Research on Machine Learning-Based Multi-source Precipitation Data Fusion

Hengliang Guo[1], Yu Fu[2], Yaohuan Yang[3], Yuanyuan Yue[2], Menggang Kou[2],

Wenyu Zhang[4, 5*]

[1]National Supercomputing Center in Zhengzhou, Zhengzhou, 450052
[2]School of Computer and Artificial Intelligence of ZZU, Zhengzhou, 450001
[3]School of Information Engineering of ZZU, Zhengzhou, 450001
[4]School of Geo-Science & Technology of ZZU, Zhengzhou, 450001
[5]College of Atmospheric Sciences of Lanzhou University, Lanzhou, 730000
{f_author}guohengliang@zzu.edu.cn,{s_author}824750862@qq.com,{t_author}yangyaohuan@zzu.edu.cn,
*{for_author}2369430089@qq.com, {fif_author}koukou8090@163.com, {six_author}zhangwy@zzu.edu.cn**

Abstract

With the development of artificial intelligence (AI) in recent years, meteorological departments have also begun to improve algorithms and revise short-term forecasts via AI, expecting to timely capture meteorological clues in massive weather data, to "prevent meteorological disasters", and "calculate precipitation faster and more accurately". At present, AI has been initially applied to the meteorological field, especially to the analysis of massive meteorological data. For instance, the AI-based data analysis technology can rapidly judge the cloud type and the meteorological prototype in satellite images. The AI-based data fusion technology contributes to more three-dimensional and refined atmosphere data, which improves the temporal and spatial resolutions of precipitation data. If the big data in AI are used to analyze typhoons and identify the typhoon track and source, the errors resulting from the naked-eye observation of images by meteorologists can be avoided, thus considerably improving the scientificity and accuracy of weather forecasts. During data fusion, the severe convective weather characteristics reflected by massive historical precipitation data can be learned through machine learning methods to predict the evolution trend of disastrous weather within the future 1 to 2 h. Furthermore, precipitation data errors are corrected through AI data analysis, and a daily precipitation fusion dataset with a spatial resolution of 1 km is obtained.

Keywords: *Artificial intelligence, data analysis, machine learning, Gaussian process regression*

1. INTRODUCTION

As the main driving force of the global water cycle, precipitation plays a critical role in the substance-energy relation [3] [6]. Precipitation data is the most important data basis for hydrological, meteorological, and ecological studies [18]. Precipitation data are usually subjected to low accuracy and spatial resolution, which, if uncertain, will lead to uncertainties in the final output. Hence, it is of great importance to improve the temporal accuracy and spatial resolution of precipitation data in various fields such as hydrology, weather, and ecology [2].

Precipitation data is mainly derived from ground observation data and radar and satellite-derived precipitation data. The precipitation observation, which can only represent the precipitation features within a certain range, is easily affected the nonuniform distribution of ground observation stations. The satellite retrieval of precipitation data integrates the merits of a large scale and a high temporal-spatial resolution, but the physical principles and algorithms of precipitation satellite retrieval are limitations, which result in a low satellite retrieval accuracy [8]. Radar-based precipitation estimation can extract the precipitation value at the highest spatial resolution, but its accuracy can be easily affected by complex geographical environments [17].

To improve the accuracy and temporal-spatial resolution of precipitation data, precipitation data fusion has been extensively explored at both home (China) and abroad in recent years. Meanwhile, some multi-source

B. Fox et al. (Eds.): IC-ICAIE 2022, AHCS 9, pp. 690-696, 2023.
https://doi.org/10.2991/978-94-6463-040-4_105

data analysis and fusion methods are prospering [12] [13] [14]. In the aspect of ground-satellite precipitation data fusion, scholars have explored different precipitation fusion models specific to different fields, including geographically weighted regression [7], optimum interpolation method [11], Kalman filtering [4], Bayesian estimation [5], probability density function (PDF) matching [15], and machine learning method [4]. To improve the spatial resolution while not affecting the accuracy, not a few experts have fused ground radar-based precipitation estimation based on ground-satellite technology fusion and proposed a basic idea of ground-satellite-radar three-source precipitation combination. First, the bias between radar and satellite-derived precipitation data is calibrated using the PDF method. Then, the optimal initial fields between ground radar and satellite-derived precipitation are fused using the BMA technology. Finally, ground observation data are integrated through the OI technology [9].

