Contents lists available at ScienceDirect

Environmental Modelling and Software

journal homepage: www.elsevier.com/locate/envsoft

A spatiotemporal estimation method for hourly rainfall based on F-SVD in the recommender system

Hua Chen [a,b,c,*], Sheng Sheng [a,b], Chong-Yu Xu [d,**], Zhiyu Li [c], Wen Zhang [e], Shaowen Wang [c], Shenglian Guo [a,b]

[a] *State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, Wuhan, China*
[b] *Hubei Provincial Key Lab of Water System Science for Sponge City Construction, Wuhan University, Wuhan, 430072, China*
[c] *Department of Geography and Geographic Information Science, University of Illinois at Urbana-Champaign, 61801, United States*
[d] *Department of Geosciences, University of Oslo, Norway*
[e] *College of Informatics, Huazhong Agricultural University, Wuhan, China*

ARTICLE INFO

Keywords:
Rainfall estimation
Spatiotemporal interpolation
Recommendation system
Singular value decomposition
Matrix factorization

ABSTRACT

In this study, a spatiotemporal estimation method based on Funk singular value decomposition (F-SVD) that considers the spatiotemporal correlation of rainfall is proposed to improve estimations from gauge observations. Hourly rainfall data of several flood events are selected to verify the proposed method by comparing with Inverse Distance Weighting (IDW) and Ordinary Kriging (OK) in Hanjiang basin, China. The results show that (1) F-SVD has the best performance in rainfall estimation, the larger the amount of rainfall event, the greater the improvement of F-SVD method as compared to OK and IDW; (2) through the combination/integration with F-SVD, the accuracy of IDW and OK can be greatly improved. Therefore, F-SVD can be employed as a practical method to estimate rainfall spatial distribution, which is essential data for regional hydrological modelling and water resource analysis.

1. Introduction

Rainfall is one of the main sources of the water system and a key component of the water cycle on the earth (Diez-Sierra and del Jesus 2017). Due to the influence of natural climate, terrain and underlying surface, the spatial and temporal distributions of rainfall on the earth's surface are uneven, where apparent characteristics of regional and temporal variation exist. The spatial and temporal distribution of rainfall is of great significance for maintaining the life and health of all biological communities (Dai et al., 2020). It is also one of the most critical data sources for hydrological scientific research, water resources management, drought and flood disaster management, and ecological environment governance (Sivakumar and Woldemeskel 2015). Currently, the most common way of rainfall observation and collection is a rainfall gauge network, of which the main features are convenient, real-time and accurate. However, since the gauges are discrete, spatial calculation methods are needed to obtain the spatially continuous rainfall distribution. The spatial interpolation method plays a vital role in the calculation of the spatial distribution of rainfall data and has been widely concerned by many scholars (Ahrens 2006; Garcia et al., 2008; Kumari et al., 2016; Morris et al., 2016).

The spatial interpolation methods for rainfall are based on Tobler's First Law of Geography (Tobler 1970) that points closer in space are more likely to have similar eigenvalues, and points farther away are less likely to have similar eigenvalues, such as Tyson Polygon (Thiessen 1911), Inverse Distance Weight (IDW) (Shepard 1968) and Kriging (Delhomme 1978), are among the most widely used methods in the spatial estimation for rainfall (Cai et al., 2018; Carrera-Hernández and Gaskin 2007; Foehn et al., 2018; Goovaerts 2000; Plouffe et al., 2015; Ryu et al., 2021; Zhang et al., 2018). While during the process of rainfall, not only points adjacent in space are more likely to have similar characteristics, but also points contiguous in time are more likely to have a consistent variation trend. Therefore, to obtain more accurate results, both the spatial dimension and the time dimension should be considered during interpolating. Many researches have been carried out on the spatiotemporal estimation method for rainfall, such as Space-time Autoregressive Moving Average model (STARMA) and Kriging interpolation method with time extension (Cliff and Ord 1975; Dalezios and

* Corresponding author. State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, Wuhan, China.
** Corresponding author.
E-mail addresses: chua@whu.edu.cn (H. Chen), chongyu.xu@geo.uio.no (C.-Y. Xu).

https://doi.org/10.1016/j.envsoft.2021.105148
Accepted 24 July 2021
Available online 28 July 2021

Adamowski 1995; Pfeifer and Deutrch 1980). Dalezios and Adamowski (1995) applied STARMA models in spatiotemporal rainfall modelling. Bargaoui and Chebbi (2009) proposed a 3-dimensional variogram (Location-Duration-Intensity) to replace the traditional 2-dimensional variogram (Location-Intensity), which can effectively consider the spatial variability of the maximum rainfall intensity in a given duration range and significantly reduce the rainfall prediction error. Spadavecchia and Williams (2009) compared simple Kriging (SK), ordinary Kriging (OK) and space-time Kriging with an external drift using a residual variogram with spatiotemporal lags in the interpolation of meteorological variables. Besides, some scholars combined the time machine learning method with spatial simulation to realize the spatiotemporal interpolation of rainfall. Xu et al. (2019) proposed a novel spatiotemporal prediction model based on the cubic spline method and the spatiotemporal echo state networks, which showed advantages in predicting meteorological series over other spatial estimation models. In general, some progress has been made in the spatiotemporal estimation method for rainfall, and the existing results have shown that the spatial-temporal interpolation methods have higher accuracies than those spatial interpolation methods without considering the time dimension.

