Article ID: 1006-8775(2023) 02-0264-13

Correction of CMPAS Precipitation Products over Complex Terrain Areas with Machine Learning Models

LI Shi-ying (李施颖)[1, 2], HUANG Xiao-long (黄晓龙)[1, 2], WU Wei (吴 薇)[1, 2],
DU Bing (杜 冰)[1, 2], JIANG Yu-he (蒋雨荷)[1, 2]
(1. Sichuan Meteorological Observation and Data Center, Chengdu 610072 China;
2. Heavy Rain and Drought-flood Disasters in Plateau and Basin Key Laboratory of Sichuan Province,
Chengdu 610072 China)

Abstract: Machine learning models were used to improve the accuracy of China Meteorological Administration Multisource Precipitation Analysis System (CMPAS) in complex terrain areas by combining rain gauge precipitation with topographic factors like altitude, slope, slope direction, slope variability, surface roughness, and meteorological factors like temperature and wind speed. The results of the correction demonstrated that the ensemble learning method has a considerably corrective effect and the three methods (Random Forest, AdaBoost, and Bagging) adopted in the study had similar results. The mean bias between CMPAS and 85% of automatic weather stations has dropped by more than 30%. The plateau region displays the largest accuracy increase, the winter season shows the greatest error reduction, and decreasing precipitation improves the correction outcome. Additionally, the heavy precipitation process' precision has improved to some degree. For individual stations, the revised CMPAS error fluctuation range is significantly reduced.

Key words: machine learning models; ensemble learning; precipitation correction; error correction; high-resolution precipitation; complex terrain

CLC number: P413 **Document code:** A

Citation: LI Shi-ying, HUANG Xiao-long, WU Wei, et al. Correction of CMPAS Precipitation Products over Complex Terrain Areas with Machine Learning Models [J]. Journal of Tropical Meteorology, 2023, 29(2): 264-276, https://doi.org/10.46267/j.1006-8775.2023.020

1 INTRODUCTION

A crucial part of weather forecasting, disaster prevention, and mitigation is the use of high-resolution, high-quality precipitation data (Harrison et al. [1]; Turk et al. [2]). These data can be evaluated to help develop more accurate high-resolution numerical weather prediction models (Lagasio et al. [3]), and monitor small - and medium-scale extreme precipitation events and the resulting flash floods, landslides, and mudslides (Hirabayashi et al. [4]; Nikolopoulos et al. [5]).

The criteria for the resolution and accuracy of precipitation products have improved with the growth of weather predictions and services (Shen et al. [6]; Hong et al. [7]; Huffman et al. [8]). The National Meteorological Information Center (NMIC) first established a baseline using precipitation data from automatic weather stations, then corrected the systematic bias of the radar and satellite precipitation products using the Probability Density Function (PDF) matching method (Simolo et al. [9]; Chen and Kumar [10]), and finally combined the radar and satellite precipitation products using the Bayesian Model Averaging method (Pan et al. [11]) to create a comprehensive and ideal joint precipitation background for the Chinese region. The spatial structure information of the estimated precipitation from the 1-km radar is further downscaled (Shen et al. [12]). Next, the combined satellite-radar precipitation products were utilized independently as a background to quantify the error estimates using statistical methods before being fused into rain gauge observations using Optimal Interpolation methods (Pan et al. [13]; Shen et al. [14]).

The NMIC used the above four methods (Pan et al. [15]) to create a three-source (gauge, satellite, and radar) fused precipitation product with spatial/temporal resolutions of 1 km h^{-1}. In China, this fused product is known as the China Meteorological Administration Multisource Precipitation Analysis System (CMPAS) (Shi et al. [16]; Pan et al. [17]), and it fully utilizes single-source precipitation products to create a complete and superior precipitation product. In addition, the NMIC has developed High Resolution China Meteorological Administration Land Data Assimilation System (HRCLDAS) (Han et al. [18]; Tie et al. [19]), which includes air temperature, 10-m wind, and specific humidity, with a spatial resolution of 1 km and a

Received 2022-09-04; **Revised** 2023-02-15; **Accepted** 2023-05-15

Funding: Program of Science and Technology Department of Sichuan Province (2022YFS0541-02); Program of Heavy Rain and Drought-flood Disasters in Plateau and Basin Key Laboratory of Sichuan Province (SCQXKJQN202121); Innovative Development Program of the China Meteorological Administration (CXFZ2021Z007)

Biography: LI Shi-ying, Engineer, primarily undertaking research on meteorological data evaluation and analysis.

Corresponding author: LI Shi-ying, e-mail: shiyingli0912@163.com

temporal resolution of 1 hour.

In complicated terrain areas, the temporal and spatial pattern of precipitation is quite complex, especially in plateau areas with sparse weather stations where there is a considerable disparity between CMPAS and rain gauge data (Li et al. [20]; Wu et al. [21]). Much research has revealed that the accuracy of the already regularly utilized precipitation fusion data still needs to be improved (Pang et al. [22]; Bai et al. [23]; Guo et al. [24]). Some studies revised precipitation products using a statistical calculation method based on precipitation occurrence and development patterns, and Yu et al. [25] improved the applicability of precipitation products in China by PDF method, but the revisions' accuracy is poor in the complex topography of western China. Topography (Wang et al. [26]), vegetation (Jia et al. [27]; Liu et al. [28]), and meteorological characteristics (Beuchat et al. [29]) including temperature and wind speed were used to improve it, and precipitation products for the Tibetan Plateau region were updated using the KNN algorithm (Huang et al. [30]; Yang et al. [31]). The findings were superior to the updated precipitation products using the PDF approach. Chen et al. [32] used different machine learning models to revise precipitation products, which are commonly used in the Chinese region, demonstrating that machine learning methods have significant revision effects on precipitation products.

Based on this, this research proposes a machine learning strategy to improve the CMPAS products' accuracy in locations with complicated terrain. Topographic features like altitude, slope, slope direction, slope variability, and surface roughness are extracted from high-precision Digital Elevation Model (DEM) data as topographic factors. These topographic factors are combined with precipitation-related temperature and wind speed as meteorological factors to pick the optimal model to revise the CMPAS products, and the revision effect is evaluated.

2 DATA AND METHODS

2.1 *Study area*

Sichuan province, China is selected for the study. In terms of topography, Sichuan is the most complex province in China, which is located in the southwestern part of the country, with high terrain in the west and low terrain in the east, sloping from northwest to southeast (Xie and Wang [33]; Lu et al. [34]). The altitude difference between the highest and lowest points in Sichuan is more than 7300 m, and the terrain is quite undulating (Huang et al. [35]).

Meteorologically, Sichuan is generally divided into three regions for analysis: the Western Sichuan Plateau, the Sichuan Basin (central and eastern Sichuan), and Panxi Area (southwestern Sichuan) (Luo et al. [36]; Zeng et al. [37]). The Western Sichuan Plateau is located on the east side of the Qinghai-Tibet Plateau, with an average altitude of more than 4000 m (Zhang [38]). The Sichuan Basin, the central region of thr province, has a total of 17 cities and is typically at 500 m or lower altitudes (Chen and Xie [39]). With an average altitude of 1300 m, the Panxi Area is a part of the Yunnan-Guizhou Plateau (Li et al. [40]).

2.2 *Data*

The CMPAS product is the subject of this revision, and the participants are the HRCLDAS temperature and wind speed products, DEM data, and automatic weather stations data. The study used the data from October 2020, January 2021, April 2021, and July 2021, representing the four seasons of autumn, winter, spring, and summer, respectively. The heavy precipitation process from June 12 to 13, 2021 is selected for the case study, and the hydrological station data from July 2021 is chosen for an independent analysis.

The CMPAS and HRCLDAS datasets are provided by the NMIC, CMA (China Meteorological Administration), with a resolution of 0.01° × 0.01° (original resolution: 1 km), and again the temporal resolution is hourly. China Meteorological Administration Multisource Precipitation Analysis System_Real Time (CMPAS_RT) from the CMPAS product is selected for the study, and CMPAS_RT is a real-time radar-satellite-gauge merged precipitation product.

The hourly surface precipitation, temperature, and wind speed data are collected from 1899 national automatic weather stations and regional automatic weather stations in Sichuan province, provided by the CMA and Sichuan Meteorological Service.

Topographic factors such as slope, slope direction, slope variability, and surface roughness of the CMAPS_RT grid points and all weather stations were extracted from the 90-m resolution DEM data released by the NMIC.