Although the multi-source precipitation data fusion has achieved certain progress, the fusion of IMERG satellite-derived daily-resolution precipitation data has been less investigated. In addition, the traditional precipitation data fusion method is not necessarily applicable to fuse mass data. The machine learning method integrates the advantages of overfitting prevention, high calculation efficiency, and accurate loss calculation [10]. However, the fusion of multi-source precipitation data based on machine learning algorithms has been less explored. Given this, the hourly precipitation data from 241 observation stations on Qilian Mountain during 2019-2020 were mainly used and analyzed together with such auxiliary variables as radar-derived precipitation and landform data. Then, the observed precipitation data was associated with precipitation and geomorphic factors in radars, thus forming a multi-source precipitation data fusion model based on Gaussian process regression [10] to improve the precipitation prediction accuracy in complex geomorphological regions.

2. RESEARCH AREA AND DATA

In this research, the precipitation data were derived from the measured hourly precipitation data of 241 observation stations on different underlying surfaces in Qilian Mountain during 2019-2020. Such observation stations included national stations, regional stations, and field stations. Satellite data came from the dataset (http://pmm.nasa.gov/data-access/downloads/) released by GFSC, with a spatial resolution of 0.25°×0.25° and a temporal resolution of 1 d. Radar-derived precipitation data were obtained from the radar-based quantitative precipitation estimation provided by Qilian Mountain, with a spatial resolution of 3 km and a temporal resolution of 1 h.

Auxiliary topographic parameters were selected from the SRTM [1] v41 fragmented data provided by the geographical spatial data cloud of the Chinese Academy of Sciences (CAS) [16]. The SRTM v41 fragmented data were subjected to format conversion, splicing, and clipping via Python to obtain a DEM data model with a spatial resolution of 90 m. The DEM data model of Qilian Mountain is shown in Figure 1.