The superiority of spatial-temporal interpolation methods for rainfall has been proved, however, compared with extensive application of the spatial estimation methods for rainfall, the existing spatiotemporal estimation methods for rainfall are too complex to be widely applied due to strong randomness and complexity of the rainfall process. In order to more conveniently and widely use the spatial-temporal estimation methods for rainfall, there are still exploration works to be worth doing for them. If the rainfall data at different times of each gauge are putting together, it can be found that an enormous two-dimensional spatiotemporal matrix is formed, where the rows are the time dimension and the columns are the spatial dimension, as shown in Fig. 1(a). From Fig. 1 (a), it can be seen that the rainfall process has a clear correlation in the time (T) and space dimension (S). It is possible to transform the spatial-temporal interpolation problem into a two-dimensional matrix solution problem. For example, if a certain value in a matrix is missing, it can be calculated through matrix factorization technology, which has been widely used in the e-commerce recommender system (Fig. 1 (b)).

A common task of the recommender system is to improve customer experience through personalized recommendations based on historical interactions and prior implicit feedback (Hu et al., 2008). These interactions are stored in the so-called “user-item interactions matrix” (Fig. 1(b)), which are used in many famous e-commerce platforms to recommend relevant products to users, such as Amazon, Taobao, Joybuy, Youtube, and Netflix etc. Its algorithms are mainly divided into two categories: collaborative filtering methods and content-based methods. And the collaborative filtering (CF) algorithm gains an advantage due to its insensitivity to content (Yu et al., 2018), which predicts user preferences for products by learning known user-item relationships (Bell and Koren 2007). Matrix Factorization (MF) technique is one of the most popular approaches for solving the problem of CF, which views user preference ratings of items as a user-item matrix and uses known user ratings of items to predict user preferences in item selection (Takacs et al., 2009). As MF in the recommender system has high prediction accuracy, it has become recognized as a mature method in environmental science, biomedicine and many other fields (Xie and Berkowitz 2006; Xue et al., 2014; Zhang et al., 2019). González-Macías et al. (2014) used the positive matrix factorization approach in identification and source apportionment of the anthropogenic heavy metals in the sediments of sea. Lee et al. (2012) applied non-negative matrix factorization to new gene expression data quantifying the molecular changes in four tissue types due to different dosages of an experimental panPPAR agonist in mouse. Yeh et al. (2018) proposed a rain removal method based on non-negative matrix factorization to improve image quality. Funk Singular Value Decomposition model (F-SVD), proposed by Funk (2006), is a variant of MF that outperforms other models in the Netflix Prize competition. The essential idea incorporated in F-SVD of MF is that users and items can be described by their latent features vectors inferred from rating matrix, and the high correspondence between user and item features leads to recommendation (Koren et al., 2009). As rainfall data can be viewed as an intrinsically related matrix, F-SVD is a good way to estimate an unknown point in the spatiotemporal rainfall matrix. From Fig. 1, it can be seen that there are similar interactions between spatiotemporal rainfall matrix and the user-item interactions matrix. In this study, F-SVD in the recommender system is regarded as a potential spatiotemporal method to estimate the rainfall for the first time, and its performance is evaluated by compared with IDW and OK methods.

This paper aims to present a new approach using information of points both adjacent in space and contiguous in time to estimate rainfall more accurately and obtain continuous spatial distribution of rainfall, which is helpful for regional hydrological modelling and spatial statistical analysis. The rest of this paper is structured as follows. Section 2 introduces the proposed spatiotemporal interpolation method based on F-SVD and its implementation steps. Section 3 introduces the study area, data and evaluation indicators. Section 4 analyzes and discusses the results of spatiotemporal interpolation. Finally, section 5 summarizes the results of the study and presents existing problems and suggestions.