Hourly precipitation data of hydrological stations were from real-time shared data with Sichuan Provincial Water Resources Departments. Through the collaborative quality control of adjacent meteorological stations and radar products, 449 stations were selected to participate in the independent evaluation.

2.3 *Analysis methods*

Through nearest-interpolation, meteorological stations and CMPAS grids were spatially matched. Wu et al. [21] demonstrated that the nearest-interpolation approach produces better results for CMPAS evaluation, and since precipitation was local and dispersive, the nearest-interpolation was selected. And the errors of station precipitation and grid precipitation were calculated, BIAS, Mean Bias (MB), Root Mean Square Error (RMSE), Relative Error (RE), and Correlation Coefficient (COR) are the primary evaluation metrics. The MB reflects the average deviation of the grid values from the observed values, the RMSE reflects the degree of dispersion of the data, the RE reflects the accuracy of the grid values, and the COR shows the degree of correlation between the grid and observed values. The BIAS, MB, RMSE, RE, and COR are calculated as

$$\mathrm{BIAS} = G_i - O_i \tag{1}$$

$$\mathrm{MB} = \frac{1}{N}\sum_{i=1}^{N}\left(G_i - O_i\right), \tag{2}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(G_i - O_i\right)^2}, \tag{3}$$

$$\mathrm{RE} = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|G_i - O_i\right|}{O_i} \tag{4}$$

$$\mathrm{COR} = \frac{\sum_{i=1}^{N}\left(G_i - \overline{G}\right)\left(O_i - \overline{O}\right)}{\sqrt{\sum_{i=1}^{N}\left(G_i - \overline{G}\right)^2}\sqrt{\sum_{i=1}^{N}\left(O_i - \overline{O}\right)^2}} \tag{5}$$

where O_i is the station observation value, G_i is the value obtained by interpolating the CMPAS products to stations and N is the total number of samples (number of stations).

The traditional score is based on the classification of observations and CMPAS_RT, as shown in Table 1.

Table 1. Precipitation dichotomy.

Type	OBSERVATIONS TRUE	OBSERVATIONS FALSE
CMPAS_RT TRUE	NA	NB
CMPAS_RT FALSE	NC	ND

The Threat Score (TS), Probability of Detection (POD), Missing Alarm Rate (MR), and False Alarm Rate (FAR) are calculated as:

$$\mathrm{TS} = \frac{\mathrm{NA}}{\mathrm{NA} + \mathrm{NB} + \mathrm{NC}} \tag{6}$$

$$\mathrm{POD} = \frac{\mathrm{NA}}{\mathrm{NA} + \mathrm{NC}} \tag{7}$$

$$\mathrm{MR} = \frac{\mathrm{NC}}{\mathrm{NA} + \mathrm{NC}} \tag{8}$$

$$\mathrm{FAR} = \frac{\mathrm{NB}}{\mathrm{NA} + \mathrm{NB}} \tag{9}$$

2.4 *Construction of the revised model*

2.4.1 DATA PREPROCESSING

The BIAS between the CMPAS_RT product and the station rain gauge data is used as the target value to participate in the machine learning correction.

As meteorological factors, temperature and wind speed are taken into consideration. Slope, slope variability, slope direction, and surface roughness extracted from the DEM data are topographic factors. Assume that there are i sets of variables associated with precipitation, and each set of variables has j factors, the meteorological and topographic factors are standardized (Chen et al. [32]) as

$$y_{ij} = \frac{x_{ij} - \overline{x_j}}{S_j} \tag{10}$$

where y_{ij} is the standardized factor value, x_{ij} is the original factor, $\overline{x_j}$ is the arithmetic mean of the jth factor, and S_j is the sample standard deviation.

All the standardized impact factors are divided into several independent principal components using Principal Component Analysis (Abdi and Williams [41]; Lasisi and Attoh-Okine [42]). The implementation of Principal Component Analysis in this study is based on the ‘scikit-learn’ of python language, and the main principles are as follows:

First, the contribution of the principal components to the precipitation results is calculated. Then, by computing the loadings between the impact factors and the principal components, the contribution of the impact factors to the precipitation results is analyzed. Finally, the principal components with a cumulative contribution of 90% are selected for the machine learning revision.

2.4.2 MODEL TRAINING AND VALIDATION

Parametric experiments on machine learning models are conducted using grid search (Bergstra and Bengio [43]) and k-fold cross-validation (Refaeilzadeh et al. [44]) methods. The grid search algorithm is a method to optimize model performance by traversing a given combination of parameters. The accuracy of each model for the test set is assessed for each pair of parameters, and the accuracy of each pair of parameters is compared through k-fold cross-validation to select the optimal parameters.

The whole training set data is averaged into k pieces. The remaining $k-1$ parts are used as the cross-validation training set, while the kth part is used as the validation set. The model is trained using the data set of k cases to produce k models under the current parameter settings, and the corresponding validation set is used to examine the prediction results of these k models to produce k correctness indicators, which are then averaged as the corresponding scores. The optimal parameters of the model are determined by scores.

2.5 *Machine learning methods*

Following the results of the experiments, three ensemble learning methods (Sagi and Rokach [45]) are chosen for revision. By establishing several models, ensemble learning solves the single prediction problem. Its working principle is to generate multiple classifiers or models that can independently learn and predict (Dong et al. [46]). These forecasts are eventually grouped into a combined forecast, which is superior to forecasting in any single category.

2.5.1 RANDOM FOREST REGRESSION

Random forests (Belgiu and Drăgut [47]) are used to resample multiple samples from the original sample and model a decision tree for each sample, and then average the predicted values of the multiple decision trees to obtain the final prediction results. First, the dataset is created by:

$$D = \left\{ \left(x_m, y_m \right), m = 1, 2 \cdots, n \right\} \tag{11}$$

where y_m is the bias between the value interpolated to the station for the CMPAS and the station rain gauge data, and x_m is the principal component.

The training dataset D_j is then drawn at random from a subset of the dataset. A random forest is created by repeatedly training N decision trees h_i, each of which is built using a random subspace partitioning approach, from which the best features are chosen for splitting. The average of each decision tree represents the projected outcome.

2.5.2 ADABOOST REGRESSION

AdaBoost (Cao et al. [48]; Rätsch et al. [49]) is an abbreviation for ‘Adaptive Boosting’. AdaBoost regression algorithm can be briefly described in three steps：

First, initialize weights. For dataset D (Eq. 8), if there are n samples, the weight for each sample x_m is initialized to $1/n$. D_1 is used for the training of the first weak learner h_1 and D_t is used for the training of the tth weak learner h_t.

Second, repeat the loop T times, recording the number of weak learners in each iteration as t, $t = 1, 2, 3, \cdots, T$. The weight distribution of the sample set x_t is changed after calculating the error rate of the learner h_t and updating the learner's current weight following the error size. In this approach, the entire training procedure is iterated.

Third, based on the learners' weight rankings after T rounds of iterations, the median weight learner is chosen as the outcome.

2.5.3 BAGGING REGRESSION

The Bootstrap aggregating, also known as the Bagging algorithm (Bauer and Kohavi [50]), serves as the foundation for more sophisticated algorithms like Random Forest. Data are put-back extracted from the original dataset D (Eq. 8). To get t Bootstrap resampling datasets, this is repeated t times. A weak learner is then obtained for each Bootstrap resampled dataset for a total of t weak learners for regression. The final result is calculated by integrating the t weak learners and taking the mean of these t weak learners.

3 RESULTS AND DISCUSSION

3.1 *Analysis of overall revision effect*

The overall evaluation scores before and after the revision are displayed in Table 2. The original TS score is 0.91, POD is 0.952, MR is 0.048, and FAR is 0.047, indicating that the model’s precipitation accuracy is already high. The three machine learning algorithms’ revised outcomes are comparable, with revised TS scores of 0.94, POD of 0.984, MR of 0.016, and FAR of 0.046. All indicators have improved to varying degrees, with the most notable decrease of 66% in MR.

The hourly precipitation is subdivided by class, and Table 3 shows that as the amount of precipitation rises, the RMSE also goes up. All of the revised RMSE decreased, with the Random Forest Regression revisions being the most successful in doing so for each class.

Table 2. TS, POD, MR, and FAR for CMPAS_RT and revised products.

Product	TS	POD	MR	FAR
CMPAS_RT	0.910	0.952	0.048	0.047
Random Forest Regression	0.940	0.984	0.016	0.046
AdaBoost Regression	0.941	0.984	0.016	0.045
Bagging Regression	0.941	0.984	0.016	0.046

Table 3. RMSE of different precipitation levels for CMPAS_RT and revised products.

