Spatial Prediction Using Random Forest Spatial Interpalation with Sample Augmentation: A Case Study for Precipitation Mapping

JIAO Sijia
Chang'an University
WU Tianjun (tjwu@chd.edu.cn)
Chang'an University
LUO Jiancheng
Chinese Academy of Sciences
ZHOU Ya'nan
Hohai University
DONG Wen
Chinese Academy of Sciences
WANG Changpeng
Chang'an University
Dong Shiying
Chang'an University

Research Article

Keywords: Data Augmentation, Random Forest, Spatial Prediction, Precipitation, Mixup, Unsampling, Small Sample

Posted Date: November 8th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2226248/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Earth Science Informatics on January 26th, 2023. See the published version at https://doi.org/10.1007/s12145-023-00936-6.

Abstract

Spatial prediction (SP) based on machine learning (ML) has been applied to soil water quality, air quality, marine environment, etc. However, there are still deficiencies in dealing with the problem of small samples. Normally, ML require large amounts of training samples in order to prevent overfitting. The data augmentation method of mixup and synthetic minority over-sampling technique (SMOTE) ignores the similarity of geographic information. Therefore, this paper proposes a modified upsampling method and combines it with the random forest spatial interpolation (RFSI) to deal with the small sample problem in geographical space. The modified unsampling mainly reflected in the following two aspects. Firstly, in the process of selecting nearest points, it is to select points with similar geographic information in some aspects of the category after classification. Secondly, the selected difference is the difference of each category. In order to verify the effectiveness of the proposed method, we select precipitation as the target factor and conduct a comparative experiment. The experimental results show that the combination of the modified upsampling method and RFSI effectively improves the accuracy of spatial prediction.

1. Introduction

Since the 1960s, spatial information technology supported by satellite positioning system, geographic information system and remote sensing has gradually develop. And a large number of data with spatial location have been collected, processed and applied(Li and Shao, 2019). Compared with other data, spatial data is difficult to use the classical statistical method of variable independence assumption because of its spatio-temporal correlation. And Newton's prediction and other methods in geometric space are not applicable. In 1970, Professor Toblert (1970) proposes the "First Law of Geography", which provide a theoretical basis for the analysis and application of spatial data. Spatial prediction(SP) has also been developed and improved.

At present, SP methods can be roughly divided into four categories: (1) deterministic prediction: inverse distance weighted(IDW) (Willmott et al., 1985), (2) geostatistics method: kriging (Matheron, 1963), (3) combination method: regression kriging(RK) (Mohanasundaram et al., 2020), (4) machine learning(ML). With the complexity of practical problems, these basic methods cannot meet the requirements. So, on the basis of them, they gradually improved and put forward many new methods. For instance, Yan et al. (2021) apply a novel multiple parameters synchronization optimization IDW algorithm which involves anisotropy(PIDW) to two spatial data of different scales. It is proved that the method can effectively improve the accuracy. Kriging also has certain expansion, such as Universal Kriging(UK) (Xuan Thanh et al., 2015), Kriging with External Drift(KED) (Berndt et al., 2014). Wu et al. (2021) uses geographical map spot as basic mapping units. In comparison with traditional regular grid-based methods, it achieves higher accuracy. On the basis of considering spatial information, RF develops into Random Forest for spatial data (RFsp) (Hengl et al., 2018) and Random Forest Spatial Interpolation (RFSI) (Sekulic et al., 2020). These method have been applied to many fields such as soil water quality, marine environment, geological exploration, air quality, etc. But they are not discussed for the small sample problem. And the number of meteorological stations set up in a region is often insufficient to study the situation of the

entire region because of the limitations of terrain, financial resources and other conditions. So, ML may have insufficient fitting in case of insufficient samples. And with the progress of productivity, the demand of social and economic life for the delicacy and timeliness of geospatial information is further highlighted. It is of great practical significance to develop spatial prediction models and improve the mapping level in the case of small samples.

The methods for small sample learning are roughly divided into three categories: small sample learning methods based on data augmentation, small sample learning methods based on metric learning, and small sample learning methods based on meta-learning (Wang, 2022). Data augmentation is to add new data to the original dataset, which can be unlabeled data or composite labeled data. In supervised data augmentation, it is divided into single sample data augmentation and multi sample data augmentation. Single sample data augmentaion is performed around the a sample itself. Multi data augmentation uses multiple samples to generate new samples, such as synthetic minority over-sampling technique(SMOTE) (Chawla et al., 2002), mixup (Zhang et al., 2017). These data augmentation methods are combined with ML to form many methods to deal with small sample learning.

In the field of geoscience, neural network(NN) (Lawrence et al., 1997) has been integrated with data augmentation technology and widely applied to the classification of hyperspectral images(HSIs). For example, Li et al. (2019) use deep convolutional neural network(CNN) to extract pixel-block pair (PBP) features, and decision fusion is utilized for final label assignment. Results demonstrate that this method can outperform support vector machine with the composite kernel (SVM-CK) (Li et al., 2019) and multiple classifier systems-based SVM with random feature selection (SVM-RFS) (Waske et al., 2010). Generative adversarial network(GAN) has been practical and effective in HSIs classification(Zhu et al., 2018). And improved Wasserstein GAN is morecapable of generating similar radar images while achieving higher structural similarity results(Lee et al., 2020). Accion et al. (2020) introduce Dual-Window Superpixel(DWS) data augmentation on the basis of CNN. Experimental results show that the method is effective in HSIs in classification. In prediction aspects, Li et al. (2022) use window offset, scaling and rotation data augmentation and deep CNN to predict subsurface mineral deposits. And this method can efficiently predict mineral prospective areas where there are few ore deposits. But its data enhancement method is to enhance samples by observing from different angles and distances. It is not applicable to station data. Yang el al. (2022) adopt cropping operations to generate sufficient training samples and utilize LeNet, AlexNet and VggNet to predict mineral deposits. LetNet can outperform other method. But its cropping data augmentation is to operate on the image. Huang el al. (2020) propose spatial autocorrelation-based mixture interpolation(SABAMIN). Compared with traditional ML, it's accuracy is improved. However, it use kriging prediction to create reliable pseudo data. In general, kriging has high theoretical requirements. And it is relatively difficult to fit the variogram.

To sum up, the combination of data augmentation and ML has been applied to the classification and prediction aspects of geosciences, especially in the classification aspcets. However, it is relatively less used in the prediction aspects. In addition, the process of generating pseudo data by kriging interpolation is relatively complex in the application process. Data augmentation methods such as clipping are not

applicable to station data. Therefore, this paper proposes the Random Forest Spatial Interpalation-Modified Unsampling(RFSI-MUS) based on the above problems. It is mainly used in the RFSI model to enhance the data of observation sample points through the modified unsampling method, as to solve the underfitting phenomenon in the RFSI model. The modified unsampling mainly reflected in the following two aspects. Firstly, in the process of selecting nearest points, it is to select points with similar geographic information in some aspects of the category after classification. Secondly, the selected difference is for each category. In order to verify the effectiveness of the method proposed in this paper, the precipitation data set of Chongqing is used to compare RFSI-MUS with Random Forest(RF), RFSI and RFSI-Mixup.

