016/j.jhydrol.2008.07.032, 2008.

Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S., et al.: The ERA-Interim reanalysis:

configuration and performance of the data assimilation system, Q. J. R. Meteorol. Soc, 137 (656), 553–597,

615   https://doi.org/10.1002/qj.828, 2011.

Friedman, J.H.: Stochastic gradient boosting, Comput. Stat. Data An., 38, 367–378, https://doi.org/10.1016/S0167-

9473(01)00065-2, 2002.

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A.,

Michaelsen, J.: The climate hazards infrared precipitation with stations - a new environmental record for monitoring extremes,

620   Sci. Data 2, 1–21. https://doi.org/10.1038/sdata.2015.66, 2015.

Ghorbanpour, A. K., Hessels, T., Moghim, S., Afshar, A.: Comparison and assessment of spatial downscaling methods for enhancing the accuracy of satellite-based precipitation over Lake Urmia Basin, J. Hydrol.,596, 126055. https://doi.org/10.1016/j.jhydrol.2021.126055, 2021.

He, X., Chaney, N., Schleiss, M., Sheffield, J.: Spatial downscaling of precipitation using adaptable random forests, Water

625   Resour. Res., 52, 8217–8237, https://doi.org/10.1002/2016WR019034, 2016.

Herrera, S., Kotlarski, S., Soares, P. M., Cardoso, R. M., Jaczewski, A., Gutiérrez, J. M., Maraun, D.: Uncertainty in gridded precipitation products: Influence of station density, interpolation method and grid resolution, Int. J. Climatol., 39(9), 3717-3729. https://doi.org/10.1002/joc.5878, 2019.

Ho, T. K.: The Random Subspace Method for Constructing Decision Forests (PDF), IEEE. T. Pattern. Anal., 20 (8): 832–844.

630   https://doi.org/10.1109/34.709601, 1998.

Hong, Z., Han, Z., Li, X., Long, D., Wang, J.: Generation of an improved precipitation data set from multisource information over the Tibetan plateau. J. Hydrometeorol., https://doi.org/10.1175/JHM-D-20-0252.1, 2021.

Hsu, K. L., Gao, X., Sorooshian, S., Gupta, H.: Precipitation Estimation from Remotely Sensed Information Using Artificial Neural Networks, J. Appl. Meteorol. 36(9), 1176–1190. https://doi.org/10.1175/1520-

635   0450(1997)036<1176:PEFRSI>2.0.CO;2, 1997.

Huffman, G.J., Bolvin, D.T., Nelkin, E.J. et al.: Integrated Multi-satellitE Retrievals for GPM (IMERG) Technical Documentation. https://docserver.gesdisc.eosdis.nasa.gov/public/project/GPM/IMERG_doc.06.pdf, 2019.

Huffman, G.J., Bolvin, D.T., Nelkin, E.J., et al.: The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-Global, Multiyear, Combined-Sensor Precipitation Estimates at Fine Scales, J. Hydrometeorol. 8 (1), 38–55.

640   https://doi.org/10.1175/JHM560.1, 2007.

Jia, S., Zhu, W., Lű, A., Yan, T.: A statistical spatial downscaling algorithm of TRMM precipitation based on NDVI and DEM in the Qaidam Basin of China, Remote Sens. Environ., 115(12), 3069-3079. https://doi.org/0.1016/j.rse.2011.06.009, 2011.

Jiang, S., Ren, L., Yang, H., Yong, B., Yang, X., Fei, Y., Ma, M.: Comprehensive evaluation of multi-satellite precipitation products with a dense rain gauge network and optimally merging their simulated hydrological flows using the Bayesian model

645   averaging method, J. Hydrol., 452-453. http://dx.doi.org/10.1016/j.jhydrol.2012.05.055, 2012.

Joyce, R., Janowiak, J., Arkin, P., Xie, P.: CMORPH: A Method that Produces Global Precipitation Estimates from Passive Microwave and Infrared Data at High Spatial and Temporal Resolution, J. Hydrometeorol. 5 (3), 487–503, https://doi.org/10.1175/1525-7541(2004) 005<0487:CAMTPG>2.0.CO;2, 2004.

Kubota, T., Shige, S., Hashizume, H., Aonashi, K., Takahashi, N., Seto, S., Hirose, M., Takayabu, Y.N., Ushio, T., Nakagawa,

650   K., Iwanami, K., Kachi, M., Okamoto, K.: Global precipitation map using satellite-borne microwave radiometers by the GSMaP project: production and validation. IEEE Trans. Geosci. Remote Sens., 45 (7), 2259–2275. https://doi.org/10.1109/TGRS.2007.895337, 2007.