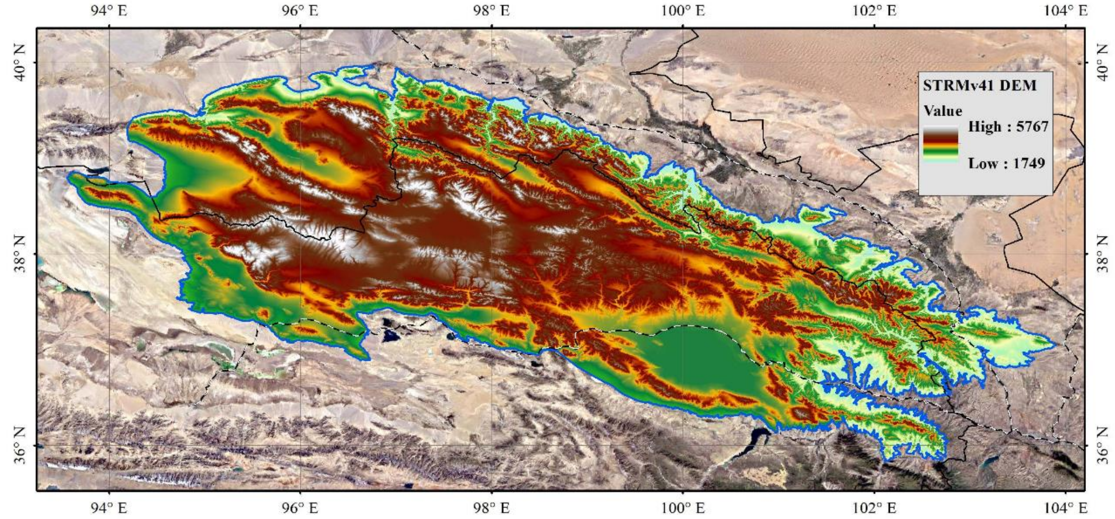

Figure 1: DEM Data Model of Qilian Mountain

3. METHODOLOGY

3.1 Multi-source precipitation data fusion model based on Gaussian process regression

Gaussian process regression is applied to computer mathematics on a certain theoretical basis. In this research, the linear kernel of linear Bayesian process regression was replaced by a kernel variable through the linear Bayesian regression technology, thus endowing the Gaussian process regression technology with favorable applicability to solving complex problems like high dimensions, small sample size, and high uncertainties.

As for the layout flow of complex function methods described by the Gaussian process from the spatial perspective of functions, its attribute is decided by the mean value function $h(x)$ and covariance function $G(x,x')$, which are expressed in the following forms, respectively:

$$h(x) = E[f(x)] \quad (1)$$

$$G(x,x') = E\big[\big(f(x) - h(x)\big)\big(f(x') - h(x')\big)\big] \quad (2)$$

where x and x' refer to random variables of R^d. GP is defined as $f(x) - GP\big(h(x), G(x,x')\big)$.

The following assumption is made for the Gaussian process:

$$\mathrm{y} = \mathrm{f(x)} + \varepsilon, \text{and } \varepsilon \sim \mathrm{N}(0, \sigma_\mathrm{n}^2) \quad (3)$$

where x is an input variable, y denotes the observation value subjected to noise pollution, f represents GP-predicted function value, and ε stands for noise.

The prior distribution of observation value y is as below:

$$y \sim N(0, K(X,X) + \sigma_n^2 I_n) \quad (4)$$

The joint prior distribution of observation values y and f_* is expressed as follows:

$$\begin{bmatrix} y \\ f_* \end{bmatrix} \sim N \left\langle 0, \begin{bmatrix} K(X,X) + \sigma_n^2 I_n & K(X,x_*) \\ K(x_*,X) & K(x_*,x_*) \end{bmatrix} \right\rangle \quad (5)$$

where $K(X,X) = K_\mathrm{n} = \big(K_{ij}\big)$ stands for a $n \times n$-order positive definite covariance matrix, and the element $K_{ij} = K\big(x_i, x_j\big)$ represents the measure of x_i - x_j correlation; $K(X,x_*) = K(x_*,X)^T$ is a covariance matrix between the input training sample sets X and x_*; I_n is a unit matrix.

Through the above equation, the prior probability density distribution of the predicted value f_* is as follows:

$$\mathrm{f}_* \mid \mathrm{X, y, x_*} \sim \mathrm{N}\left(\overline{\mathrm{f}_*}, \mathrm{cov(f_*)}\right) \quad (6)$$

where

$$\overline{f_*} = K(x_*,X)[K(X,X) + \sigma_n^2 I_n] - 1 \quad (7)$$

$$\begin{aligned} cov(f_*) &= K(x_*,x_*) \\ &- K(x_*,X)[K(X,X) \\ &+ \sigma_n^2 I_n]^{-1} K(X,x_*) \end{aligned} \quad (8)$$

The covariance function most extensively selected in the Gaussian process is a square exponential covariance function.

$$K(x,x') = \sigma_f^2 exp\big(-(x - x')^T M^{-1}(x - x')\big) \quad (9)$$

To determine the Gaussian process model, the hyperparameter set $\theta = \{M, \sigma_f^2, \sigma_n^2\}$ should be solved, which is usually estimated using the maximum likelihood method for hyperparameter training.

3.2 Accuracy evaluation

The evaluation indexes used in this research included correlation coefficient R, ot-mean-square error, and $Bias$, are expressed as follows:

$$R = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(O_i - \bar{O})}{\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}\sqrt{\sum_{i=1}^{n}(O_i - \bar{O})^2}} \quad (10)$$

$$RMSE = \sqrt{\sum_{i=1}^{n}(Y_i - O_i)^2 / n} \quad (11)$$

$$Bias = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} O_i} - 1 \quad (12)$$

where *n* represents the number of samples from a station; *Y* stands for the value of IMERG precipitation data; *O* is the value of observed precipitation data; $Bias$ is the average bias level between two groups of data.