Product	Precipitation (mm h^{-1})				
	0.1-1.9	2-4.9	5-9.9	10-19.9	≥20
CMPAS_RT	0.284	0.846	1.614	3.463	6.495
Random Forest Regression	0.274	0.828	1.597	3.414	6.239
AdaBoost Regression	0.277	0.834	1.614	3.420	6.275
Bagging Regression	0.281	0.833	1.605	3.456	6.299

3.2 *Analysis of different stations*

Rain gauge precipitation from 1899 automatic weather stations across Sichuan is statistically evaluated with CMPAS_RT and the revised products from three methods. Figs. 1-4 depict the spatial distribution of the indicators. The distribution of the number of stations for each indicator before and after the revision is also counted (Fig. 5).

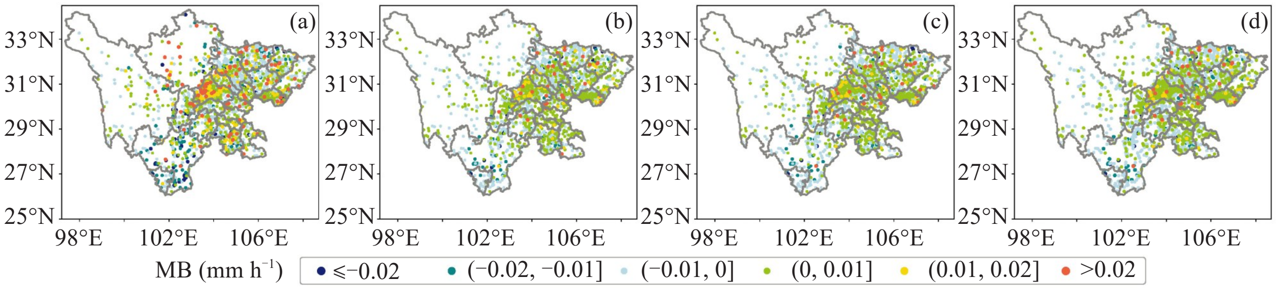

Figure 1. MB distribution of CMPAS_RT and revised products from each station (a: CMPAS_RT; b: Random Forest Regression; c: AdaBoost Regression; d: Bagging Regression).

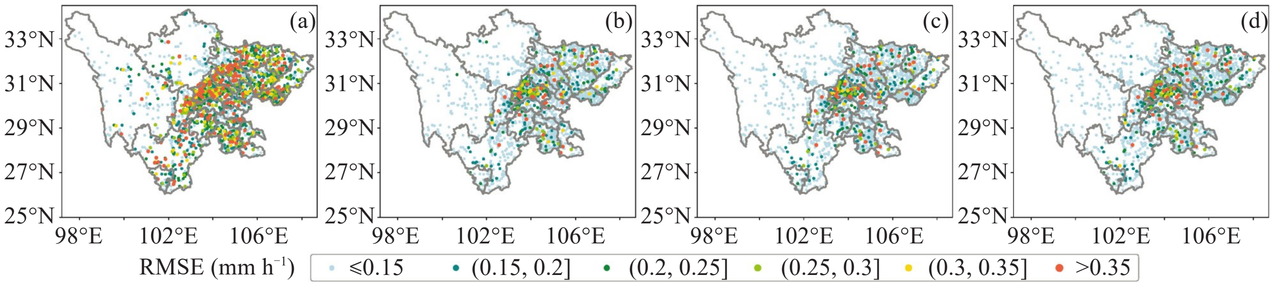

Figure 2. RMSE distribution of CMPAS_RT and revised products from each station (a: CMPAS_RT; b: Random Forest Regression; c: AdaBoost Regression; d: Bagging Regression).

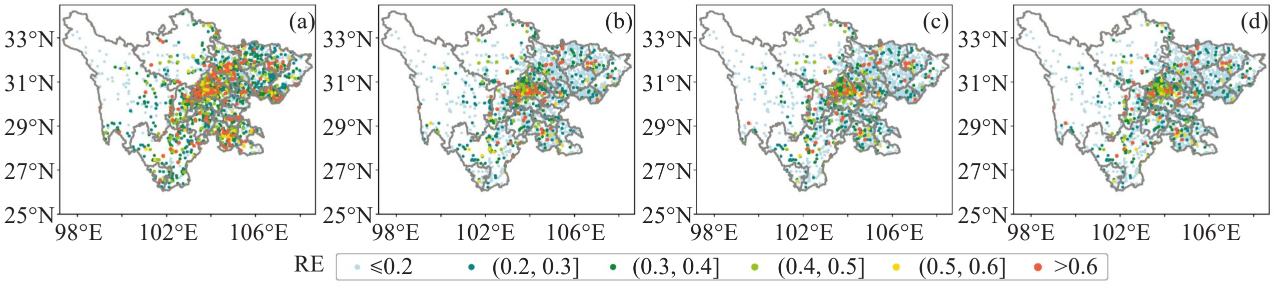

Figure 3. RE distribution of CMPAS_RT and revised products from each station. (a: CMPAS_RT; b: Random Forest Regression; c: AdaBoost Regression; d: Bagging Regression).

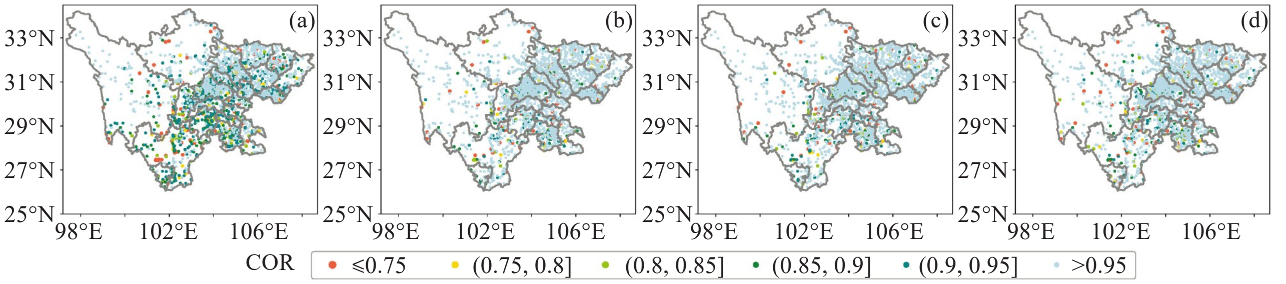

Figure 4. COR distribution of CMPAS_RT and revised products from each station (a: CMPAS_RT; b: Random Forest Regression; c: AdaBoost Regression; d: Bagging Regression).
(Source of map: Sichuan Bureau of Surveying, Mapping and Geoinformation. Map approval number: Chuan S[2021]00059, http://scsm.mnr.gov.cn/nbzdt.htm)

The overall revised results are similar for all three methods. In terms of MB, it is mainly concentrated in the range of −0.01−0.01mm h^{-1} in CMPAS_RT, with 980 stations having an MB of less than 0 mm h^{-1}, indicating that CMPAS_RT underestimates precipitation for most stations. The three methods greatly minimize the MB between the products and stations. Following the adjustment, the MB of roughly 800 stations is −0.005−0 mm h^{-1}, and that of nearly 600 stations is 0−0.005 mm h^{-1}. The number of stations with MB between −0.01 mm h^{-1} and 0.01 mm h^{-1} is reduced by about 30%. In overview, 95% of the stations are revised to reduce MB, and 85% of the stations reduce MB by more than 30%.

For the RMSE, the stations with a large RMSE in CMPAS_RT are mostly concentrated in the basin, with the RMSE of 939 stations concentrated in the range of

0.12 – 0.2 mm h^{-1}. The three approaches produce comparable spatial distributions of the RMSE corrected. The RMSE of roughly 900 stations is 0 – 0.04 mm h^{-1}, and it greatly reduces throughout the basin. After correction, the RMSE of 82% of the stations decreased, and the RMSE of 80% of the stations decreased by more than 20%. It demonstrates that the three revision techniques are superior for modifying discrete data.

The stations with large RE in CMPAS_RT are mostly concentrated within Sichuan Basin, with 54% of stations having an RE of 0.15 – 0.35. The revised RE within the basin is significantly decreased, with 547 stations having an RE of 0–0.05 and 50% having an RE of less than 0.1. In brief, a total of 77% of stations in Sichuan have lowered RE, with 62% having reduced RE by more than 20%.