Kumar, A., Ramsankaran, R., Brocca, L., Munoz-Arriola, F.: A machine learning approach for improving near-real-time satellite-based rainfall estimates by integrating soil moisture, Remote Sens., 11(19), 2221-, https://doi.org/10.3390/rs11192221,

655   2019.

Lei, H., Li, H., Zhao, H., Ao, T., Li, X.: Comprehensive evaluation of satellite and reanalysis precipitation products over the eastern Tibetan plateau characterized by a high diversity of topographies, Atmos. Res., 259, https://doi.org/10.1016/j.atmosres.2021.105661, 2021.

Lu, X., Tang, G., Wang, X., Liu, Y., Zhang, Y.: Correcting GPM IMERG precipitation data over the Tianshan mountains in

660   China, J. Hydrol., 575(2019),1239-1252, https://doi.org/10.1016/j.jhydrol.2019.06.019, 2019.

Ma, Y., Yang, H., Yang, C., Yuan, Y., Tang, G., Yao, Y., Di, L., Li, C., Han, Z., Liu, R.: Performance of optimally merged multisatellite precipitation products using the dynamic Bayesian model averaging scheme over the Tibetan plateau. J. Geophys. Res.-Atmos., 123, https://doi.org/10.1002/2017JD026648, 2017.

Massari, C., Maggioni, V., Barbetta, S., Brocca, L., Ciabatta, L., Camici, S., Moramarco, T., Coccia, G., Todini, E.:

665   Complementing near-real time satellite rainfall products with satellite soil moisture-derived rainfall through a Bayesian inversion approach, J. Hydrol. 573, 341–351, https://doi.org/10.1016/j.jhydrol.2019.03.038, 2019.

Nerini, D., Zulkafli, Z., Wang, L.P., Onof, C., Buytaert, W., Lavadocasimiro, W., Guyot, J: A comparative analysis of TRMM-rain gauge data merging techniques at the daily time scale for distributed rainfall-runoff modeling applications, J. Hydrometeorol. 16, 2153–2168, https://doi.org/10.1175/JHM-D-14-0197.1, 2015.

670   Nguyen, G. V., Le, X. H., Van, L. N., Jung, S., Yeon, M., Lee, G.: Application of Random Forest Algorithm for Merging Multiple Satellite Precipitation Products across South Korea, Remote Sens., 13(20), 4033, https://doi.org/10.3390/rs13204033, 2021.

Pham, Q. B., Yang, T. C., Kuo, C. M., Tseng, H. W., Yu, P. S.: Combing random forest and least square support vector regression for improving extreme rainfall downscaling, Water, 11(3), 451, https://doi.org/10.3390/w11030451,2019.

675   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., Toll, D.: The Global Land Data Assimilation System, B. Am. Meteorol. Soc., 85(3), 381-394. https://doi.org/10.1175/BAMS-85-3-381, 2004.

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., Rigol-Sanchez, J. P.: An assessment of the effectiveness of a random forest classifier for land-cover classification, ISPRS J. Photogramm., 67, 93-104,

680   https://doi.org/10.1016/j.isprsjprs.2011.11.002, 2012.

Sharifi, E., Saghafian, B., Steinacker, R.: Downscaling satellite precipitation estimates with multiple linear regression, artificial neural networks, and spline interpolation techniques, Geophys. Res.-Atmos., 124, 789–805. https://doi.org/10.1029/2018JD028795, 2019.

Sharifi, E., Steinacker, R., & Saghafian, B.: Assessment of GPM-IMERG and other precipitation products against gauge data under different topographic and climatic conditions in Iran: Preliminary results, Remote Sens., 8(2), 135. https://doi.org/10.3390/rs8020135, 2016.

Shen, Y., Pan, Y., Yu, J., Zhao, P., Zhou, Z.: Quality assessment of hourly merged precipitation product over China. Trans. Atmos. Sci. 36 (1), 37–46. https://doi.org/10.13878 /j.cnki.dqkxxb.2013.01.005 (in Chinese), 2013.

Shen, Y., Zhao, P., Pan, Y., Yu, J.: A high spatiotemporal gauge-satellite merged precipitation analysis over China, J. Geophys. Res.-Atmos. 119 (6), 3063–3075. https://doi.org/10.1002/2013JD020686, 2014.

Shen, Z., Yong, B.: Downscaling the GPM-based satellite precipitation retrievals using gradient boosting decision tree approach over Mainland China, J. Hydrol., 602, 126803, https://doi.org/10.1016/j.jhydrol.2021.126803, 2021.