2. Study Area And Data Set

2.1 Study area

Chongqing is selected as the study area to validate the performance of the proposed method. Chongqing is located in Southwest China and the upper reaches of the Yangtze River. It is surrounded by Daba Mountain, Wuhui mountain, Wuling Mountain and Daluta mountain in the north, East and south. The landform is dominated by hills and mountains, with a large slope area. It is known as "mountain city". The following map (see Fig. 1) shows distribution of digital elevation model (DEM).

2.2 Meteorological data

The raw station data of precipitation were collected from China Meteorological Information Center (http://data.cma.cn/). This paper downloads the daily data of 11 stations of Chongqing in January 2018. The daily rainfall measurements are in units of 1/10 of a mm. To obtain the precipitation in January, the daily precipitation is averaged. The map(Fig. 1) shows the locations of 11 stations.

2.3 Auxiliary data for multiple covariates

Multiple covariates is important to improve model performance. For example, latitude, longitude and altitude as covariates to predict temperature (Mohsenzadeh Karimi et al., 2020). Beheren et al. (2018) use educlidean distance as covariates. Land surface temperature (LST) is also widely used for the prediction of precipitation (Alvarez et al., 2014). Therefore, this paper selects elevation, humidity, LST, NDVI and GPM precipitation as covariates to predict precipitation. The following Table 1 is specific information of multiple-covariates and their data source.

Table 1
Multiple-covariates for the prediction

Type	Spatial resolution	Temporal resolution	Data source
Elevation	30m	\	GDEMV2
NDVI	1km	Monthly	MOD13A3
LST	1km	8-days	MOD11A2
GPM	0.1°	Monthly	V06(GPM_3IMERGM)
Humidity	1km	Daily	\

Elevation were collected from Geospatial Data Cloud (http://www.gscloud.cn/). Spatial resolution is 30m. GPM precipitation, NDVI, LST were collected from National Aeronautics and Space Administration(NASA) (https://pmm.nasa.gov/). Spatial resolution of GPM precipiatation is 0.1°. In order to unify the spatial resolution, this paper adopts the resampling technology to unify the covariates spatial resolution to 1km. In addition, the time resolution of HUM is every day. LST time resolution is 8 days. LST data includes daytime surface temperature (LSTd) and nighttime surface temperature (LSTn). To obtain the monthly average humidity and LST, average the Humidity of every day and the LST of every 8 days.

3. Methodology

This paper adopts the four SP methods of RF, RFSI, RFSI-Mixup and RFSI-MUS to predict the monthly average precipitation of Chongqing. In order to verify them, leave-one-out cross validation (LOOCV) is used to evaluate acuracy criteria. Figure 2 show the flow chart of SP using these methods.

3.1 Random Forest and Random Forest Spatial Interpolation

RF(Breiman, 2001) algorithm is an integrated learning method based on bagging proposed by Breiman in 2001. It can be used for data classification and regression prediction by constructing multiple decision trees to deal with the relationship between independent variables and dependent variables. RF builds a large number of tree models. The importance of various eigenvalues is integrated and screened, and the importance of different eigenvalues is fully considered to select the optimal sample eigenvalue to find the optimal solution. In order to obtain the final predicted value, average the values of all predicted values. RF model predictions can be written as

$$\hat{z}(s_0) = f\left(x_1\left(s_0\right), x_2\left(s_0\right), \cdots, x_m\left(s_0\right)\right) \quad (1)$$

Considering that nearby observations carry information about the value at a prediction location, RFSI model is proposed. It is as follows:

$$\hat{z}(s_0) = f\left(x_1\left(s_0\right), x_2\left(s_0\right), \cdots x_m\left(s_0\right), d_1, z\left(s_1\right), \cdots, d_n, z\left(s_n\right)\right)$$

2

where the $x_i(s_0)$(i = 1,2,...,m) are covariates at location s_0, the$\hat{z}\left(s_0\right)$is prediction at location s_0, the $z(s_j)$(j = 1,...,n) and d_j are the j-th nearest observation and euclidean distance from s_0.

3.2 Random Forest Spatial Interpolation with Mixup (RFSI-Mixup)

RFSI-Mixup is mainly used to deal with the problem that the number of observation points is too small in the process of training the RFSI model, which leads to the lack of fit to the prediction. It mainly uses the mixup method on the basis of RFSI to expand the sample size of training points in pairs, so as to solve the problem of small observation points in the RFSI training process. The mixup mainly uses E.Q. (3)-(4) for data enhancement.

$$x_i^{Mixup} = \varphi x_i\left(s_k\right) + \left(1-\varphi\right) x_i\left(s_l\right)$$

3

$$z^{Mixup} = \varphi z\left(s_k\right) + \left(1-\varphi\right) z\left(s_l\right)$$

4

where $\mathrm{x}_{\mathrm{i}}^{\mathrm{Mixup}}$, z^{Mixup} is new covariate and observation used mixup. And $\varphi \in (0,1), \mathrm{k},\mathrm{l} \in \mathrm{i}$. As the pseudo data generated in mixup combination is kept between two points, it may not be suitable for too large labels. Therefore, when the data of one station is twice that of other stations, this paper adopts the following label combination method:

$$z^{Mixup} = z\left(s_k\right) + z\left(s_l\right)$$

5

Pseudocode of algorithm for SP based on RFSI-Mixup model is shown in Table 2.

Table 2
Pseudo code of SP algorithm based on RFSI-Mixup

Input: Mtry, min.node.size, sample fraction, num.trees of RF parameters of 500. Covariates and observations of chongqing, L = 0.05,0.1,…,0.95.
Output: predictions, uncertainty, MAE, RMSE, R^2, CCC
1. Observations and its covariates devided 11 groups,$obs_k(k = 1,2,\ldots,11)$.
2. for l in 1 to 19
for j in 1 to 500
data_test = obs_k
data_train = obs_{-k}
Use Eq. (3)-(5) for data_train to augmentation.
Train RFSI.
Calculate MAE of data_test.
end
end
3. Select the RFSI parameter corresponding to the minimum MAE.
4. Train optimal RFSI model.
5. for i in 1 to 11
Implement 2–4.
end
6. LOOCV for $obs_k(k = 1,2,\ldots,11)$.
7. Calculate MAE, RMSE, R^2, CCC.
8. Predict all points precipitation and uncertainty of Chongqing.
9. Mapping of predictions and uncertainty.

3.3 Random Forest Spatial Interpalation with Modified Unsampling (RFSI-MUS)

As mixup only combines data in pairs, it does not consider the spatial location relationship between points. In other words, the greater the similarity of the points whose spatial positions are close to each other. In the combination process, the combination value fluctuates between two points resulting in the

inability to meet the diversity of the expanded sample. That is, the predicted value will be too large or too small in the process of boundary point prediction, so the error will be too large. As for the shortcomings of mixup, this paper proposes the RFSI-MUS model. Firstly, the observation points are classified according to the similarity of precipitation, spatial distance and covariates. To ensure that there is not much information about observation points, each category should contain at least three observation points. Secondly, the similarity points of verification points are selected in the calss according to precipitation, spatial distance, etc. Finally, according to E.Q.(6)-(7) enhance data. Compared with the unsampling, this paper uses the difference of each class, and ensures that its covariates and precipitation are positive in the process of selecting random numbers.

$$x_i^{MUS} = x_i\left(s_k\right) \pm rand\left(0, a\right) \Delta x_i$$

6

$$z^{MUS} = z\left(s_k\right) \pm rand\left(0, a\right) \Delta z$$

7

where $\Delta \mathrm{x_i}$, $\Delta \mathrm{z}$ is differences of class, $\mathrm{x_i^{MUS}}$, z^{MUS} is new covariate and observation used modified unsampling. *a* is maxium number that satisfies the condition of $\mathrm{x_i^{MUS}}$, $\mathrm{z^{MUS}} > 0$. The classification and of observation points pseudo code based on Chongqing are shown in the Fig. 4 and Table 3. The selection of covariates depends on their importance.