Most of the stations in Sichuan have a COR greater than 0.85, indicating that CMPAS_RT is substantially connected with non-independent stations. The COR for the majority of the stations after the modification is 0.9 or higher, accounting for 70% of the total. Since the COR is already high before the correction, there is a minor rise after the revision.

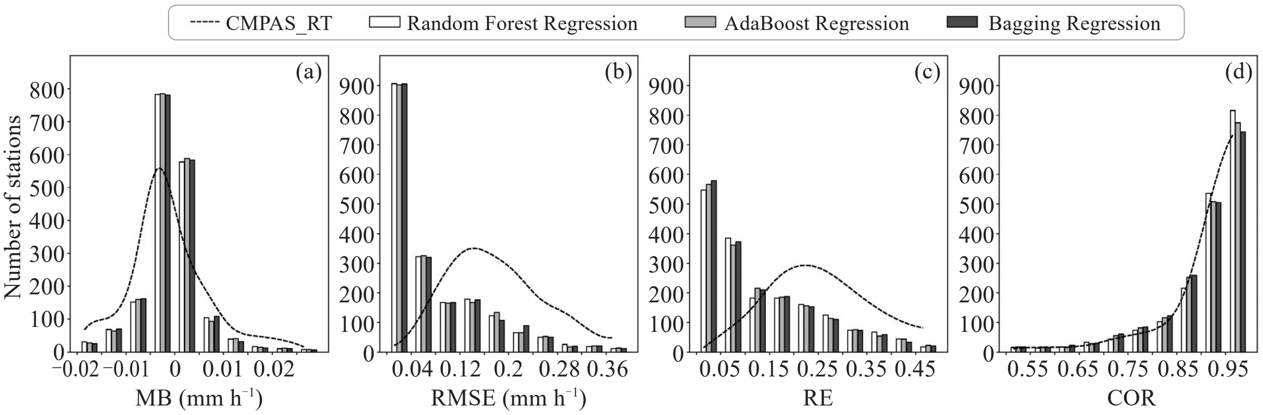

Figure 5. Comparison of indicators of CMPAS_RT with revised products (a: MB; b: RMSE; c: RE; d: COR).

3.3 *Analysis of different regions*

For analysis, Sichuan province is split into three regions: the Western Sichuan plateau, the Panxi Area, and the Sichuan Basin. The RE of CMPAS_RT and the revised products for different regions are shown in Table 4. Affected by the vast number of stations in the Sichuan Basin and the complicated station environment, the average RE of CMPAS_RT in the Sichuan Basin is 0.38; the RE of the Western Sichuan Plateau is the smallest, which is 0.266, and AdaBoost regression has the best correction impact on the RE of the whole Sichuan, which is 0.136 for the Western Sichuan Plateau, 0.184 and 0.169 for the Panxi Area and Sichuan Basin, respectively. The PDF of the RE change rate for the revised CMPAS_RT products is shown in Fig. 6a-6c for various regions, with the Western Sichuan Plateau showing the most obvious effects of the revision, where the RE is reduced by more than 90% for about 30% of the stations. Bagging regression performs best in the range of 70%–80% of RE reduction for the Panxi Area and the Sichuan Basin, and the proportion of stations in this range is the biggest, at 20% and 23%, respectively.

Table 5 shows the RMSE of CMPAS_RT and the revised products for different regions, which is greater in the Sichuan Basin than in the Panxi area or the Western Sichuan Plateau. With a reduction of 61% and 62% in the RMSE for the Panxi Area and the Sichuan Basin, respectively, Bagging Regression performed marginally better than the other approaches. The three machine learning techniques produced comparable outcomes for the Western Sichuan Plateau, showing a significant 77% reduction.

The PDF of the RMSE change rate is shown in Fig. 6d-6f. On the Western Sichuan Plateau, the RMSE is decreased by more than 90% at roughly 30% of the stations, and the Random Forest revision is ideal for this interval. Besides, the RMSE is reduced by 70%–80% at approximately 20% of the stations, and Adaboost regression is optimal in this interval. For the Panxi Area, 20% of the stations have a 70% – 80% reduction in RMSE, Bagging Regression worked best in the interval where the RMSE is decreased by 60% – 70%, while Random Forest Regression performs the best in the interval where the RMSE is dropped by 30%–60%. The most significant outcome of the correction is the reduction in RMSE by more than 70% at approximately 70% of the stations on the Western Sichuan Plateau.

The COR for various regions is illustrated in Table 6. Due to the reduced precipitation, the original COR is higher for the Western Sichuan Plateau at 0.925 and lower for the Panxi Area at 0.857. The PDF of the COR change rate for the revised products is displayed in Fig. 6g-6i, Random Forest has a good effect on revising the COR, more than 50% of the stations on the Western Sichuan Plateau have a rise of roughly 0–10% in COR, and a small number of stations in the Panxi Area see an increase of more than 10%. In the Sichuan basin, the COR varies less.

Table 4. RE for CMPAS_RT and revised products in different regions.

Region	RE			
	CMPAS_RT	Random Forest Regression	AdaBoost Regression	Bagging Regression
Western Sichuan Plateau	0.266	0.140	0.136	0.154
Panxi Area	0.357	0.184	0.184	0.193
Sichuan Basin	0.380	0.170	0.169	0.175

Table 5. RMSE for CMPAS_RT and revised products in different regions.

Region	RMSE (mm h^{-1})			
	CMPAS_RT	Random Forest Regression	AdaBoost Regression	Bagging Regression
Western Sichuan Plateau	0.141	0.032	0.032	0.033
Panxi Area	0.197	0.079	0.080	0.076
Sichuan Basin	0.241	0.093	0.093	0.091

Table 6. COR for CMPAS_RT and revised products in different regions.

Region	COR			
	CMPAS_RT	Random Forest Regression	AdaBoost Regression	Bagging Regression
Western Sichuan Plateau	0.925	0.941	0.938	0.934
Panxi Area	0.857	0.892	0.881	0.884
Sichuan Basin	0.917	0.921	0.920	0.918

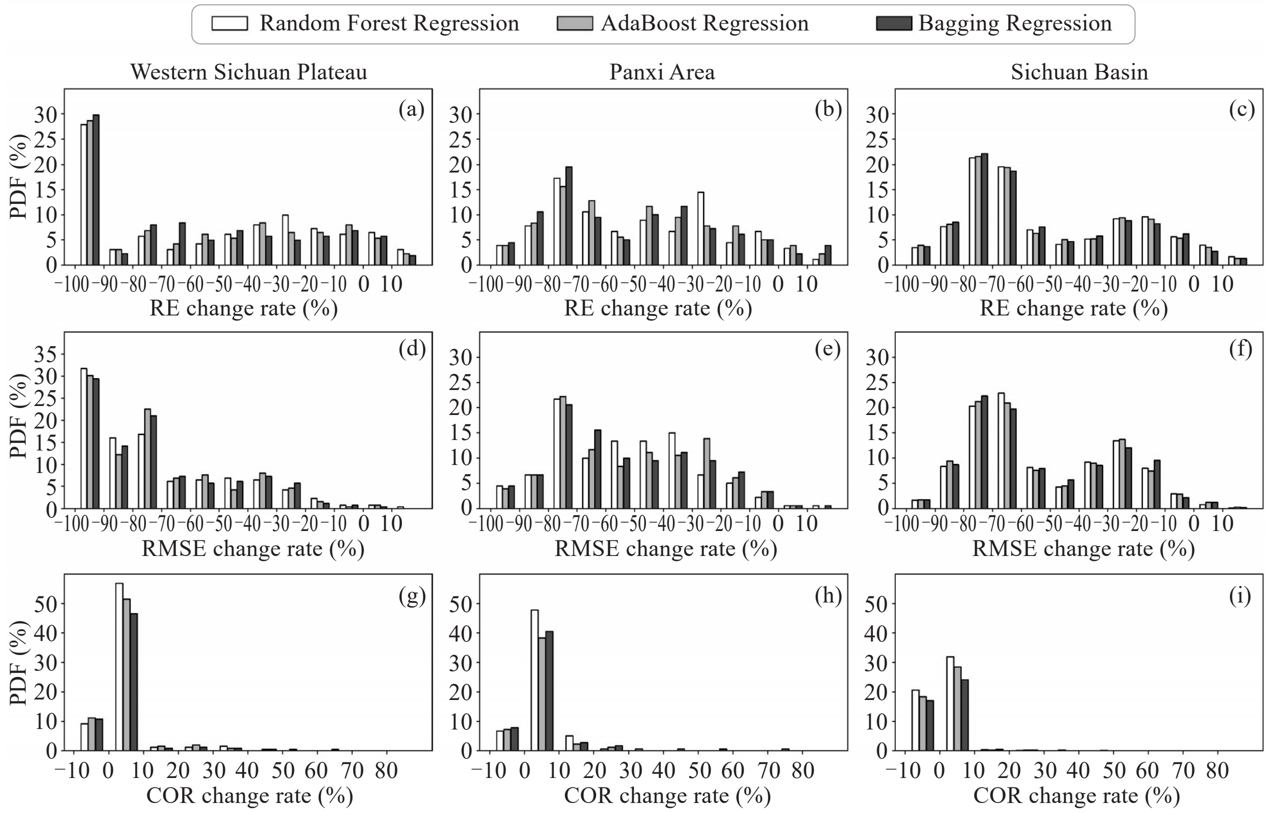

Figure 6. PDF of indicators' change rate of the revised CMPAS_RT products in different regions (a, b, c: RE; d, e, f: RMSE; g, h, i: COR).