4. RESULTS AND ANALYSIS

4.1 The correlation analysis between precipitation and auxiliary variables

Variables mostly strongly correlated were chosen to explore the main influencing factors of precipitation and establish a multi-source precipitation data fusion model. The daily ground precipitation data and radar-derived precipitation data during one-time precipitation from July 6 to July 11, 2020, were collected, and their correlation coefficients with auxiliary ground parameters were obtained, followed by the correlation data analysis (Figure 3). It could be discovered from Figure 3 that when the correlation coefficient between ground precipitation and radar-derived data was the maximum, the correlation coefficient was always greater than 0.8 except on the first precipitation day, namely, July 6. The constructed concrete multi-source precipitation data fusion model is listed in Table 1.

Table 1: Multi-Source Precipitation Data Fusion Model

Date	Model
2020.07.06	$\widehat{Prec_{XGBoost,06}} = f_{XBGoost}(Rader, Lon, Lat, DEM)$
2020.07.07	$\widehat{Prec_{XGBoost,07}} = f_{XBGoost}(Rader, Lon)$
2020.07.08	$\widehat{Prec_{XGBoost,08}} = f_{XBGoost}(Rader, Lon, Lat)$
2020.07.09	$\widehat{Prec_{XGBoost,09}} = f_{XBGoost}(Rader, Lon, Lat)$
2020.07.10	$\widehat{Prec_{XGBoost,10}} = f_{XBGoost}(Rader, Lon, Lat, DEM)$

4.2 Daily precipitation data fusion results

In this research, the satellite-derived precipitation data on July 6, 2020, were chosen for tests. To keep the temporal consistency between high-resolution (1 km × 1 km) DEM data and satellite-derived precipitation data, the satellite-derived precipitation data and ground observation data were fused using the point-surface fusion method (Scheme 1) and station bias correction method (Scheme 2). Meanwhile, the fusion results of daily precipitation data were calculated through MARS, RF, and GPR. The fusion results obtained by Schemes 1 and 2 are displayed in Figures 2 and 3.

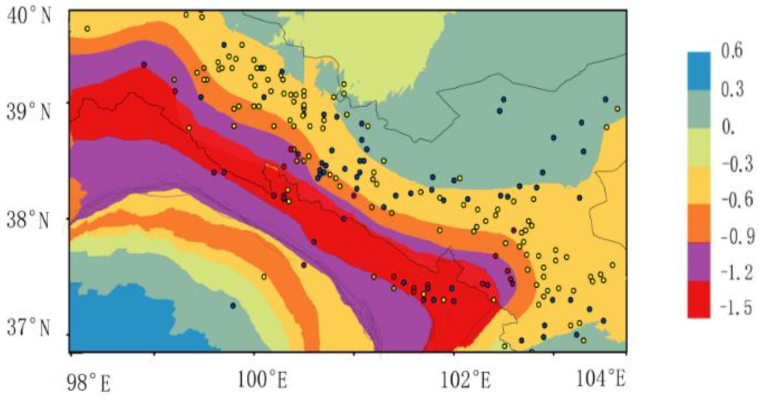

Figure 2: Fusion Data Distribution Obtained through Point-Surface Fusion Method Based on GPR

It could be observed from Figure 2 that the bias between MERRA2-derived precipitation data and observation data in Qilian Mountain presented a spatial distribution feature of "small on south and north slopes and large on ridges" along the mountain. Qilian Mountain is close to the Hexi corridor and adjoins the Qaidam basin and the Yellow River basin, where the altitude of south and north slopes is lower than that of ridges, accompanied by relatively simple landforms. The precipitation data bias in Qilian Mountain is associated with the altitude, climate, and underlying surfaces in this area.