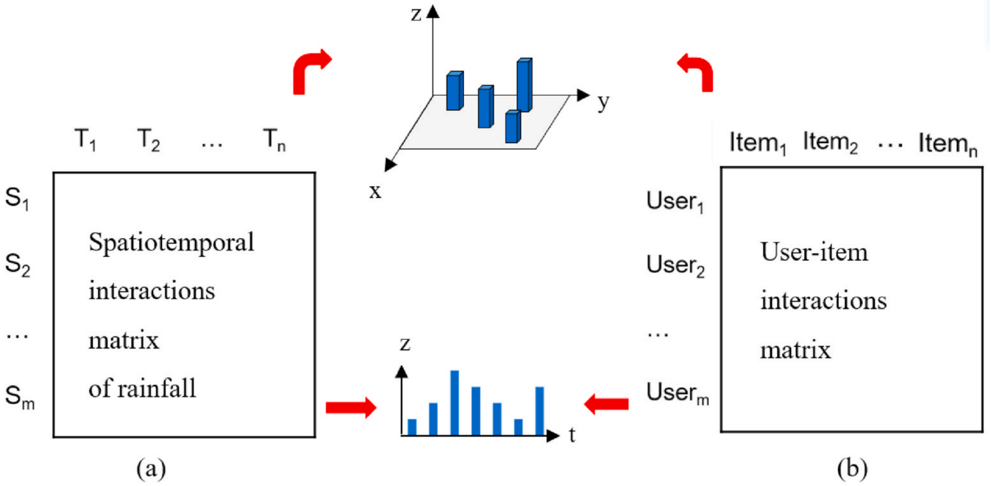

z
y
x
T1 T2 … Tn
S1
S2
…
Sm
Spatiotemporal interactions matrix of rainfall
z
t
Item1 Item2 … Itemn
User1
User2
…
Userm
User-item interactions matrix
(a)
(b)

Fig. 1. Spatiotemporal interactions matrix of rainfall and user-item interactions matrix.

2. Spatiotemporal estimation method based on F-SVD

2.1. Rainfall spatiotemporal estimation expressed by F-SVD

Rainfall data can be viewed as an intrinsically related matrix, with columns representing time and rows representing space, respectively. There is an implicit interactions relationship between time and space, which cannot be figured out directly, thus latent factors are needed to establish an indirect relationship. There is an example to illustrate the basic ideas of F-SVD to achieve spatiotemporal estimation of rainfall, as shown in Fig. 2. The rainfall records of 5 gauges with 5 moments form a spatiotemporal two-dimensional matrix R, and the rainfall at gauge S_5 at moment T_5 is supposed to be unknown marked with red box in R. Firstly assuming there are latent factors affecting the spatiotemporal distribution of the rainfall at each gauge and each moment, they can't be observed directly from R. While they can be decomposed from R by using F-SVD. In this case the derived latent spatial factors matrix X and temporal factors matrix Yconsist of 5 row vectors x_i and column vectors y_i ($\mathrm{i} = 1, \ldots,5$), which represent the interactions relationships of latent factors in space and time. Then, it can be assumed that rainfall values in R can be derived from latent factors X and Y by multiplying them. For example, there exist latent factors x_5 and y_5 for gauge S_5 and moment T_5 marked with red box in X and Y, which can be deduced from spatiotemporal interactions relationships in R by using F-SVD. The supposed unknown values at gauge S_5 and moment T_5 can be estimated by multiplying latent factors x_5 and y_5, whose estimation is 5.1 mm.

By comparing the estimation matrix R' and original matrix R, it can be seen that there are estimation errors between them, while they are relatively small. For example, for unknown points, the relative error is -8.9 %, which indicates the high accuracy of F-SVD in rainfall estimation. Of course, this is a specific case, and the applicability of the F-SVD method needs to be further verified in the following sections.

2.2. Proposed spatiotemporal estimation model based on F-SVD

The objective of the proposed spatiotemporal estimation method based on F-SVD, is to use the current time and historical information to interpolate rainfall at the target points. The steps are as follows, where the transformation of variables involved is shown in Fig. 3.

(1) For the estimation of rainfall at position i at moment j, a spatiotemporal interactions matrix R sized of $m \times n$ which consists of rainfall of m positions at n moments is needed. The positions and moments corresponding to the rows and columns of the matrix have to be figured out first and the method is shown below.

In matrix R the m positions consists of target position and rainfall gauges around, which belong to the optimal set chosen by the spatial uniformity L after running through all the possible permutations. Randomly select $m-1$ gauges from all available gauges and combine them with the target position to form a set of m points. For the formed set, L is calculate based on the spatial distribution of points drawn according to the latitudes and longitudes. L is a measure of spatial relationship of point set, and the larger the L value, the more evenly distributed the points. By listing all possible combinations and calculating the spatial uniformity, the point set with the largest L is the optimal set. L can be defined as

$$L = \frac{4a}{\pi A} \tag{1}$$

In the equation above, A represents the area of the grid rectangle that contains all the points, and a indicates the total area of exclusive circles, which is defined for each point in the set as a circle with the center of itself and a radius of half the distance from the nearest adjacent point.

As for the set of previous moments T_n involved, considering the efficiency and accuracy of matrix factorization, it is determined by N according to the following rules:

$$T_n \quad \begin{cases} n = tstart, tstart + 1, \ldots j, & j - tstart < N \\ n = j - N, j - N + 1, \ldots j, & j - start \geq N \end{cases} \tag{2}$$

where *tstart* denotes the starting time of rainfall event. If the rainfall event does not last long, then n ranges from the starting time *tstart* to the interpolated time j. Else if the event lasts longer than N, for moment j to which over N hours passed from the starting time, n includes N moments before j and j itself.