3.4 *Analysis of different seasons*

The RE of CMPAS_RT and the revised product for each season are compared in Table 7, with CMPAS_RT having the biggest RE in summer (0.59) and the smallest RE in winter (0.208). Bagging regression performs somewhat better in spring, summer, and autumn, with similar results for the three methods of revision in winter. The average RE is lowered by 80% in winter, 62% in spring, 60% in summer, and 22% in autumn. The findings of the Random Forest revisions are chosen for further analysis since the three machine learning revision results are comparable for all seasons (Fig. 7). Fig. 7a-7c demonstrates that the autumn revision's RE decrease is lower, and for the Sichuan Basin, the RE reduction is mostly focused at around 20%.

The RMSE is presented in Table 8, with the maximum RMSE of 0.541 mm h^{-1} in summer and the smallest RMSE of 0.031 mm h^{-1} in winter. The overall effect of the three revision approaches is similar throughout all seasons, with a considerable reduction in RMSE. The biggest reduction in RMSE is recorded in winter (71%), spring (69%), summer (65%), and the least in autumn (31%).

The PDF of the RMSE change rate for the modified CMPAS_RT products for each region during different seasons is shown in Fig. 7d-7f. For the majority of stations and during all seasons, the RMSE drops by over 90% for the Western Sichuan Plateau. For stations in the Panxi Area, the reduction in RMSE is greater in the spring, summer, and winter months than in the autumn. The majority of the stations in the Sichuan Basin see a decrease of more than 90% in RMSE in spring and summer, a reduction concentrated in the range of 60% to 90% in winter, while the revision effect in autumn is limited, with a decline of 40% to 60% in RMSE.

The COR of CMPAS_RT and the revised products during various seasons are shown in Table 9. CMPAS_RT's COR is somewhat greater in summer at 0.935 and marginally lower in the fall at 0.862. The COR increases modestly in all seasons following the adjustment because it is already high before it, with an increase of 2.2% in autumn, 1.6% in winter, 1.2% in spring, and 0.8% in summer. Fig. 7g, 7h, and 7i displays the probability distribution of COR for the updated CMPAS_RT products during various seasons. With the majority of stations exhibiting a 0–5% increase in the revised COR change, the trend of the revised change in the three areas is generally similar across the four seasons.

In the Panxi Area and Sichuan Basin, the revised effect of CMPAS_RT is somewhat weaker for autumn. Reviewing the weather and climate profiles for October 2020 reveals that the majority of the basin had more than 15 days of precipitation, and a large percentage of the Panxi Area had between 10 and 18 days, all of which were higher than in the same period in the typical year. The machine learning method to construct the error relationship between precipitation products and observed values may be impacted by the month's high precipitation, leading to a marginally decreased accuracy.

Table 7. RE of CMPAS_RT and revised product for different seasons.

Season	RE			
	CMPAS_RT	Random Forest Regression	AdaBoost Regression	Bagging Regression
Spring	0.298	0.117	0.114	0.112
Summer	0.590	0.250	0.238	0.234
Autumn	0.285	0.221	0.226	0.222
Winter	0.208	0.040	0.040	0.041

Table 8. RMSE of CMPAS_RT and revised product for different seasons.

Season	RMSE (mm h^{-1})			
	CMPAS_RT	Random Forest Regression	AdaBoost Regression	Bagging Regression
Spring	0.123	0.038	0.038	0.038
Summer	0.541	0.190	0.193	0.194
Autumn	0.138	0.064	0.065	0.066
Winter	0.031	0.009	0.009	0.010

Table 9. COR of CMPAS_RT and revised product for different seasons.

Season	COR			
	CMPAS_RT	Random Forest Regression	AdaBoost Regression	Bagging Regression
Spring	0.930	0.941	0.938	0.932
Summer	0.935	0.943	0.940	0.940
Autumn	0.862	0.881	0.875	0.874
Winter	0.932	0.947	0.942	0.944

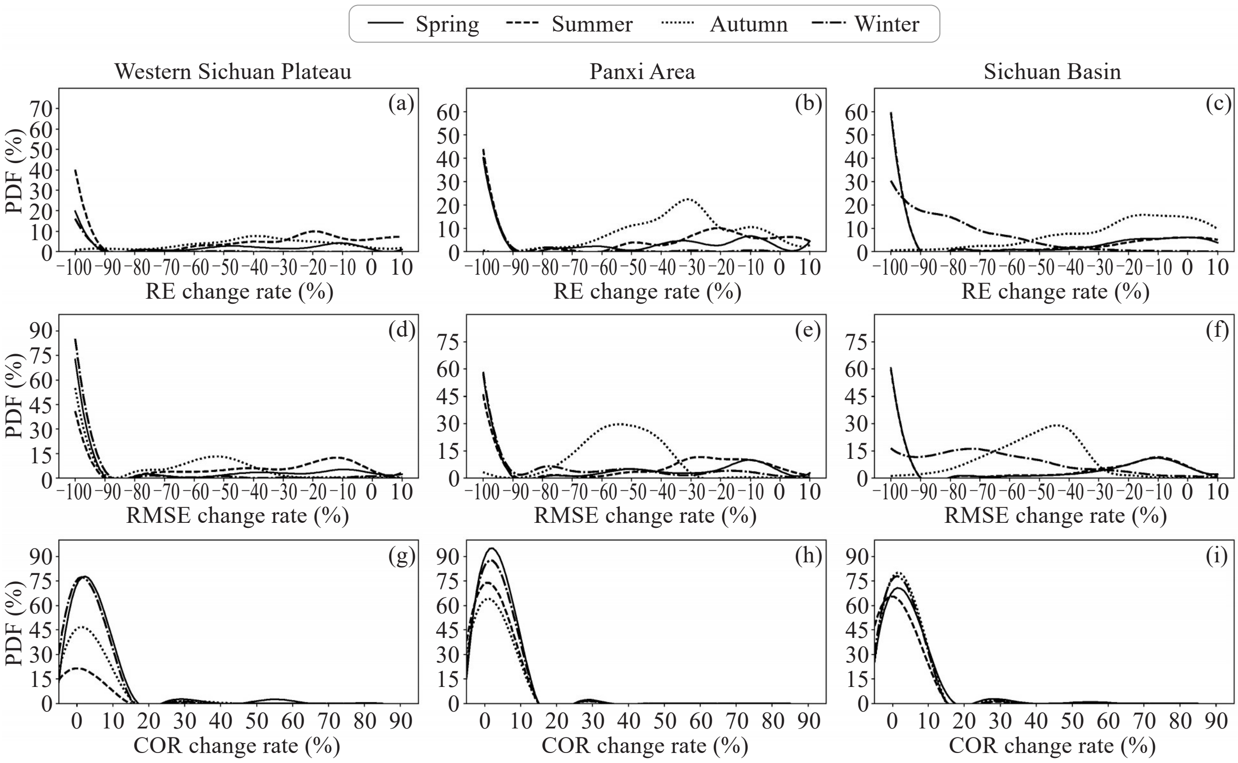

Figure 7. PDF of RE, RMSE, and COR change rate of the revised CMPAS_RT products in different seasons (a, b, c: RE; d, e, f: RMSE; g, h, i: COR).