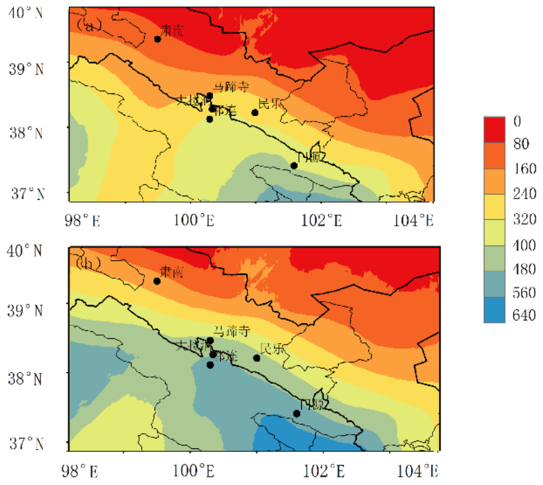

Figure 3: Fusion Data Distribution Obtained through Station Bias Calibration Method Based on GPR

It could be seen from Figure 3 that the annual precipitation fields of precipitation data and observation data on Qilian Mountain showed identical variation trends. The precipitation was distributed along the northwest-southeast direction, namely, the trend of the mountain, which corresponded to the altitude very well. From the north to the south, the precipitation showed a rising trend in a steplike fashion. The changes in the satellite-derived precipitation data reflected the change features of precipitation in this area.

4.3 Accuracy evaluation of fusion results

With the precipitation data from ground observation stations as true values, the two fusion schemes were quantitatively evaluated through the leave one-out cross validation method. The station verification results under Schemes 1 and 2 are listed in Tables 2 and 3, respectively.

Table 2: Station Verification Results of Scheme 1

Algorithm	Model accuracy			Verification accuracy		
	R	*RMSE*	*Bias* (%)	*R*	*RMSE*	*Bias* (%)
MARS	0.95	34.81	0.25	0.73	34.83	-0.08
RF	0.96	13.05	0.05	0.77	31.34	0.02
GPR	0.98	15.21	0.08	0.79	30.21	-0.02

Table 3: Station Verification Results of Scheme 2

Algorithm	Model accuracy			Verification accuracy		
	R	*RMSE*	*Bias* (%)	*R*	*RMSE*	*Bias* (%)
MARS	0.97	18.77	0.05	0.66	40.12	-0.02
RF	0.87	25.49	0.01	0.68	36.41	-0.01
GPR	0.08	17.43	0.16	0.77	32.24	-0.01

5. CONCLUSIONS

With Qilian Mountain as the main research area and the rainfall process on July 6, 2020, as the main research object, auxiliary variables like radar-derived precipitation data, satellite-derived precipitation data, soil longitude and latitude, and DEM were combined to comprehensively figure out the correlations of ground observation data, radar-derived precipitation data and satellite-derived precipitation data with auxiliary variables of the ground surface. On this theoretical basis, a machine learning algorithm was used to construct a multiple nonlinear regression model and a precipitation data fusion model applicable to the climatic conditions in Qilian Mountain. Next, the spatial distribution of residuals obtained by the Gaussian process regression model was estimated through the adaptive multi-spline regression method and random forest algorithm. Finally, the daily fusion precipitation data with a spatial resolution of 1 km were obtained, followed by the accuracy test using ground observation data. The research conclusions were drawn as follows:

(1) Ground observation precipitation data presented an evident positive correlation with radar-derived precipitation data, and its correlation with auxiliary ground surface parameters was changed with the ground precipitation process. For instance, the correlation coefficient between ground observation data and latitude change reached 0.49, but it turned into a negative value (-0.43) on July 9.

(2) The fusion results obtained by the RF method would generate massive traces, while those obtained by the MARS method were relatively ambiguous. However, the fusion results obtained by the newly proposed GPR method displayed reasonable changes, and its detailed data information presented a higher quality than the results obtained through the direct interpolation of station data information. The accuracy of point-surface fusion results acquired through the three algorithms exceeded the accuracy of the fusion result obtained through error correction, and some interpolation traces in error correction conclusions were reduced.