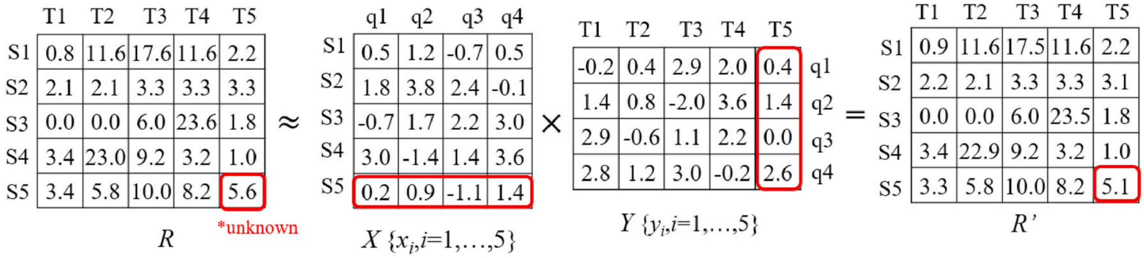
(2) After obtaining the correspondence between matrix rows and positions, matrix columns and moments, the rainfall data used for interpolation need to be filled into the matrix accordingly. For the $m-1$ rainfall gauges involved, the rainfall data before and at the interpolation time are filled directly into the related locations in the matrix. For the target position, if there are observation records before the interpolation time, then the observed data are directly filled into the corresponding locations, in which way only F-SVD is used for interpolation; else the traditional method such as IDW has to be used to interpolate the historical rainfall first, and then the historical interpolation result is filled into the matrix, in which way the F-SVD is integrated with traditional method. After the data filling is completed, only the position corresponding to the target point and the interpolated time in the matrix is a null value.

(3) Based on the F-SVD model, the matrix R is decomposed into spatial feature matrix X and temporal feature matrix Y by computing the relationships of q latent features in time and space through minimizing squared error on all known rainfall. Moreover, in case of the phenomenon of over-fitting, regularization method is introduced to the objective function:

$$E^2{}_{i,j} = \left(R_{i,j} - R'_{i,j}\right)^2 = \left(R_{i,j} - \sum_{q=1}^{q} X_{i,q} Y_{q,j}\right)^2 \tag{3}$$

$$\min : SSE = \sum_{i=1}^{m+1} \sum_{j=1}^{n} E^2_{i,j} + \lambda \sum_{i,q} \left|X_{i,q}\right|^2 + \lambda \sum_{q,j} \left|Y_{q,j}\right|^2 \tag{4}$$

where SSE denotes the loss function and λ is a hyper-parameter that controls the degree of regularization.



	T1	T2	T3	T4	T5
S1	0.8	11.6	17.6	11.6	2.2
S2	2.1	2.1	3.3	3.3	3.3
S3	0.0	0.0	6.0	23.6	1.8
S4	3.4	23.0	9.2	3.2	1.0
S5	3.4	5.8	10.0	8.2	5.6

R *unknown

≈

	q1	q2	q3	q4
S1	0.5	1.2	-0.7	0.5
S2	1.8	3.8	2.4	-0.1
S3	-0.7	1.7	2.2	3.0
S4	3.0	-1.4	1.4	3.6
S5	0.2	0.9	-1.1	1.4

$X\ \{x_i, i=1,\ldots,5\}$

×

T1	T2	T3	T4	T5	
-0.2	0.4	2.9	2.0	0.4	q1
1.4	0.8	-2.0	3.6	1.4	q2
2.9	-0.6	1.1	2.2	0.0	q3
2.8	1.2	3.0	-0.2	2.6	q4

$Y\ \{y_i, i=1,\ldots,5\}$

=

	T1	T2	T3	T4	T5
S1	0.9	11.6	17.5	11.6	2.2
S2	2.2	2.1	3.3	3.3	3.1
S3	0.0	0.0	6.0	23.5	1.8
S4	3.4	22.9	9.2	3.2	1.0
S5	3.3	5.8	10.0	8.2	5.1

R'

Fig. 2. Example of rainfall spatiotemporal estimation using F-SVD.

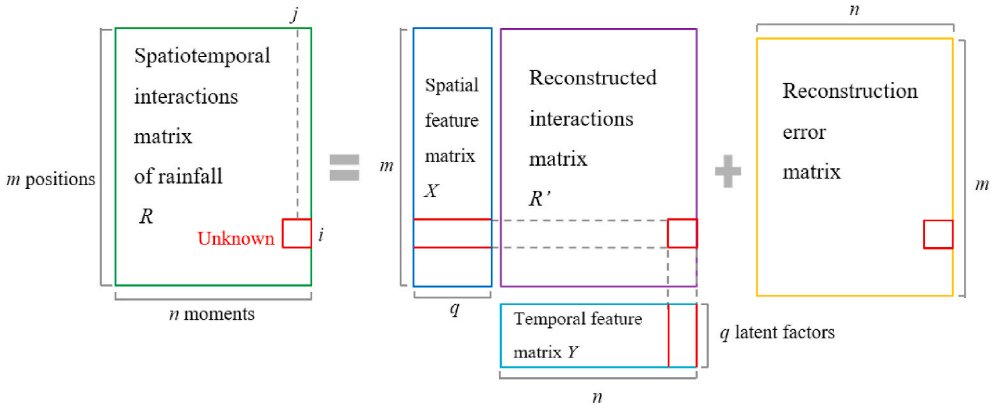

j
Spatiotemporal interactions matrix of rainfall
R
Unknown
i
m positions
n moments
m
Spatial feature matrix
X
q
Reconstructed interactions matrix
R'
n
Reconstruction error matrix
m
Temporal feature matrix Y
q latent factors
n

Fig. 3. The sketch diagram of the proposed spatiotemporal estimation model based on F-SVD.