Tan, J., Xie, X., Zuo, J., Xing, X., Liu, B., Xia, Q.: Coupling random forest and inverse distance weighting to generate climate surfaces of precipitation and temperature with multiple-covariates, J. Hydrol., 598(7), 126270, https://doi.org/10.1016/j.jhydrol.2021.126270, 2021.

Tang, G.Q., Behrangi, A., Long, D., Li, C.M., Hong, Y.: Accounting for spatiotemporal errors of gauges: a critical step to evaluate gridded precipitation products, J. Hydrol. 559, 294–306, https://doi.org/10.1016/j.jhydrol.2018.02.057, 2018.

Tao, Y., Gao, X., Hsu, K., Sorooshian, S., Ihler, A.: A deep neural network modeling framework to reduce bias in satellite precipitation products, J. Hydrometeorol., 160114111258006, https://doi.org/10.1175/JHM-D-15-0075.1, 2016.

Wang, R., Lu, S., Li, Q.: Multi-criteria comprehensive study on predictive algorithm of hourly heating energy consumption for residential buildings, Sustain. Cities Soc., 49, 101623, https://doi.org/10.1016/j.scs.2019.101623, 2019.

Woldemeskel, F. M., Sivakumar, B., Sharma, A.: Merging gauge and satellite rainfall with specification of associated uncertainty across Australia, J. Hydrol., 499, 167-176, https://doi.org/10.1016/j.jhydrol.2013.06.039, 2013.

World Meteorological Organization (WMO): Guide to Hydrometeorological Practices, 402, World Meteorological Organization Issue 168 of WMO (Series), 1965.

Wu, H., Yang, Q., Liu, J., Wang, G.: A spatiotemporal deep fusion model for merging satellite and gauge precipitation in China, J. Hydrol., 584, 124664, https://doi.org/10.1016/j.jhydrol.2020.124664, 2020.

Wu, Z., Zhang, Y., Sun, Z., Lin, Q., He, H.: Improvement of a combination of TMPA (or IMERG) and ground-based precipitation and application to a typical region of the east China plain, Sci. Total Environ., 640-641(NOV.1), 1165-1175, 2018.

Xiao, S., Zou, L., Xia, J.: Bias correction framework for satellite precipitation products using a rain/no rain discriminative model, Sci. Total Environ., https://doi.org/10.1016/j.scitotenv.2021.151679, 2021.

Xie, P., Xiong, A.Y.: A conceptual model for constructing high-resolution gauge-satellite merged precipitation analyses, J. Geophys. Res.- Atmos., 116(D21), https://doi.org/10.1029/2011JD016118, 2011.

Xu, Q., Chen, J., Peart, M.R., Ng, C.N., Hau, B.C., Law, W.W.: Exploration of severities of rainfall and runoff extremes in

715   ungauged catchments: a case study of Lai Chi Wo in Hong Kong, China, Sci. Total Environ. 634, 640–649. https://doi.org/10.1016/j.scitotenv.2018.04.024, 2018.

Yang, Z., Hsu, K., Sorooshian, S., Xu, X., Dan, B., Yuan, Z., Koen.: Merging high-resolution satellite-based precipitation fields and point-scale rain gauge measurements - a case study in Chile. J. Geophys. Res.- Atmos., 122, 5267–5284, https://doi.org/10.1002/2016JD026177, 2017.

720   Yilmaz, K.K., Hogue, T.S., Hsu, K.L., Sorooshian, S., Gupta, H.V., Wagener, T.: Intercomparison of rain gauge, radar, and satellite-based precipitation estimates with emphasis on hydrologic forecasting, J. Hydrometeorol., 6 (4), 497–517. https://doi.org/10.1175/JHM431.1, 2005.

Yin, J., Guo, S., Gu, L., Zeng, Z., Xu, C.Y.: Blending multi-satellite, atmospheric reanalysis and gauge precipitation products to facilitate hydrological modelling, J. Hydrol., 593(1), https://doi.org/10.1016/j.jhydrol.2020.125878, 2020.

725   Zhang, L., Li, X., Zheng, D., Zhang, K., Ge, Y.: Merging multiple satellite-based precipitation products and gauge observations using a novel double machine learning approach, J. Hydrol., 594(6), 125969, https://doi.org/10.1016/j.jhydrol.2021.125969, 2021.

Zhou, Z., Guo, B., Xing, W., Zhou, J., Xu, Y.: Comprehensive evaluation of latest GPM era IMERG and GSMaP precipitation products over mainland China, Atmos. Res., 246, https://doi.org/10.1016/j.atmosres.2020.105132, 2020.