Table 3
Pseudo code of spatial prediction algorithm based on RFSI-MUS

Input: Mtry, min.node.size, sample fraction, num.trees of RF parameters of 500. Covariates, observations of Chongqing.
Output: predictions, uncertainty, MAE, RMSE, R^2, CCC
1. Classification based on similarity of covariates, distance and precipitation.
2. Calculate differences, maximum(*M*), minimum(*m*) for each class.
3. Observations and its covariates devided 11 groups, obs_k(k = 1,2,…,11)
4. Select nearest points $nobs_k$ in the class for obs_k. 5. for i in 1:m Extract random number. if $\mathrm{nobs}_k^i = M_i$ use E.Q. (6) or (7) '+' to data augmentation. else if $\mathrm{nobs}_k^i = m_i$ use E.Q. (6) or (7) '-' to data augmentation else use (6)-(7) to data augmentation. end 6. The data after data augmentation and obs_k combination into obs_k^{MUS}.
7. for j in 1:500
data_test = obs_k
Use obs_k^{MUS} to trian RFSI.
Calculate MAE of data_test.
end
8. Select the RFSI parameter corresponding to the minimum MAE.

Input: Mtry, min.node.size, sample fraction, num.trees of RF parameters of 500. Covariates, observations of Chongqing.
9. Train optimal RFSI model.
10. for l in 1 to 11
Implement 4–9.
end
11. LOOCV for obs_k(k = 1,2,…,11)
12. Calculate MAE, RMSE, R^2, CCC,
13. Predict all points precipitation and uncertainty of Chongqing.
14. Mapping of predictions values and uncertainty.

3.4 Accuracy Assessment and Uncertainty Analysis

To verify the validity and accuracy of prediction of the RFSI-MUS model, This paper uses LOOCV method to evaluate the model performance. And it use the following performance criteria. It is Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Coefficient of Determination (R^2), Concordance Correlation Coefficient (CCC). These performance criteria can be written as

$$MAE = \frac{1}{N}\sum_{r=1}^{N}\left|\hat{z}\left(s_r\right) - z\left(s_r\right)\right|$$

8

$$RMSE = \sqrt{\frac{1}{N}\sum_{r=1}^{N}\left(\hat{z}\left(s_r\right) - z\left(s_r\right)\right)^2}$$

9

$$R^2 = \left[1 - \frac{SSE}{SST}\right]\%$$

10

$$CCC = \frac{2\rho\sigma_{\hat{z}}\sigma_z}{\sigma_{\hat{z}}^2 + \sigma_z^2 + \left(\mu_{\hat{z}} - \mu_z\right)^2}$$

11

Furthermore, in order to verify the validity of RFSI-MUS in selecting nearest points and classification, we discuss the differences of the whole study area and the nearest points which are only the similarity of spatial distance. The models are RFSI-MUS-D and RFSI-MUS-N respectively. In addition, in order to measure the uncertainty of the whole study area, the following performance criteria is adopted as the error measurement in this paper.

$$\sigma\left(s_0\right) = \frac{\hat{z}_{0.841}\left(s_0\right) - \hat{z}_{0.159}\left(s_0\right)}{2}$$

12

where the$\hat{\mathrm{z}}\left(\mathrm{s_r}\right)$ and $z\left(\mathrm{s_r}\right)$ are the prediction and observation at cross-validation location s_r, N is total number of cross-validation location. *SSE* is the sum of squared errors at cross-validation locations and *SST* is the total sum of squares $\sigma_{\hat{\mathrm{z}}}^2$, $\sigma_{\mathrm{z}}^2$are the predicted and observed variance.$\mu_{\hat{\mathrm{z}}}$, μ_{z} are the predicted and observed mean. ρ is s the correlation coefficient between prediction and observation.

4. Results And Analysis

4.1 The prediction result of each method

The SP based on RF, RFSI, RFSI-Mixup, RFSI-MUS, RFSI-MUS-D,RFSI-MUS-N models showed in Fig. 5. All results showed highest predicted precipitation in the east of Chongqing, and low predicted precipitation in the west of Chongqing. According to DEM of Fig. 1, the elevation in the west of Chongqing is low, while that in the east is high. This trend is consistent with the characteristics that the higher the elevation is, the higher the precipitation is.

Compared six precipitation prediction, the prediction is more high in small area south-east of Chongqing based RF, RFSI and RFSI-mixup. Especially in the area where the precipitation of the station data is not high, the predicted precipitation is high. But in some corresponding regions it does not have high precipitation based on RFSI-MUS. And combine two models RFSI-MUS-D and RFSI-MUS-N, They have effectively improved the situation. So, RFSI-MUS is a good choice for SP.

4.2 Evaluation and Analysis of Results

According to E.Q. (12), Fig. 6 provides the spatial distribution of prediction uncertainty by different methods. From this figure, the uncertainty of RF is relatively high. The difference in uncertainty is mainly concentrated in the southeast and northeast of Chongqing. It can be seen from the station map in Fig. 1 that there are few station points in this area, which makes the uncertainty in this area relatively high. By contrast, the uncertainty of RFSI-MUS is relatively low. Compared RFSI-MUS-D and RFSI-MUS-N, the uncertainty of RFSI-MUS is also relatively low. Therefore, the nearest points of the similarity of spatial distance, precipitation, covariates and differences of each class selected by RFSI are valid.

The cross validation results of the proposed RFSI-MUS model with RF, RFSI, RFSI-Mixup, RFSI-MUS-D and RFSI-MUS-N are shown in Table 4. Since the CCC and R^2 of RF and RFSI are near 0, they are not shown in the table. It is found by comparison that RFSI-MUS is the largest and RF is the smallest in terms of predicted accuracy. In terms of MAE, RFSI-MUS is the smallest and RF is highest. In addition, it can be seen from the correlation plot between the observed values and the predicted values in Fig. 7 that RF, RFSI, RFSI-Mixup, RFSI-MUS-D and RFSI-MUS-N are relatively dispersed compared with RFSI-MUS. It is also confirmed that the RF, RFSI, and RFSI-Mixup, RFSI-MUS-D and RFSI-MUS-N in Table 3 have higher RMSE and lower R^2 and CCC.