3.5 *Analysis of heavy precipitation event*

On June 12-13, 2021, the northeastern part of the Sichuan Basin had heavy rainfall, with the center of the precipitation occurring there. The hourly rainfall intensity and cumulative rainfall throughout this process were strong. When the 24-hour cumulative precipitation from CMPAS_RT is compared to the rain gauge data (Fig. 8a and 8b), the CMPAS_RT precipitation area, direction, and rainband pattern are very comparable with the observations. Since the three machine learning techniques produce similar results, the updated random forest findings are chosen for in-depth investigation.

The MB of hourly rainfall before and after the revision is illustrated in Fig. 8c and Fig. 8d. The region with the highest MB prior to the adjustment is in the northeastern half of the basin's heavy precipitation zone where 18% of the stations have MB greater than 0.02 mm h^{-1} or less than −0.02 mm h^{-1}. With only 9% of the stations having a considerable MB, the MB inside the area of severe precipitation has significantly decreased since the revision.

For analysis, the top five stations in terms of 24-hour cumulative precipitation are chosen (Fig. 9). Four stations recorded 24-hour cumulative precipitation totals that are higher than that of CMPAS_RT, showing that CMPAS_RT does somewhat underestimate heavy precipitation. The MB is decreased to varying degrees after the revision, with a maximum reduction of 58%. The COR at the five stations also slightly increased following the correction.

3.6 *Analysis of independent data*

The data from 449 quality-controlled hydrological stations in July 2021 are selected for the independent analysis, and the station distribution is shown in Fig. 10a. Fig. 10b-10f displays a comparison of the indicators before and after the adjustment. The independent revision is most impacted by Bagging Regression. The range of MB variations after the revision significantly

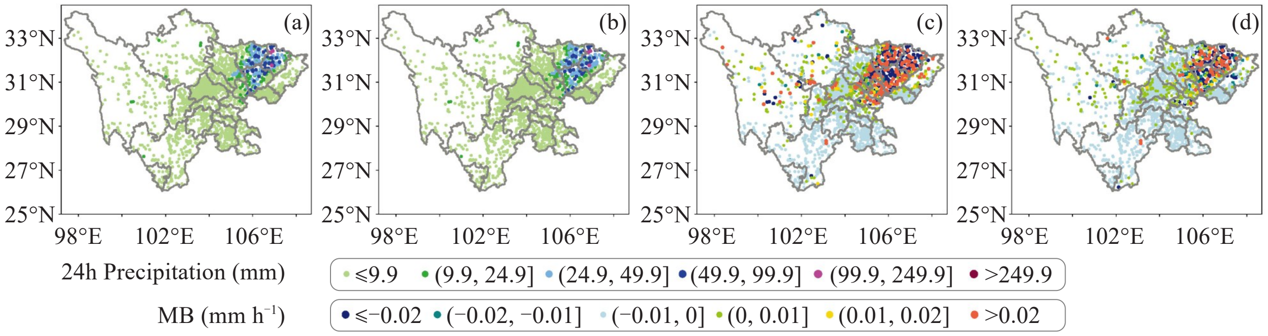

Figure 8. Spatial distribution of 24-hour cumulative precipitation and MB (a: 24-hour cumulative precipitation of weather stations; b: 24-hour cumulative precipitation of CMPAS_RT; c: MB distribution of CMPAS_RT; d: MB distribution of the revised product).

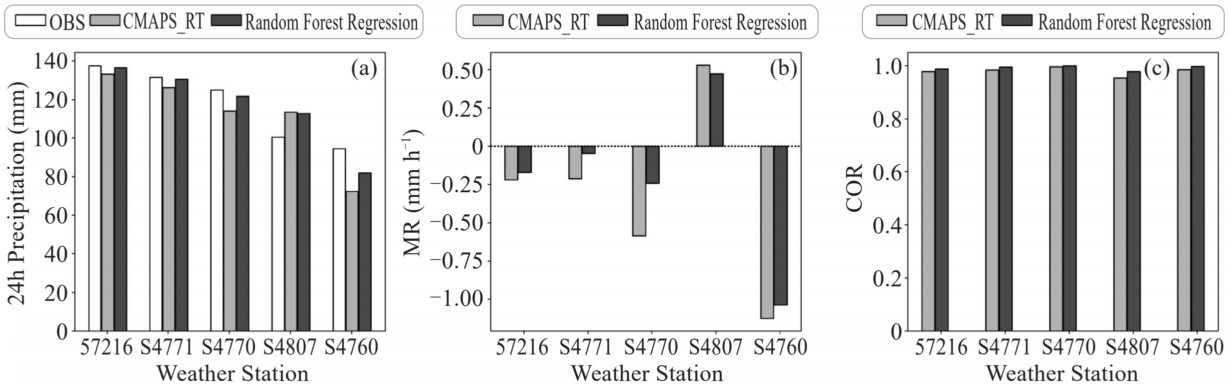

Figure 9. Revisions of different heavy precipitation stations (a: 24-hour cumulative precipitation; b: MB; c: COR).

decreased, although the median change in MB was slightly different. The median COR between hydrological stations and CMPAS_RT is slightly lower than that of meteorological stations, at 0.6, because the accuracy of precipitation data at hydrological stations is 0.5 mm and that of CMPAS_RT is 0.1 mm, which is subject to some error. The range of fluctuation of COR is dramatically reduced after the revision, with the median increasing to 0.63.

The majority of the CMPAS_RT and hydrological stations' TS and POD fall within the range of 0.7−1. Following the revision, the median has slightly increased, and the TS and POD ranges are primarily in the range of 0.8−1. The MR is largely between 0 and 0.3 before the adjustment, and between 0 and 0.2 after it, with less missing rate.

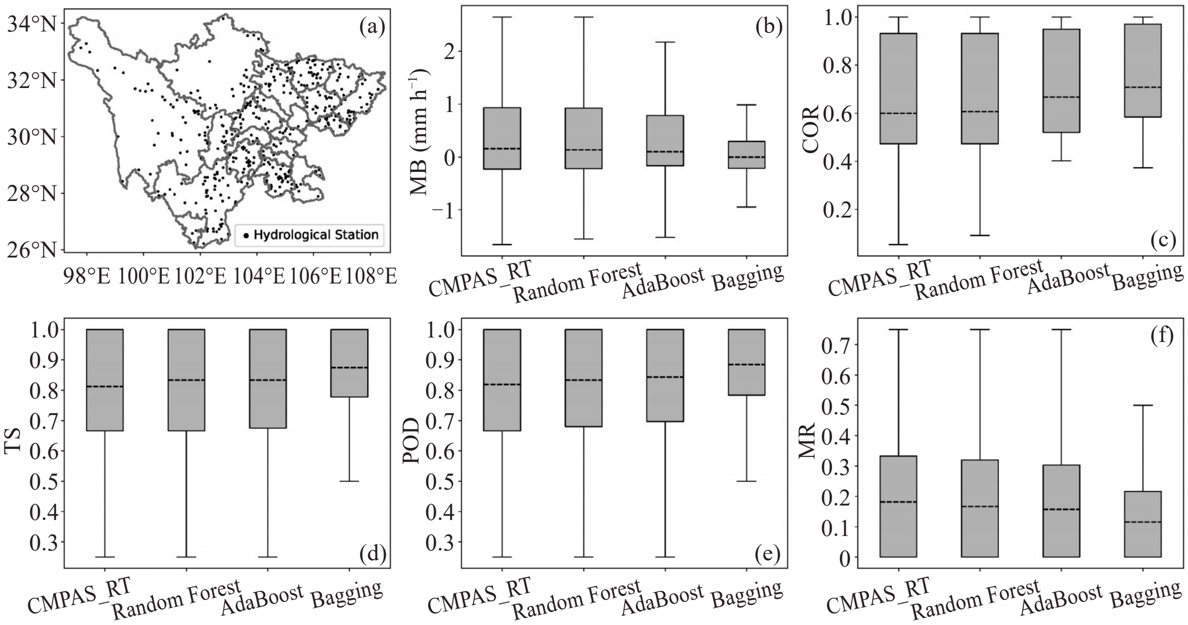

Figure 10. Revision effect of hydrological stations (a: Spatial distribution of hydrological stations; b: MB; c: COR; d: TS; e: POD; f: MR).