In order to minimize *SSE*, stochastic gradient descent (SGD) algorithm is chosen to solve the optimization problem above. A summary of this method is available as a flowchart here in Fig. 4. $X_{i,q}$ and $Y_{q,j}$ will decrease in the direction of the fastest one in which the gradient descent and therefore the optimal solution can be inferred. To learn the optimum value of spatial feature vector $X_{i\cdot}$, rainfall of all known time $P_{i,j}(j=1,2\ldots n)$ are used to factorized, that is to say, the value of each component in spatial feature vector $X_{i\cdot}$ is related to all temporal feature vector $Y_{\cdot j}(j=1,2\ldots n)$ that extracted from historical rainfall information. The equations are as follow:

$$X_{i,q} = X_{i,q} - \alpha \frac{\partial E_{i,j}^2}{\partial X_{i,q}} = X_{i,q} + 2\alpha\left(E_{i,j}Y_{q,j} - \lambda X_{i,q}\right) \tag{5}$$

$$Y_{q,j} = Y_{q,j} - \alpha \frac{\partial E_{i,j}^2}{\partial Y_{q,j}} = Y_{q,j} + 2\alpha\left(E_{i,j}X_{i,q} - \lambda Y_{q,j}\right) \tag{6}$$

In the equations above, α indicates the learning rate in machine learning.

(4) The spatial feature matrix X and temporal feature matrix Y are multiplied to obtain the optimal reconstructed interactions matrix $R^{'}$ and each element in it has a value. A one to one correspondence exists between the elements in R and $R^{'}$, that is, the value of the element in spatiotemporal interactions matrix R is equal to that of reconstructed interactions matrix $R^{'}$ plus reconstruction error matrix. Hence, the value of row i, column j in $R^{'}$ is the estimated rainfall of point i at moment j.

2.3. Evaluation methods

In this study, two widely used spatial interpolation approaches including IDW and OK are adopted as benchmark methods for comparison without considering the temporal change trend. F-SVD can estimate the rainfall value of one site by using the spatiotemporal matrix of rainfall, which can be applied to the estimation of missing rainfall value or the test of rainfall abnormal value for sites. When it is applied to the interpolation of unknown points in space, it needs to be combined with the existing spatial interpolation methods to obtain more accurate spatial interpolation results. In order to evaluate the performance of the combination of F-SVD with the spatial interpolation methods, this study considered the combined use of F-SVD with IDW and OK, respectively named F-SVD-IDW and F-SVD-OK.

The leave-one-out cross validation method was adopted to assess the accuracy. In this process, each time a record of one gauge from the dataset was removed and then be assumed using the information of all the gauges left. Then the interpolation results were compared to the observations to evaluate the estimation error using four statistical measures, namely root-mean-square error (RMSE), mean average error (MAE), percentage error (PERC) and two-sample Kolmogorov-Smirnov test statistic (KS). Among these statistical measures, the two-sample Kolmogorov-Smirnov test is a non-parametric test that compares whether there is a significant difference between two samples based on the empirical distribution function, and it is applicable and even for small sample sizes (Engmann and Cousineau 2011). The calculation formulas of each indicator are as follows:

$$RSME = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(z_i^{sim} - z_i^{obs}\right)^2} \tag{7}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|z_i^{sim} - z_i^{obs}\right| \tag{8}$$

$$PERC = \frac{1}{n_1 + n_2}\left(\sum_{i=1}^{n_1}\left|\frac{z_i^{sim} - z_i^{obs}}{z_i^{obs}}\right| + n_2\right) \tag{9}$$

$$\begin{cases} \sup\limits_{x\in R}\left|F_1(x) - F_2(x)\right| \le d_p, KS_i = 0 \\ \sup\limits_{x\in R}\left|F_1(x) - F_2(x)\right| > d_p, KS_i = 1 \\ KS = \frac{1}{n}\sum_{i=1}^{n} KS_i \end{cases} \tag{10}$$

where z_i^{obs} and z_i^{sim} denote the observed value and the interpolated value at the i-th gauge; n_1 and n_2 represent the number of records which are non-zero and records where a measured zero is not predicted; F_1 and F_2 indicate the distribution functions of calculation sequence and observation sequence of the i-th gauge; d_p is the critical value at the significance level $p = 5\%$.