Table 4
Accuracy metrics of four prediction methods based on LOOCV

Method	MAE	RMSE	R^2	CCC
RF	3.02697	4.16993	\	\
RFSI	2.84463	3.8273	\	\
RFSI-Mixup	1.4193	1.8027	0.7818	0.8465
RFSI-MUS	0.9848	1.6225	0.8233	0.8752
RFSI-MUS-D	1.2054	2.054	0.7167	0.7894
RFSI-MUS-N	2.4151	2.9138	0.43	0.5944

Further analysis shows that RF only uses the intrinsic covariates of each point, and no other prediction factors are introduced. In addition, when the number of observation points is small, insufficient fitting may occur. RFSI introduces the observation value of the nearest point and the distance to the prediction point on the basis of RF, but the accuracy is still not greatly improved. RFSI-Mixup adopts data augmentation for sample points on the basis of RFSI. Compared with RF and RFSI, the precision has been further improved. But mixup is only a simple combination of sample points, so it does not make full use of geographic information. Based on the shortcomings of mixup method, the proposed RFSI-MUS effectively uses spatial information. So, The spatial prediction error of RFSI-MUS is relatively low. And compared with RFSI-MUS-D and RFSI-MUS-N, and also effectively verify the effectiveness of the two points mentioned in RFSI-MUS. To sum up, RFSI-MUS is a good choice in the combination of SP and data augmentation.

5. Conclusion

In order to obtain accurate SP results, this paper proposes RFSI-MUS that data augmentation SP model based on modified upsampling. It not only considers the selection of nearest points of spatial location, precipitation and covariate similarity in the process of data augmentation, but also considers the

difference by using the classification method to increase the diversity of samples. In order to verify the accuracy of the RFSI-MUS model, this paper conducts a comparative experiment based on the precipitation in Chongqing. And compares RFSI-MUS with RF, RFSI, RFSI-Mixup, RFSI-MUS-D and RFSI-MUS-N models. From the perspective of prediction mapping effect, uncertainty, and cross validation accuracy analysis, the RFSI-MUS method is effective in SP.

In summary, RFSI-MUS has a number of important advantages over RFSI and RFSI-Mixup. Firstly, compared with RFSI method, RFSI-MUS handles the case of few sample points. Because various reasons such as geography and financial resources in life, sample points are difficult to meet the requirements of ML. Therefore, the proposal of RFSI-MUS potentially solves this problem. Secondly, compared with mixup data augmentation methods, RFSI-MUS data augmentation takes into account the first law of geography, and is no longer a simple mechanical data augmentation. Thirdly, RFSI-MUS makes use of the characteristics that covariates such as altitude and LST are continuous variables. During data augmentation, they will not exceed the range of the region, so as to avoid that the covariates are too large or too small to meet the terrain conditions.

In spatial prediction, the RF method is widely used. And it has expanded with the complexity of the data set, such as RFsp and RFSI. However, these models do not take into account the small number of sample points. Therefore, the RFSI-MUS proposed in this paper mainly aims at the small number of sample points. And the experiment proves its effectiveness. However, RFSI-MUS is also expanded in other aspects. For example, in the process of selecting nearest points, only a simple natural segment method is used to divide the similarity of covariates. Therefore, we can discuss more effective methods to depict the similarity of covariates later. In the process of data augmentation, although the influence of nearby points is considered, it is still a relatively simple combination method. The combination method under various factors can be considered in future research.

Declarations

Author's Contribution

Jiao Sijia carried out the data preparation, performed the experiments, experimental analysis, and wrote the manuscript. Wu Tianjun outlined there search topic, proposed there search methodology, and designed the experiments. All authors have read and agreed to the published version of the manuscript.

Conflict of Interest

There is no conflict of interest.

Availability of Data and Materials

For data and materials in this paper, please contact 2020112038@chd.edu.cn

Funding

This work was supported in part by the Science and Technology Project of Inner Mongolia Autonomous Region under Grant 2021ZD0045, the Project of Chongqing Agricultural Industry Digital Map under Grant 21C00346, National Natural Science Foundation of China under Grant 42071316 and 12001057, Key Research and Development Program of Shaanxi under Grant 2021NY-170, Fundamental Research Funds for the Central Universities, CHD under Grant 300102120201, 300102122101, and 300102269103, National Key Research and Development Program, under Grant 2021YFB3900905 and 2021YFB3901300.

References

1. Accion A, Arguello F, Heras DB (2020) Dual-Window Superpixel Data Augmentation for Hyperspectral Image Classification. Applied Sciences-Basel 10(24): 8833.
2. Alvarez O, Guo Q, Klinger RC, Li W, Doherty P (2014) Comparison of elevation and remote sensing derived products as auxiliary data for climate surface interpolation. International Journal of Climatology 34(7): 2258-2268.
3. Behrens T, Schmidt K, Rossel RAV, Gries P, Scholten T, MacMillan R A (2018) Spatial modelling with Euclidean distance fields and machine learning. European Journal of Soil Science 69(5): 757-770.
4. Berndt C, Rabiei E, Haberlandt U (2014) Geostatistical merging of rain gauge and radar data for high temporal resolutions and various station density scenarios. Journal of Hydrology 508: 88-101.
5. Breiman L (2001) Random Forests. Machine Learning 45(1): 5-32.
6. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research 16: 321-357.
7. Hengl T, Nussbaum M, Wright MN, Heuvelink GBM, Graeler B (2018) Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. Peerj 6: e5518.
8. Huang C, Shibuya A (2020) High Accuracy Geochemical Map Generation Method by a Spatial Autocorrelation-Based Mixture Interpolation Using Remote Sensing Data. Remote Sensing 12(12): 1991.
9. Lawrence S, Giles CL, Tsoi AC, Back AD (1997) Face recognition: A convolutional neural-network approach. IEEE transactions on neural networks 8(1): 98-113.
10. Lee H, Kim J, Kim EK, Kim S (2020) Wasserstein Generative Adversarial Networks Based Data Augmentation for Radar Data Analysis. Applied Sciences-Basel 10(4): 1449.
11. Li HT, Shao ZD (2019) Review of spatial interpolation analysis algorithm. Computer Systems Applications 28(07): 1-8.
12. Li W, Chen C, Zhang MM, Li HC, Du Q (2019) Data Augmentation for Hyperspectral Image Classification With Deep CNN. Ieee Geoscience and Remote Sensing Letters 16(4): 593-597.
13. Li YS, Peng C, Ran XJ, Xue LF, Chai SL (2022) Soil geochemical prospecting prediction method based on deep convolutional neural networks-Taking Daqiao Gold Deposit in Gansu Province, China as an example. China Geology 5(1): 71-83.
14. Matheron G (1963) Principles of geostatistics. Economic geology 58(8): 1246-1266.