4 CONCLUSIONS

In this study, three ensemble learning approaches are chosen to adjust the CMPAS product by using high-precision DEM data to extract topographic features such as elevation, slope, slope direction, slope variability, and surface roughness as topographic factors, combined with precipitation-related air temperature and wind speed as meteorological factors. Different stations, different geographical regions, different seasons, and precipitation processes are used to study the corrective effect, and hydrological stations are introduced for an independent analysis.

The three machine learning methods' revised findings for Sichuan are comparable, showing varying degrees of improvement in each error indicator and the most notable improvement in MR, proving that the three machine learning techniques have a significant revision impact on the missing report situation.

The statistical analysis of the station revision's impact reveals that the revised MB, RMSE, and RE have all been greatly reduced, indicating that the machine learning approach is successful in revising discrete data. There is a slight increase in COR following the adjustment since they are already high previously. 85% of the automatic weather stations see an MB reduction of more than 30%, 80% of the stations see an RMSE drop of more than 20%, and the RE of 62% of the stations is decreased by over 20%. 70% of the stations have a revised COR of 0.9 or higher.

The Sichuan Basin, the Panxi Area, and the Western Sichuan Plateau are three separated zones for analysis. When it comes to reducing the RMSE for the Panxi Area and the Sichuan Basin, Bagging regression performs marginally better than the other two machine learning techniques. The results are comparable for all three approaches in the Western Sichuan Plateau; the RMSE of CMPAS_RT is decreased by 77%, making it the best adjustment among the three regions, and more than 70% of stations experience this reduction. Adaboost regression is the most effective in revising the RE in all three regions, and the revision remains especially successful in the Western Sichuan Plateau where the RE is reduced by far more than 90% in roughly 30% of the stations. Random Forest Regression has better results in revising the COR, with over 50% of the stations on the Western Sichuan Plateau increasing the COR by around 0−10%.

The findings of the seasonal revisions demonstrate that the overall revisions are similar for all three methods, with considerable reductions in both RMSE and RE. Winter sees the greatest decrease in the RMSE and RE, followed in turn by spring, summer, and autumn. The four seasons are broadly consistent with the corrected trends in COR, with the majority of stations indicating an increase of 0−5%.

Regarding precipitation processes, the CMPAS understates heavy precipitation in the non-independent evaluation. The revised CMPAS is close to the rain gauge precipitation to varying degrees, and the revised MB within the heavy precipitation zone has been greatly reduced. The median of each indicator has improved marginally and the range of fluctuation of the indicators has greatly decreased when using the hydrological stations for independent evaluation.

To review, the ensemble learning model has a significant revision effect in revising the CMPAS precipitation products by using topographic and meteorological factors, with the most considerable correction outcome in the Western Sichuan Plateau, and the less the precipitation, the better the revision outcome, and the accuracy of the heavy precipitation process is also improved to some extent. The error fluctuation range of the revised CMPAS is substantially smaller for independent stations.

A limitation of this study is that there are few stations in complex terrain and the data is susceptible to the surrounding environment. Furthermore, station maintenance is more difficult. Therefore, there is some compromise with the reliability and quality of the observed precipitation data. Further research is needed to develop a more accurate and efficient revised model to enhance the accuracy of CMPAS.

REFERENCES

[1] HARRISON D L, SCOVELL R W, KITCHEN M. High-resolution precipitation estimates for hydrological uses [J]. Water Management, 2009, 162(2): 125-135, https://doi.org/10.1680/wama.2009.162.2.125

[2] TURK F J, ARKIN P, EBERT E E, et al. Evaluating High-Resolution Precipitation Products [J]. Bulletin of the American Meteorological Society, 2008, 89(12): 1911-1916, https://doi.org/10.1175/2008BAMS2652.1

[3] LAGASIO M, PARODI A, PULVIRENTI L, et al. A synergistic use of a high-resolution Numerical Weather Prediction model and high-resolution Earth Observation products to improve precipitation forecast [J]. Remote Sensing, 2019, 11(20): 2387, https://doi. org / 10.3390 / rs11202387

[4] HIRABAYASHI Y, KANAE S, EMORI S, et al. Global projections of changing risks of floods and droughts in a changing climate [J]. Hydrological Sciences Journal , 2008, 53(4): 754-772, https://doi.org/10.1623/hysj.53.4.754

[5] NIKOLOPOULOS E I, BARTSOTAS N S, ANAGNOSTOU E N, et al. Using high-resolution numerical weather forecasts to improve remotely sensed rainfall estimates: The case of the 2013 colorado flash flood [J]. Journal of Hydrometeorology, 2015, 16(4): 1742-1751, https://doi.org/10.1175/JHM-D-14-0207.1

[6] SHEN Y, XIONG A Y, WANG Y, et al. Performance of high-resolution satellite precipitation products over China [J]. Journal of Geophysical Research: Atmospheres, 2010, 115(D2): D02114, https://doi.org/10.1029/2009JD012097

[7] HONG Y, HSU K L, SOROOSHIAN S, et al. Precipitation Estimation from Remotely Sensed Imagery using an Artificial Neural Network Cloud Classification System [J].

Journal of Applied Meteorology and Climatology, 2004, 43 (12): 1834-1853, https://doi.org/10.1175/JAM2173.1

[8] 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]. Journal of Hydrometeorology, 2007, 8(1): 38-55,https://doi.org/10.1175/JHM560.1

[9] SIMOLO C, BRUNETTI M, MAUGERI M, et al. Improving estimation of missing values in daily precipitation series by a probability density function-preserving approach [J]. International Journal of Climatology, 2010, 30(10): 1564-1576, https://doi. org / 10.1002/joc.1992

[10] CHEN M Y, KUMAR A. Influence of ENSO SSTs on the spread of the probability density function for precipitation and land surface temperature [J]. Climate Dynamics, 2015, 45(3-4): 965-974, https://doi. org / 10.1007 / s00382-014-2336-9

[11] PAN Y, SHEN Y, YU J J, et al. An experiment of high-resolution gauge-radar-satellite combined precipitation retrieval based on the Bayesian merging method [J]. Acta Meteorologica Sinica, 2015, 73(1): 177-186 (in Chinese), https://doi.org/10.11676/qxxb2015.010

[12] SHEN Y, HONG Z, PAN Y, et al. China's 1 km Merged Gauge, Radar and Satellite Experimental Precipitation Dataset [J]. Remote Sensing, 2018, 10(2): 264, https://doi.org/10.3390/rs10020264

[13] PAN Y, SHEN Y, YU J J, et al. Analysis of the combined gauge-satellite hourly precipitation over China based on the OI technique [J]. Acta Meteorologica Sinica, 2012, 70 (6): 1381-1389 (in Chinese), https://doi. org / 10.11676 / qxxb2012.116

[14] SHEN Y, ZHAO P, PAN Y, et al. A high spatiotemporal gauge-satellite merged precipitation analysis over China [J]. Journal of Geophysical Research: Atmospheres, 2014, 119(6): 3063-3075, https://doi.org/10.1002/2013JD020686

[15] PAN Y, GU J X, YU J J, et al. Test of merging methods for multi-source observed precipitation products at high resolution over China [J]. Acta Meteorologica Sinica, 2018, 76(5): 755-766 (in Chinese), https://doi. org / 10.11676/qxxb2018.034

[16] SHI C X, PAN Y, GU J X, et al. A review of multi-source meteorological data fusion products [J]. Acta Meteorologica Sinica, 2019, 77(4): 774-783 (in Chinese), https://doi.org/10.11676/qxxb2019.043

[17] PAN Y, GU J X, XU B, et al. Advances in multi-source precipitation merging research [J]. Advances in Meteorological Science and Technology, 2018, 8(1): 143-152 (in Chinese), https://doi. org / 10.3969 / j. issn. 2095-1973.2018.01.019

[18] HAN S, SHI C X, JIANG Z W, et al. Development and progress of High Resolution CMA Land Surface Data Assimilation System [J]. Advances in Meteorological Science and Technology, 2018, 8(1): 102-108 (in Chinese), https://doi. org / 10.3969 / j. issn. 2095-1973.2018.01.013

[19] TIE, R A, SHI C X, WAN G, et al. CLDASSD: Reconstructing fine textures of temperature field using super-resolution technology [J]. Advances in Atmospheric Sciences, 2022, 39(1): 117-130, https://doi. org / 10.1007 / s00376-021-0438-y

[20] LI S Y, HUANG X L, WU W, et al. Evaluation of CMPAS precipitation products over Sichuan, China [J]. Atmospheric and Oceanic Science Letters, 2022, 15(2): 1674-2834, https://doi.org/10.1016/j.aosl.2021.100129