For all measures, the smaller the value, the better the results. The low values of the first three measures indicate that the errors of the interpolation results are small, and the closer the fourth measure is to 0, the fewer gauges with significant errors in the interpolation and the measured sequence.

3. Study region and data

The study region is the upstream of the Hanjiang basin (Fig. 5), which is the largest tributary in the Yangtze River and the water source of the Middle Route Project of South to North Water Transfer, China (Chen et al., 2007). It originates from Qinling Mountain and is located in the southeast of China between east longitude of 106°15′–112°00′ and north latitude of 31°40′–34°20’. The entire drainage area of the study region is about 96,000 km^2. Influenced by geographical factors, the basin has a subtropical monsoon climate with humid air and abundant

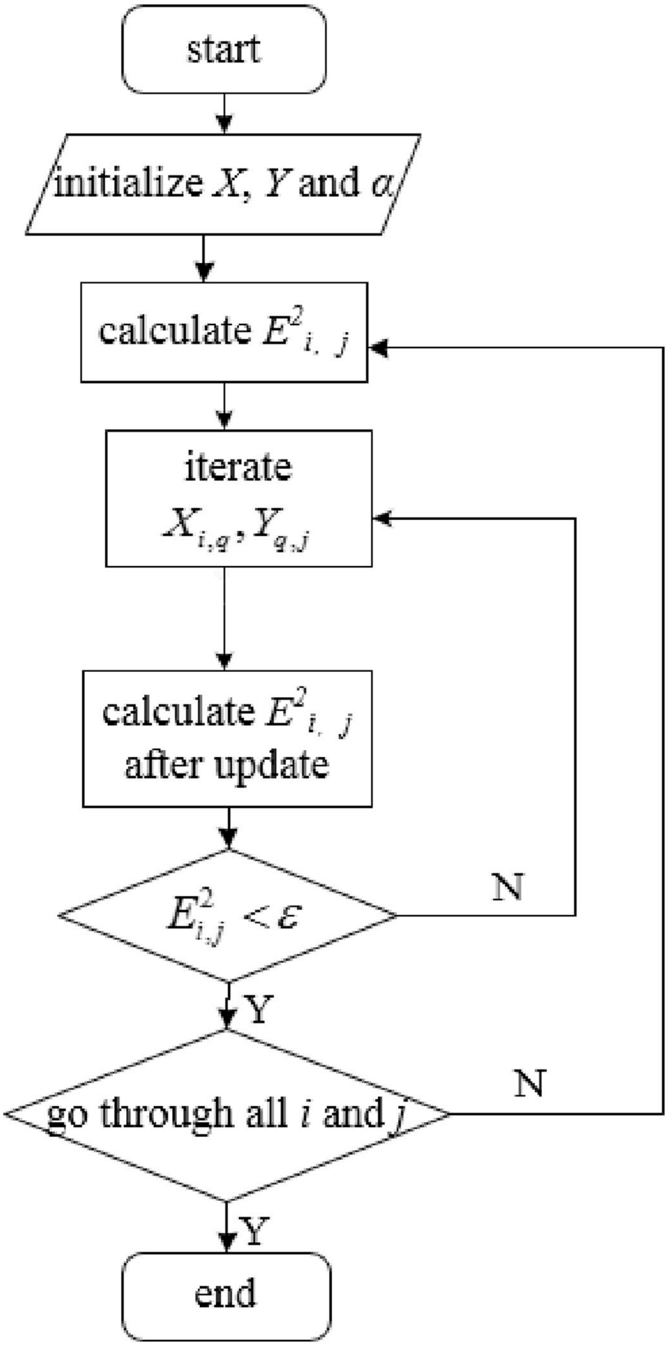


Fig. 4. Flow chart of stochastic gradient descent method.

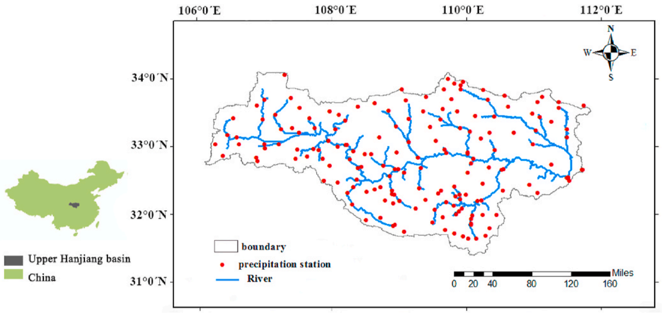


Fig. 5. Study area and geographical distribution of rainfall gauges in Hanjiang basin.

rainfall. The annual average rainfall is approximately 830 mm, decreasing from south to north.

The spatial distribution of 176 rainfall gauges is shown in Fig. 5, hourly rainfall data from several flood events were selected during the period from 2012 to 2018 under different meteorological conditions in Hanjiang basin. The starting and finishing time of rainfall events were determined according to the division of flood event based on hydrological hydrographs observed at the outlet streamflow gauge of the study area. The 20 selected events as well as some statistics are listed in Table 1. It can be seen from the table that the selected rainfall events are distributed from April to October, spanning the three seasons of spring, summer and autumn, including the main rainfall season in the Hanjiang River Basin (May to October); thus have a good representation. The duration of different rainfall events spans a wide range, with the shortest being only 16 h and the longest reaching 106 h. The percentage of gauges with a cumulative observed rainfall of zero varies from less than 5 to more than 20, which represents the different spatial concentration of rainfall.