15. Mohanasundaram S, Udmale P, Shrestha S, Baghel T, Doshi SC, Narasimhan B, Kumar GS (2020) A new trend function-based regression kriging for spatial modeling of groundwater hydraulic heads under the sparse distribution of measurement sites. Acta Geophysica 68(3): 751-772.
16. Mohsenzadeh Karimi S, Kisi O, Porrajabali M, Rouhani-Nia F, Shiri J (2020) Evaluation of the support vector machine, random forest and geo-statistical methodologies for predicting long-term air temperature. ISH Journal of Hydraulic Engineering 26(4): 376-386.
17. Sekulic A, Kilibarda M, Heuvelink GBM, Nikolic M, Bajat B (2020) Random Forest Spatial Interpolation. Remote Sensing 12(10): 1687.
18. Tobler WR (1970) A computer movie simulating urban growth in the Detroit region. Economic geography 46(sup1): 234-240.
19. Wang H (2022) Research on Few-Shot Image Recognition Technology Based on Data Augmentation and Metric Learning. master thesis University of Electronic Science and Technology.
20. Waske B, van der Linden S, Benediktsson JA, Rabe A, Hostert P (2010) Sensitivity of support vector machines to random feature selection in classification of hyperspectral data. IEEE Transactions on Geoscience and Remote Sensing 48(7): 2880-2889.
21. Willmott CJ, Rowe CM, Philpot WD (1985) Small-scale climate maps: A sensitivity analysis of some common assumptions associated with grid-point interpolation and contouring. The American Cartographer 12(1): 5-16.
22. Wu TJ, Luo JC, Gao LJ, Sun YW, Yang YP, Zhou YN, Dong W, Zhang X (2021) Geoparcel-Based Spatial Prediction Method for Grassland Fractional Vegetation Cover Mapping. Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing 14: 9241-9253.
23. Xuan Thanh N, Ba Tung N, Khac Phong D, Quang Hung B, Thi Nhat Thanh N, Van Quynh V, Thanh Ha L (2015) Spatial Interpolation of Meteorologic Variables in Vietnam using the Kriging Method. Journal of Information Processing Systems 11(1): 134-147.
24. Yan JB, Wu B, He QH (2021) An anisotropic IDW interpolation method with multiple parameters cooperative optimization. Acta Geodetica et Cartographica Sinica 50(5): 675-684.
25. Yang N, Zhang Z, Yang J, Hong Z (2022) Applications of data augmentation in mineral prospectivity prediction based on convolutional neural networks. Computers & Geosciences 165: 105075.
26. Zhang HY, Cisse M, Dauphin YN, Lopez-Paz D (2017) mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412.
27. Zhu L, Chen YS, Ghamisi P, Benediktsson JA (2018) Generative adversarial networks for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing 56(9): 5046-5063.