(3) When the GPR method was used to realize the point-surface fusion of satellite-derived precipitation data and ground observation data, the accuracy of fusion results ($R = 0.69$ and $Bias = -0.03\%$) was considerably improved in comparison with the original satellite-derived precipitation data.

REFERENCES

[1] Jarvis, H. I. Reuter, A. Nelson, et al. Hole-filled SRTM for the globe Version4[DB/OL]. https://srtm.csi.cgiar.org, CGIAR-CSI SRTM 90m Database, 2008.

[2] CHAO L, ZHANG K, LI Z, et al. Geographically weighted regression-based methods for merging satellite and gauge precipitation[J]. Journal of Hydrology, 2018, 558: 275-289.

[3] GOOVAERTS P. Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall [J]. Journal of Hydrology, 2000, 228(1-2): 113-129.

[4] H Chen, Chandrasekar V, Cifelli R, et al. A Machine Learning System for Precipitation Estimation Using Satellite and Ground Radar Network Observations [J]. IEEE Transactions on Geoscience and Remote Sensing, 2019, PP (99): 1-13.

[5] HILL D J. Assimilation of weather radar and binary ubiquitous sensor measurements for quantitative precipitation estimation[J]. Journal of Hydroinformatics, 2015, 17(4): 598-613.

[6] LI M, SHAO Q. An improved statistical approach to merge satellite rainfall estimates and raingauge data [J]. Journal of Hydrology, 2010, 385(1-4): 51-64.

[7] Li Y L, Xiong L H, Yan L. A geographically weighted regression Kriging approach for TRMM-rain gauge data merging and its application in hydrological forecasting [J]. Resources and Environment in the Yangtze Basin, 2017, 26 (09): 1359-1368.

[8] Liu Y B, Fu Q N, Song P, Zhao X S, Dou C C. Satellite retrieval of precipitation: an overview [J]. Advanced in Earth Science, 2011, 26 (11): 1162-1172.

[9] Pan Y, She 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(01): 177-186.

[10] Rasmussen, Carl, Edward, et al. Gaussian Processes for Machine Learning (GPML) Toolbox [J]. Journal of Machine Learning Research, 2010.

[11] Shen Y, Pan Y, Xu B, Yu J J. Parameter improvements of hourly automatic weather stations precipitation analysis by optimal interpolation over China [J]. Journal of Chengdu University of Information Technology, 2012, 27 (02): 219-224.

[12] 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.

[13] Tomoo U, Kazushi S, Takuji K, et al. A Kalman Filter Approach to the Global Satellite Mapping of Precipitation (GSMaP) from Combined Passive Microwave and Infrared Radiometric Data (2. Global Satellite Mapping of Precipitation (GSMaP) Project, Precipitation Measurements from Space) [J]. Journal of the Meteorological Society of Japan. Ser. II, 2009.

[14] Wu Z, Zhang Y, Sun Z, et al. Improvement of A Combination of TMPA (or IMERG) and Ground-based Precipitation and Application to a Typical Region of the East China Plain. Science of The Total Environment, 2018, 640/641: 1165-1175.

[15] Xie P, Xiong A Y. A conceptual model for constructing high-resolution gauge-satellite merged precipitation analyses[J]. Journal of Geophysical Research Atmospheres, 2011, 116(D21).

[16] Yang Q Q, Gao C, Zha X Y, Zhang P J. Changes in upstream climate and runoffs of Huaihe River under RCP scenario [J]. Journal of Anhui Agricultural Sciences, 2020, 48(03): 209-214.

[17] Yilmaz K K, Adler R F, Tian Y, et al. Evaluation of a Satellite-Based Global Flood Monitoring System. International Journal of Remote Sensing, 2010, 31(14): 3763-3782.

[18] Zhu G F, Pu T, Zhang T, Liu H L, Zhang X B, Liang F. The accuracy of TRMM precipitation data in Hengduan mountainous region, China [J]. Scientia Geographica Sinica, 2013, 33(09): 1125-1131.