4. Results and discussions

4.1. Sensitivity analysis on the value of m and N

The F-SVD method requires two parameters m and N to determine the size of the matrix when interpolating, where m is the number of gauges involved in interpolation and is related to the number of rows of the matrix; combining N and interpolation time j can determine the number of columns in the matrix. To obtain the best interpolation result, various values of these two parameters are selected for calculation and comparison. Table 2 lists the average spatial uniformity of all stations when m takes different values. As the value of m increases, the spatial uniformity increases first and then decreases. When the value of m is 20, L reaches the maximum value, which means the distribution of the surrounding stations is the most uniform and can well reflect the spatial information in all directions around the interpolation point in this case. Therefore, the value of m in this study is assigned 20.

For the determination of the value of N, to ensure the efficiency of the calculation, three typical long-term events were selected, and the accuracy of interpolation is calculated in different situations. These three events are No. 1, No. 14 and No. 19, respectively, and their duration exceeds 50 h, which are long-lasting rainfall events. Here RSME

Table 1
Information of selected rainfall events with hourly time step data.

Event	Period	Duration	Season	No-rain fraction
NO.	(–)	(h)	(–)	(%)
1	19 Jul-22Jul 2012	60	Summer	26.54
2	7 Sep-8 Sep 2012	30	Autumn	17.90
3	24 May-26 May 2013	64	Spring	4.32
4	24 Jun-25 Jun 2013	32	Summer	14.81
5	17 Jul-20 Jul 2013	88	Summer	8.64
6	18 Apr-19 Apr 2014	32	Spring	21.60
7	1 Sep-2 Sep 2014	32	Autumn	11.11
8	26 Sep-28 Sep 2014	48	Autumn	9.26
9	7 May-8 May 2015	16	Spring	15.43
10	23 Jun-25 Jun 2015	54	Summer	8.64
11	23 Sep-24 Sep 2015	24	Autumn	15.43
12	22 Jun-25 Jun 2016	62	Summer	6.17
13	13 Jul-15 Jul 2016	38	Summer	4.32
14	24 Sep-28 Sep 2016	106	Autumn	9.88
15	2 May-3 May 2017	34	Spring	11.11
16	3 Jun-6 Jun 2017	68	Summer	3.09
17	5 Oct-7 Oct 2017	58	Autumn	3.09
18	23 Sep-27 Sep 2017	106	Autumn	3.70
19	25 May-27 May 2018	50	Spring	9.88
20	17 Jun-19 Jun 2018	44	Summer	2.47

*The proportion of gauges with a cumulative rainfall of zero is listed as no-rain fraction.

Table 2
Average spatial uniformity of all gauges with *m* of different values.

m	16	18	19	20	21	22	24
L	0.2952	0.3061	0.3081	0.3084	0.3062	0.3025	0.2833

is used as the evaluation indicator, and the result is shown in Fig. 6. It is found that as the value of *N* becomes larger, RSME decreases first and then increases, reaching the minimum value, that is, the highest accuracy, when *N* is 24. So *N* is assigned 24 in this study.

4.2. Models evaluation on all rainfall events

The overall estimation results of the five models in 20 rainfall events are comprehensively evaluated using the four different indicators and shown in Table 3. It can be seen from Table 3 that the accuracy of the results of the five methods is within a reasonable range, indicating that they can be well applied to rainfall estimation in the Hanjiang River Basin. The evaluation results using four selected indicators are consistent, with the accuracy of F-SVD being the highest, F-SVD-IDW and F-SVD-OK being the second and third highest, IDW being the fourth highest, and OK being the lowest. The result of IDW is better than OK with a small gap, which is similar to previous research (Hadi and Tombul 2018; Yang et al., 2015). Besides, the difference in accuracy between the five methods is more obvious judged by MAE and RSME since these two indicators are directly related to the amount of rainfall and heavy rainfall usually has a great impact on the values of them. And PERC and KS are less likely to be influenced by the magnitude of rainfall; thus the difference is smaller (Wasko et al., 2013).

The cumulative rainfall of 20 events in the basin and the cross-validation interpolation results of the five methods are shown in Fig. 7. It can be seen that rainfall is mainly concentrated in the southwest of the basin. This is due to the high terrain in the southwest, the warm and humid air flows along the windward slope of the mountain range, plus the influence of the local climate, forming a strong rainfall center in the southwest of the basin. With RSME as the evaluation index, the sites with large cumulative rainfall, mainly distributed in the southwest, have large interpolation errors as the value of RSME is directly affected by the amount of rainfall. The sites with small interpolation errors are mainly distributed in the northern part of the basin since the rainfall in the north is very small or even zero. Comparing the distribution of gauges errors using the five methods, methods that combined with F-SVD have more yellow points in the northern part and less blue points in the southern part of the basin than OK and IDW, indicating that the use of F-SVD will improve the accuracy no matter the rainfall is large or small.