Figures

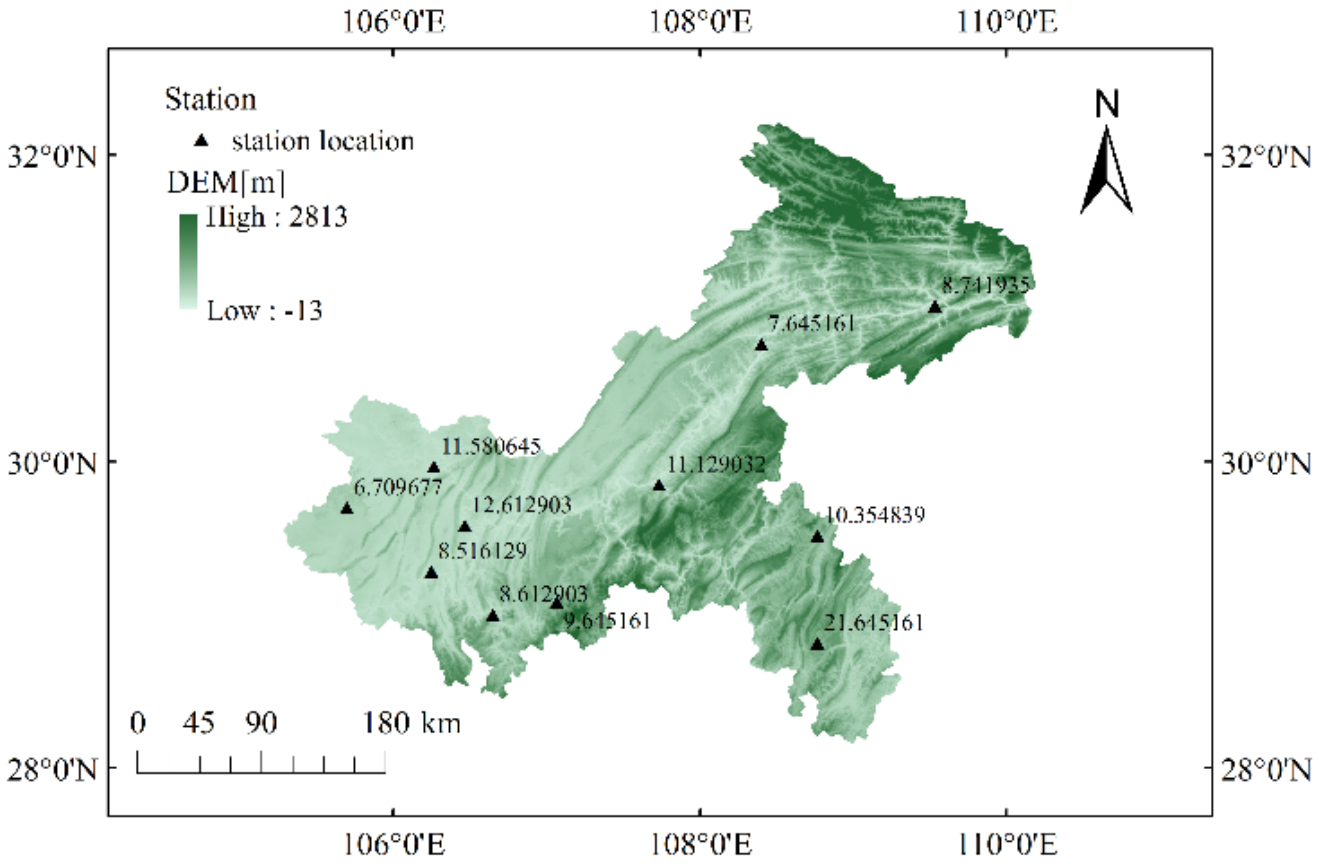


Figure 1

DEM and Station Location Map of Chongqing

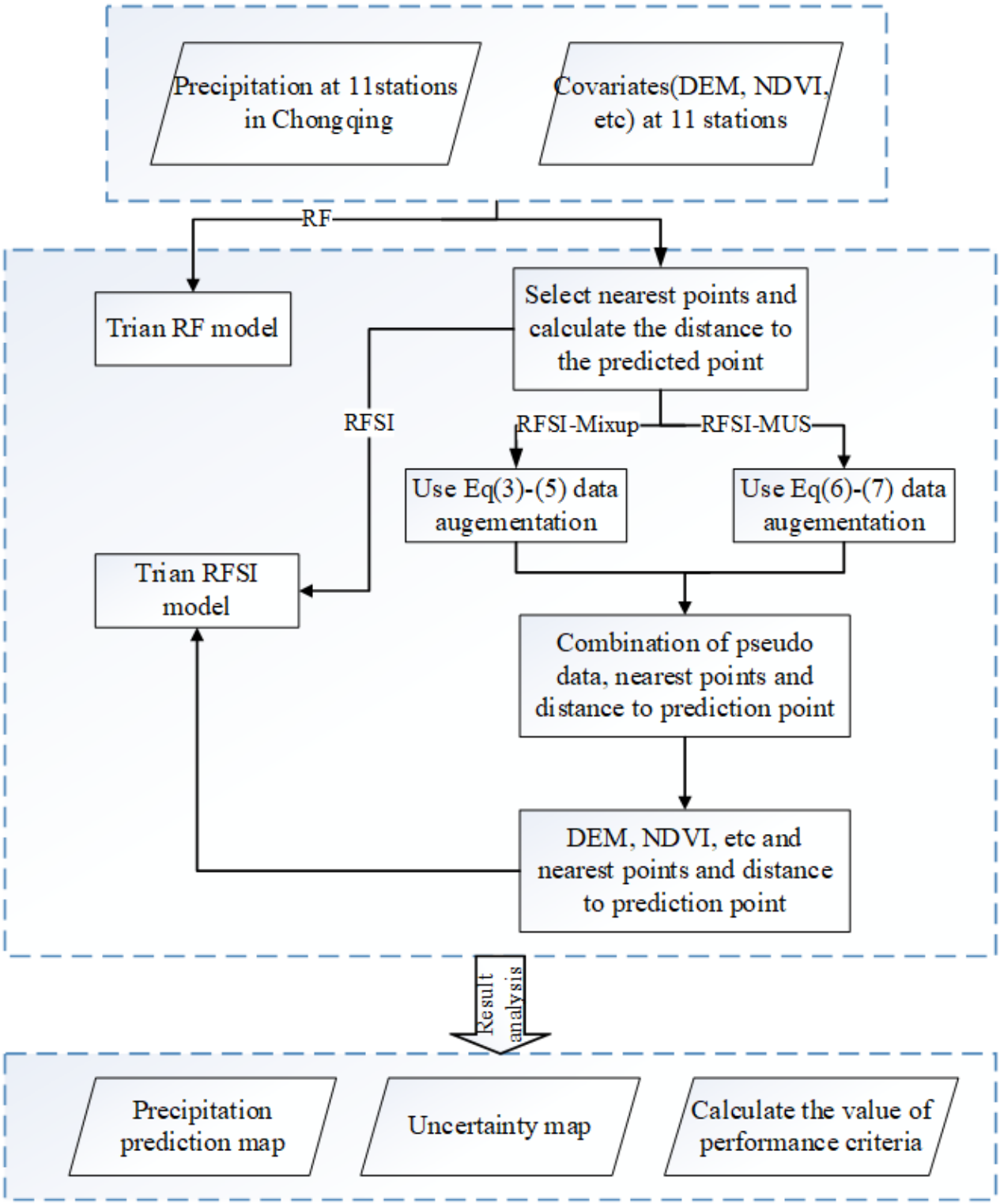
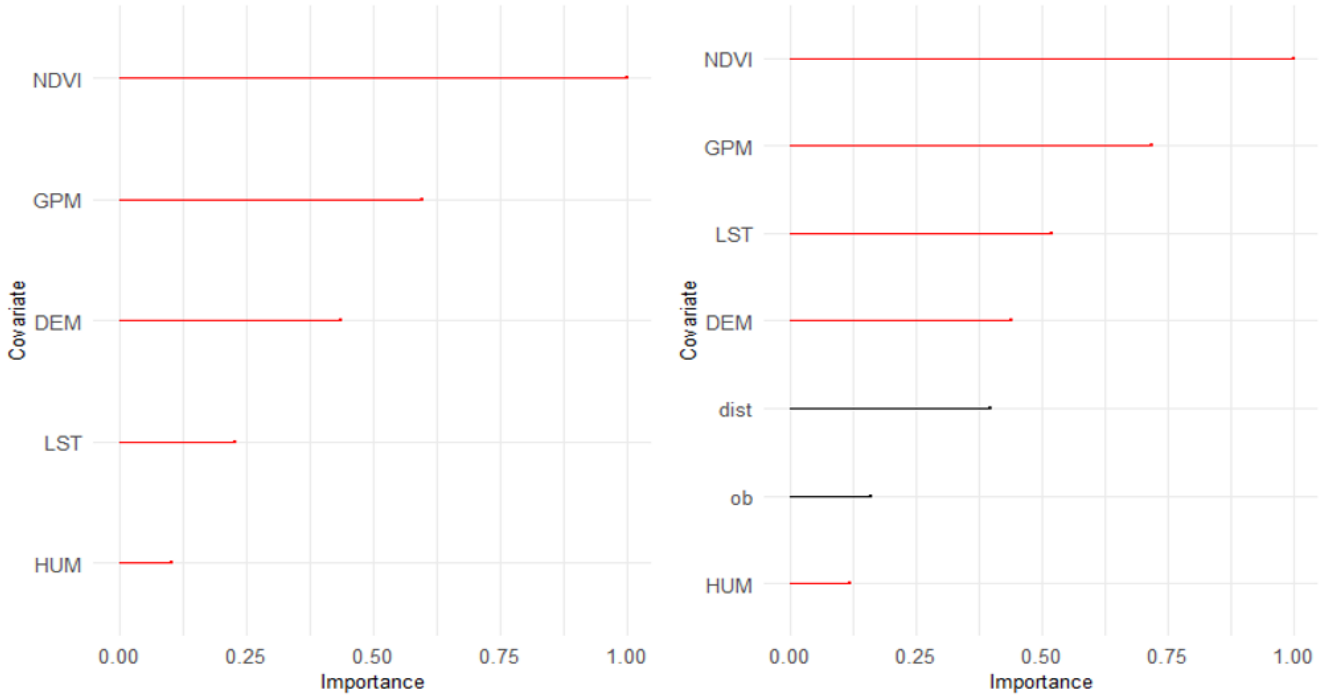


Figure 2

Schematic representation of this paper based on Chongqing precipiatation



(a)RF covariates importance (b) RFSI covariates importance

Figure 3

Covariates importance for the precipitation of Chongqing: (a)RF, (b)RFSI. The importance index is scaled to a maximum of 1.