[21] WU W, HUANG X L, XU X L, et al. Application assessment of merged precipitation analysis products in Sichuan Province [J]. Desert and Oasis Meteorology, 2021, 15(4): 1-8 (in Chinese), https://doi. org / 10.12057 / j. issn.1002-0799.2021.04.001

[22] PANG Z H, SHI C X, GU J X, et al. Assessment of a gauge-radar-satellite merged hourly precipitation product for accurately monitoring the characteristics of the super-strong Meiyu precipitation over the Yangtze River basin in 2020 [J]. Remote Sensing, 2021, 13(19): 3850, https://doi.org/10.3390/rs13193850

[23] BAI L, WEN Y Q, SHI C X, et al. Which precipitation product works best in the Qinghai-Tibet Plateau, multi-source blended data, global / regional reanalysis data, or satellite retrieved precipitation data? [J]. Remote Sensing, 2020, 12(4): 683, https://doi.org/10.3390/rs12040683

[24] GUO X, LONG K J, FAN J L, et al. Comparative assessment of four merged precipitation products in a sustained heavy rainfall process in Sichuan [J]. Plateau and Mountain Meteorology Research, 2021, 41(2): 42-52 (in Chinese), https://doi. org / 10.3969 / j. issn. 1674-2184.2021.02.005

[25] YU J J, SHEN Y, PAN Y, et al. Improvement of satellite-based precipitation estimates over China based on probability density function matching method [J]. Journal of Applied Meteorological Science, 2013, 24(5): 544-553 (in Chinese), https://doi. org/ 10.3969 / j. issn. 1001-7313.2013.05.004

[26] WANG Y D, NAN Z T, CHEN H, et al. Correction of CMORPH daily precipitation data over the Qinghai-Tibetan Plateau with K-Nearest Neighbor model [J]. Remote Sensing Technology and Application, 2016, 31 (3): 607-616 (in Chinese), https://doi. org / 10.11873 / j. issn.1004-0323.2016.3.0607

[27] JIA S F, ZHU W B, LŰ A F, et al. A statistical spatial downscaling algorithm of TRMM precipitation based on NDVI and DEM in the Qaidam Basin of China [J]. Remote Sensing of Environment, 2011, 115(12): 3069-3079, https://doi.org/10.1016/j.rse.2011.06.009

[28] LIU Y Q, FAN G Z, ZHOU D W, et al. Variability of NDVI in winter and spring on the tibetan plateau and their relationship with summer precipitation [J]. Acta Meteorologica Sinica, 2007, 65(6): 959-967 (in Chinese), https://doi.org/10.11676/qxxb2007.090

[29] BEUCHAT X , SCHAEFLI B , SOUTTER M, et al. Toward a robust method for subdaily rainfall downscaling from daily data [J]. Water Resources Research, 2011, 47 (9): W09524, https://doi.org/10.1029/2010WR010342

[30] HUANG M M, LIN R S, HUANG S, et al. A novel approach for precipitation forecast via improved K-nearest neighbor algorithm [J]. Advanced Engineering Informatics, 2017, 33: 89-95, https://doi. org / 10.1016 / j. aei.2017.05.003

[31] YANG Z D, LIU P, YANG Y. Convective / stratiform precipitation classification using ground-based Doppler radar data based on the K-nearest neighbor algorithm [J]. Remote Sensing, 2019, 11(19): 2277, https://doi. org / 10.3390/rs11192277

[32] CHEN H, NING C, NAN Z T, et al. Correction of the

daily precipitation data over the Tibetan Plateau with machine learning models [J]. Journal of Glaciology and Geocryology, 2017, 9(3): 583-592 (in Chinese), https://doi.org/10.7522/j.issn.1000-0240.2017.0065

[33] XIE Y Y, WANG J J. Preliminary study on the deviation and cause of precipitation prediction of GRAPES kilometer scale model in southwest complex terrain area [J]. Acta Meteorologica Sinica, 2021, 79(5): 732-749 (in Chinese), https://doi.org/10.11676/qxxb2021.053

[34] LU X N, HONG J, WANG L L, et al. Drought risk assessment in complex landform area [J]. Transactions of the Chinese Society of Agricultural Engineering, 2015, 31 (1): 162-169 (in Chinese), https://doi. org / 10.3969 / j. issn.1002-6819.2015.01.023

[35] HUANG X L, XU X L, WU W, et al. Analysis of terrain characteristics of meteorological stations in Sichuan Province based on DEM [J]. Plateau and Mountain Meteorology Research, 2022, 42(1): 135-142 (in Chinese), https://doi. org / 10.3969 / j. issn. 1674-2184.2022.01.019

[36] LUO Y, CHEN C, ZHANG T Y, et al. Analysis on the characteristics of atmospheric self-cleaning ability index in Sichuan Province from 1981 to 2017 [J]. China Environmental Science, 2021, 41(2): 527-536 (in Chinese), https://doi. org / 10.19674 / j. cnki. issn1000-6923.2021.0059

[37] ZENG B, CHEN Y, WANG Q, et al. Temporal and spatial characteristics of different classes and various durations of precipitation in Sichuan Province from 1961 to 2016 [J]. Journal of Glaciology and Geocryology, 2019, 41(2): 444-456 (in Chinese), https://doi.org/10.7522/j.issn.1000-0240.2019.0012

[38] ZHANG P Z. A review on active tectonics and deep crustal processes of the Western Sichuan region, eastern margin of the Tibetan Plateau [J]. Tectonophysics, 2013, 584: 7-22, https://doi.org/10.1016/j.tecto.2012.02.021

[39] CHEN Y, XIE S D. Temporal and spatial visibility trends in the Sichuan Basin, China, 1973 to 2010 [J]. Atmospheric Research, 2012, 112: 25-34, https://doi.org/10.1016/j.atmosres.2012.04.009

[40] LI P, CHEN T T, LIU S Q. Spatiotemporal dynamics and drivers of farmland changes in Panxi Mountainous Region, China [J]. Sustainability, 2016, 8(11): 1209, https://doi.org/10.3390/su8111209

[41] ABDI H, WILLIAMS L J. Principal component analysis [J]. Wiley Interdisciplinary Reviews: Computational statistics, 2010, 2(4): 433-459, https://doi. org / 10.1002 / wics.101

[42] LASISI A, ATTOH-OKINE N. Principal components analysis and track quality index: A machine learning approach [J]. Transportation Research Part C: Emerging Technologies, 2018, 91: 230-248, https://doi.org/10.1016/j.trc.2018.04.001

[43] BERGSTRA J, BENGIO Y. Random search for hyper-parameter optimization [J]. Journal of Machine Learning Research, 2012, 13(1): 281-305.

[44] REFAEILZADEH P, TANG L, LIU H, et al. Cross-validation [M]// LIU L, ÖZSU M T (eds), Encyclopedia of Database Systems. New York: Springer, 2009: 532-538.

[45] SAGI O, ROKACH L. Ensemble learning: A survey [J]. WIREs Data Mining and Knowledge Discovery, 2018, 8 (4): e1249, https://doi.org/10.1002/widm.1249

[46] DONG X B, YU Z W, CAO W M, et al. A survey on ensemble learning [J]. Frontiers of Computer Science, 2020, 14(2): 241-258, https://doi.org/10.1007/s11704-019-8208-z

[47] BELGIU M, DRĂGUŢ L. Random forest in remote sensing: A review of applications and future directions [J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2016, 114: 24-31, https://doi. org / 10.1016 / j. isprsjprs.2016.01.011

[48] CAO Y, MIAO Q G, LIU J C, et al. Advance and prospects of AdaBoost algorithm [J]. Acta Automatica Sinica, 2013, 39(6): 745-758, https://doi. org / 10.1016 / S1874-1029(13)60052-X

[49] RÄTSCH G, ONODA T, MÜLLER K R. Soft Margins for AdaBoost [J]. Machine Learning, 2001, 42(3): 287-320, https://doi.org/10.1023/A:1007618119488

[50] BAUER E, KOHAVI R. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants [J]. Machine Learning, 1999, 36(1): 105-139, https://doi.org/10.1023/A:1007515423169

Citation: LI Shi-ying, HUANG Xiao-long, WU Wei, et al. Correction of CMPAS Precipitation Products over Complex Terrain Areas with Machine Learning Models [J]. Journal of Tropical Meteorology, 2023, 29(2): 264-276, https://doi.org/10.46267/j.1006-8775.2023.020