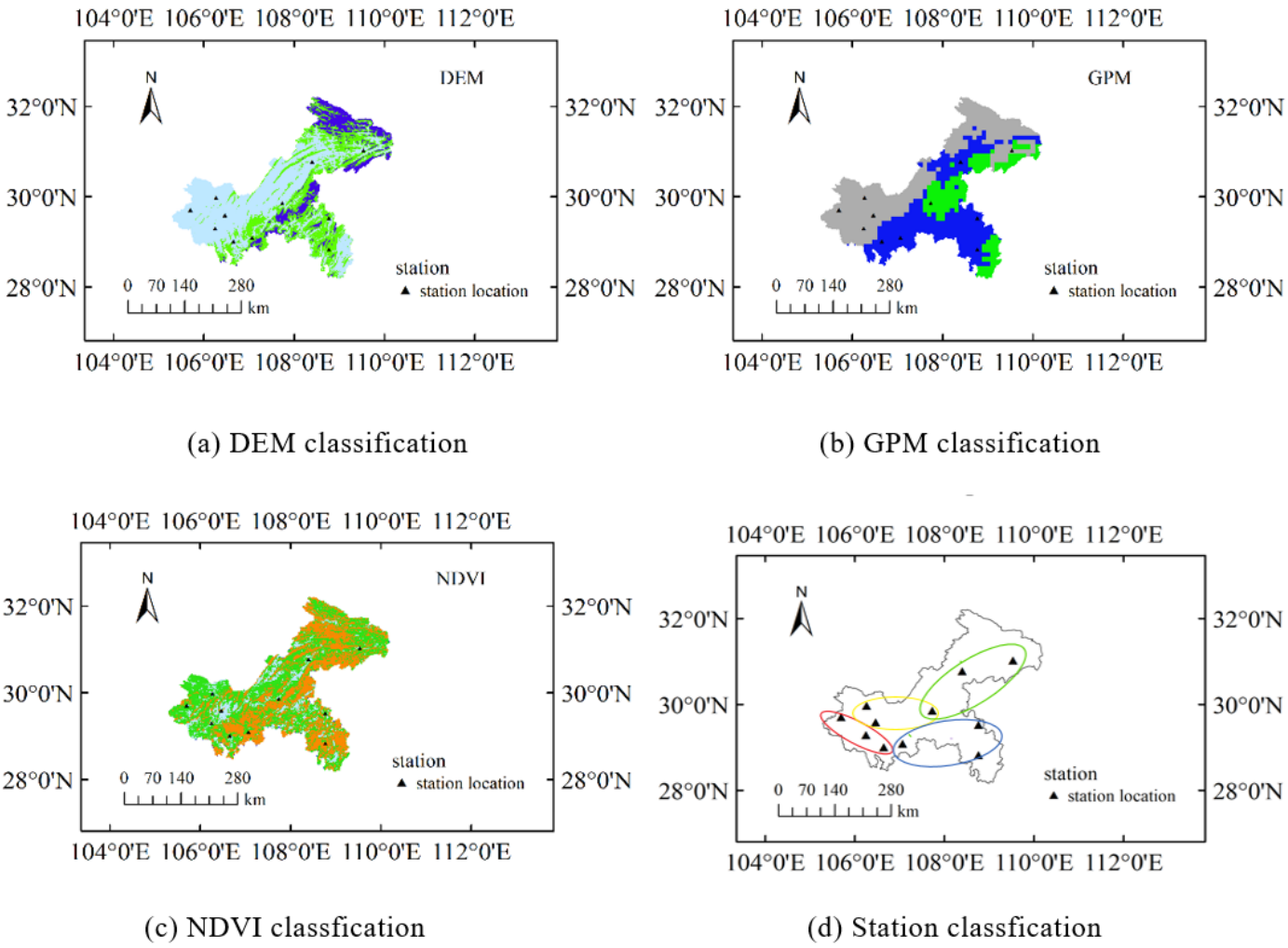


Figure 4

Covariates and station classfication for Chongqing. (d) is obtained from (a), (b), (c), precipitation and spatial distance.

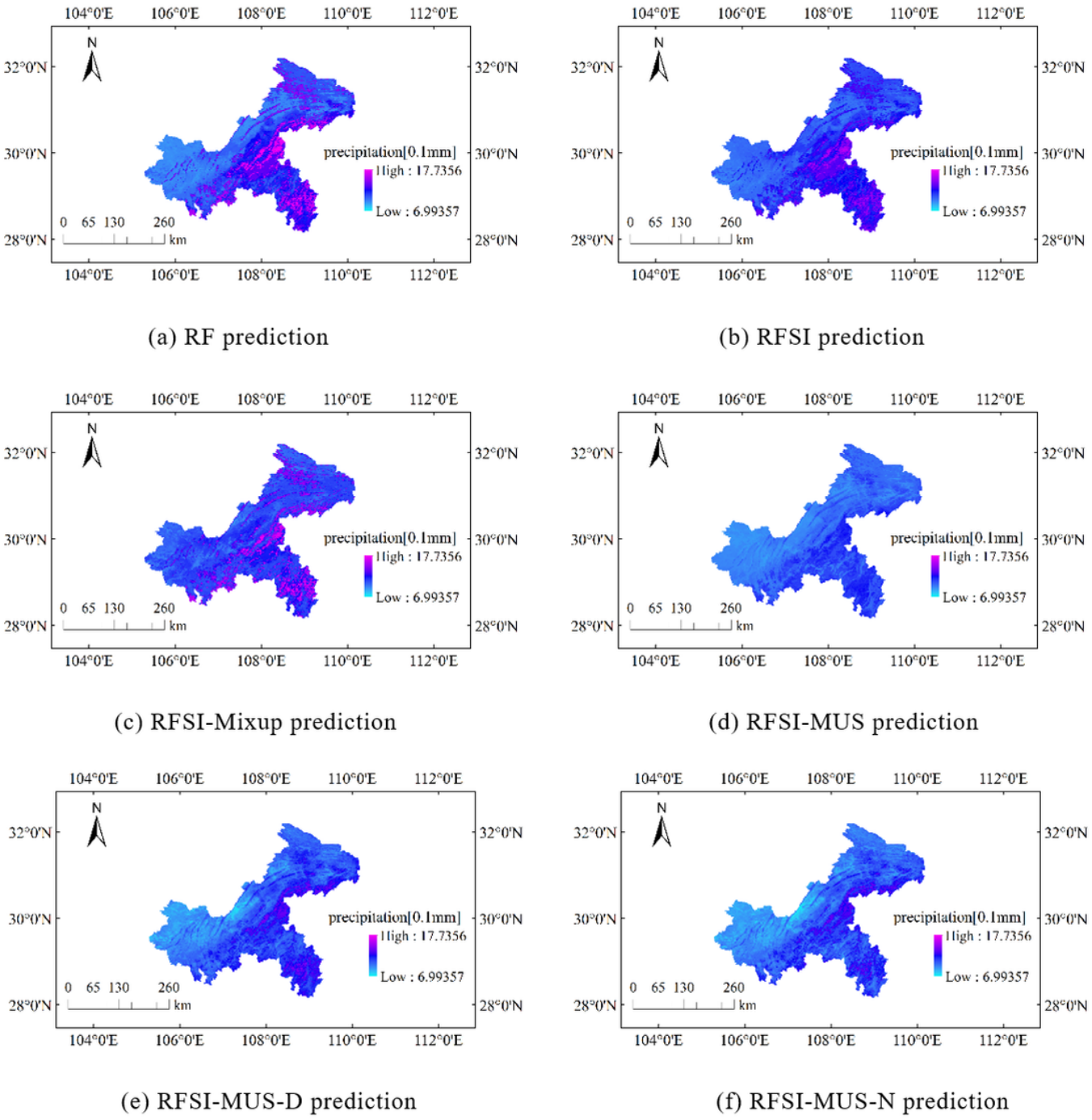


Figure 5

Spatial prediction results of Chongqing precipitation: (a) RF, (b) RFSI,(c) RFSI-Mixup, (d)RFSI-MUS, (e) RFSI-MUS-D, (f) RFSI-MUS-N

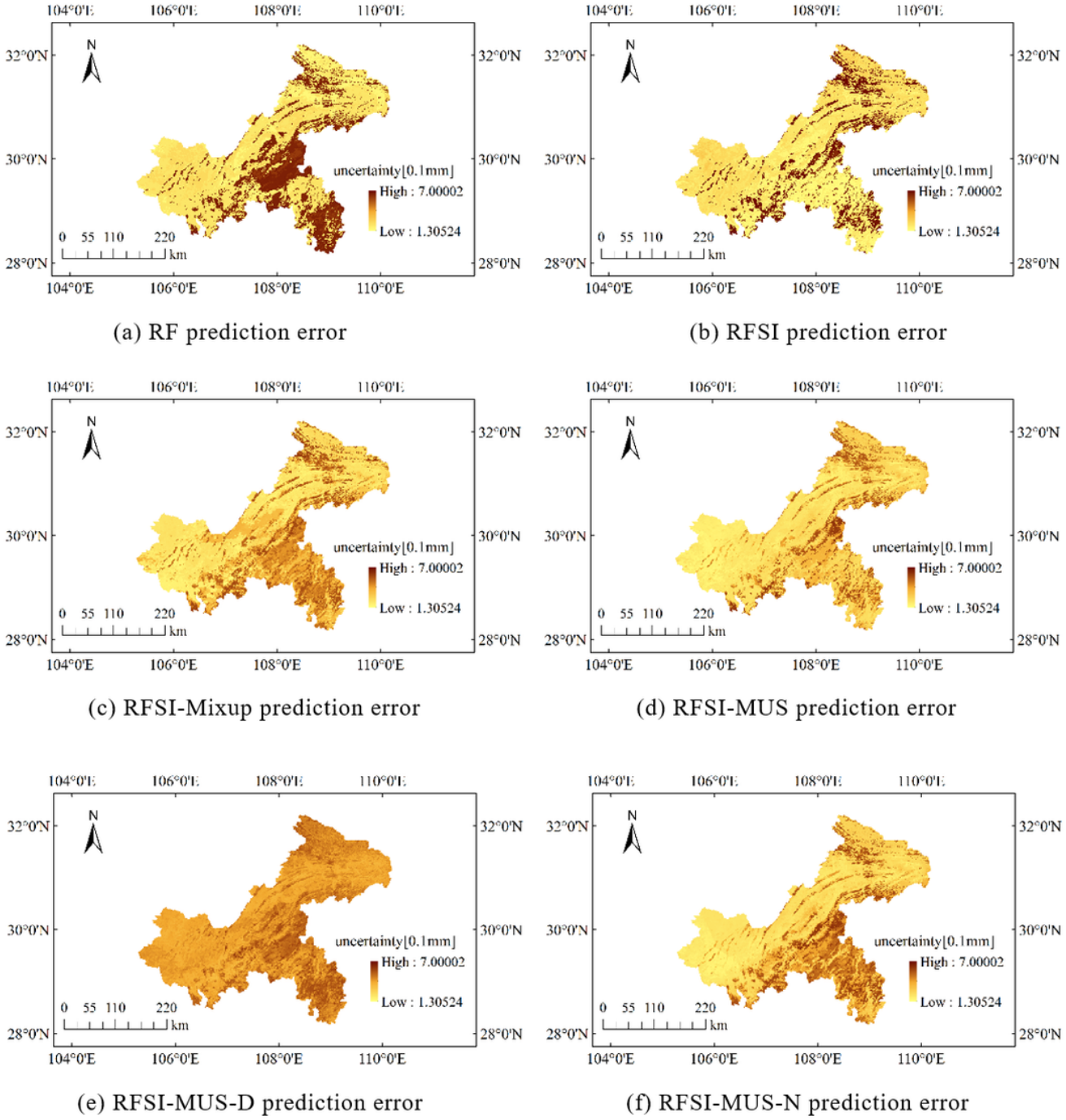


Figure 6

Prediction standard error for Chongqing precipitation: (a) RF, (b) RFSI, (c) RF-Mixup, (d) RFSI-MUS (e) RFSI-MUS-D, (f) RFSI-MUS-N

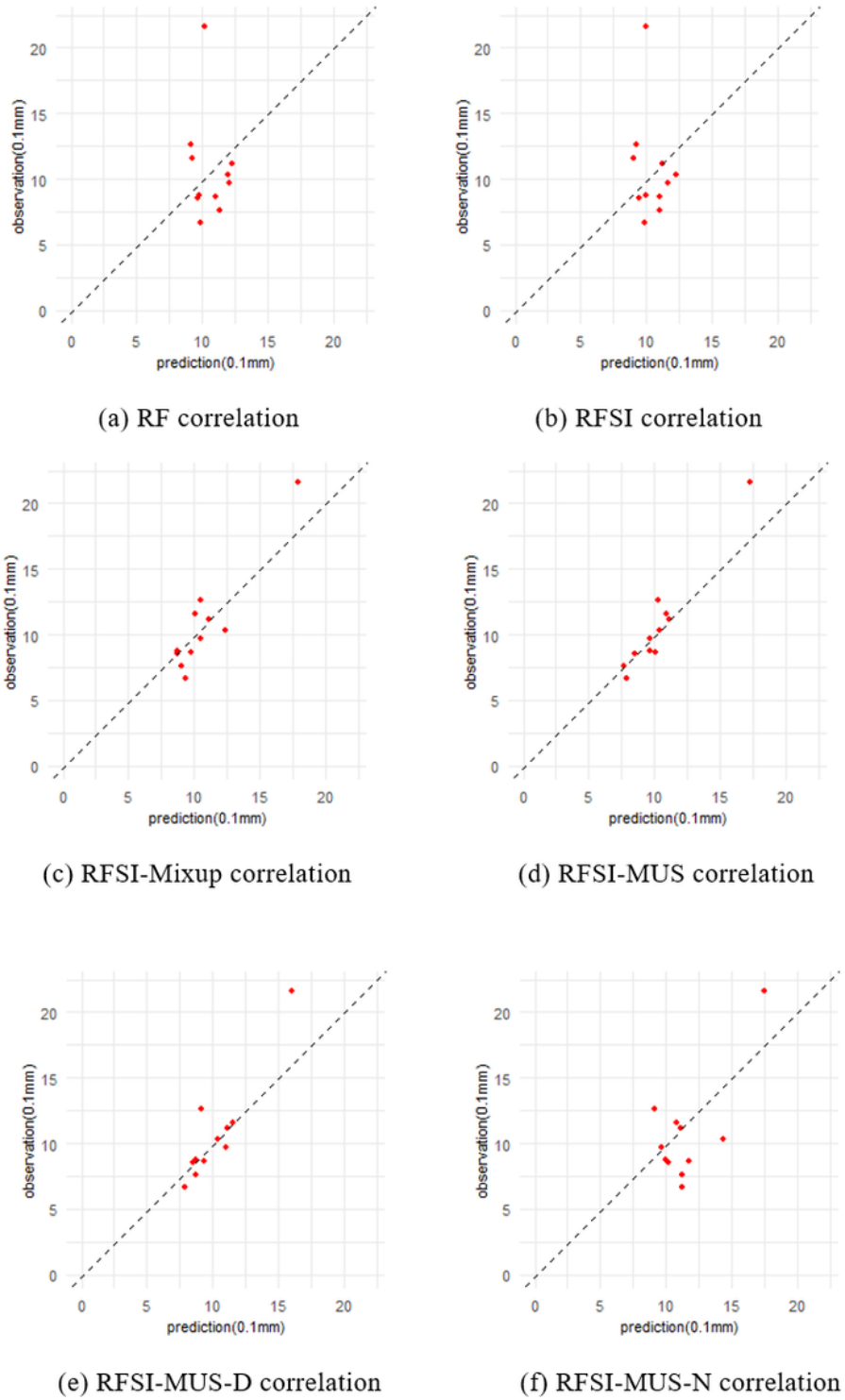


Figure 7

Correlation plots based on observations and predictions: (a) RF, (b) RFSI, (c) RFSI-Mixup,(d) RFSI-MUS, (e) RFSI-MUS-D, (f) RFSI-MUS-N correlation