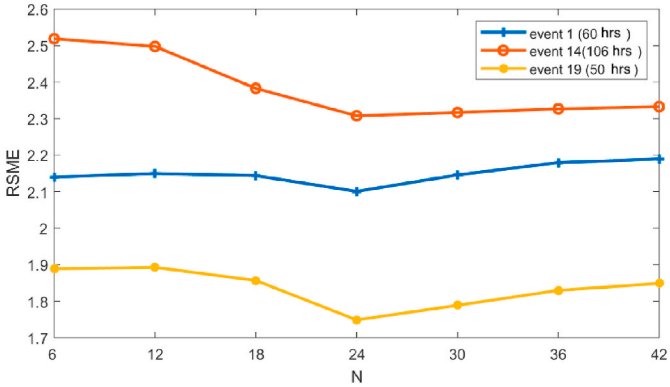


Fig. 6. RSME of 3 long-duration rainfall events under different values of *N*.

Table 3
Models performances on all rainfall events.

Methods	MAE (mm)	RSME (mm)	KS(−)	PERC (−)
OK	2.190	3.751	0.360	0.936
IDW	2.163	3.669	0.257	0.911
F-SVD-OK	1.263	2.357	0.204	0.839
F-SVD-IDW	1.220	2.294	0.180	0.825
F-SVD	1.123	2.141	0.178	0.795

The areal rainfall characters of 20 events calculated by the interpolation results are shown in Table 4, among them the calculation results that are closest to the actual rainfall records are marked with a green background, and the farthest ones are marked with orange. It can be seen that the average of areal rainfall ranges from 0.5 mm to 3 mm. The standard deviation reflects the fluctuation of rainfall at different times. No. 5 event has an average areal rainfall of 1.22 mm, which is of medium-scale, but the standard deviation is the lowest, indicating that the rainfall process is relatively smooth with few cases of sudden increase and decrease. The average areal rainfall of the No.9 event is 1.86 mm and it only lasts for 16 h, but the standard deviation reaches 2.01, indicating that rainfall mainly concentrated in several periods. The characteristic values of the events calculated by the five interpolation methods are not much different from the measured values. Among the 20 selected rainfalls, for both rainfall characteristic indicators, at least 16 of the rainfall characteristic values calculated by F-SVD are the closest to the actual measurement, and that calculated by OK are the farthest. Besides, through combination with F-SVD, both IDW and OK perform better and generate more accurate rainfall characteristics than before.

4.3. Models evaluation on representative rainfall events

To better evaluate the interpolation ability of the F-SVD model, the magnitude of the event is sorted, and three typical events of heavy, moderate and small rain are selected for comparative analysis. They are in turn the No. 7 small rain event (7 May-8 May 2015), No. 20 moderate rain event (17 Jun-19 Jun 2018) and No. 18 heavy rain event (23 Sep-27 Sep 2017). The results are shown in Fig. 8. In the figure, (a) is the cumulative observed rainfall at each site, (b), (c), and (d) are the MAE of F-SVD, IDW, and OK, respectively, and (e) and (f) are the difference in accuracy between SVD and the other two methods, where the blue dots indicate the accuracy of F-SVD is higher, and the red dot indicates that of F-SVD is lower. It can be seen that for rainfall events of different magnitudes, the interpolation error of gauges with heavy rainfall is large. Besides, the blue dots in figures (e) and (f) are more than the red points, indicating the number of gauges, whose interpolation accuracy using F-SVD is higher than using IDW and OK, is more. But for the gauges with small cumulative rainfall (<5 mm), the accuracy is not improved. The difference in accuracy between F-SVD and the other two methods is small during small rain, but as the rainfall magnitude becomes larger, the difference gradually increases. In the heavy rain event, the historical rainfall at most gauges is not zero, thus the F-SVD method can effectively extract the spatial and temporal feature information from the historical rainfall for interpolation, resulting in a noticeable improvement in accuracy.

The results of three representative rainfall events using all five interpolation methods evaluated by different indicators are listed in Table 5. It can be seen that the accuracy of IDW is higher than that of OK, and through combination with F-SVD, the accuracy of IDW and OK are greatly improved. The accuracy of F-SVD is highest, which shows good estimation ability than traditional interpolators. As the magnitude of rainfall increases, the difference in accuracy between the five methods becomes larger. The interpolation error of No. 7 small rain event is the lowest judging from MAE, RSME and PERC. Besides, the evaluation result of No.20 moderate rain event using all indicators is worser than No. 18 heavy rain event. As can be seen from Table 1, N0.20 event lasts for 106 h and No.18 event only lasts for 44 h. For N0.20 event, the rainfall is very scattered and is extremely low in many moments, and the trend and regularity are not obvious, contributing to a larger

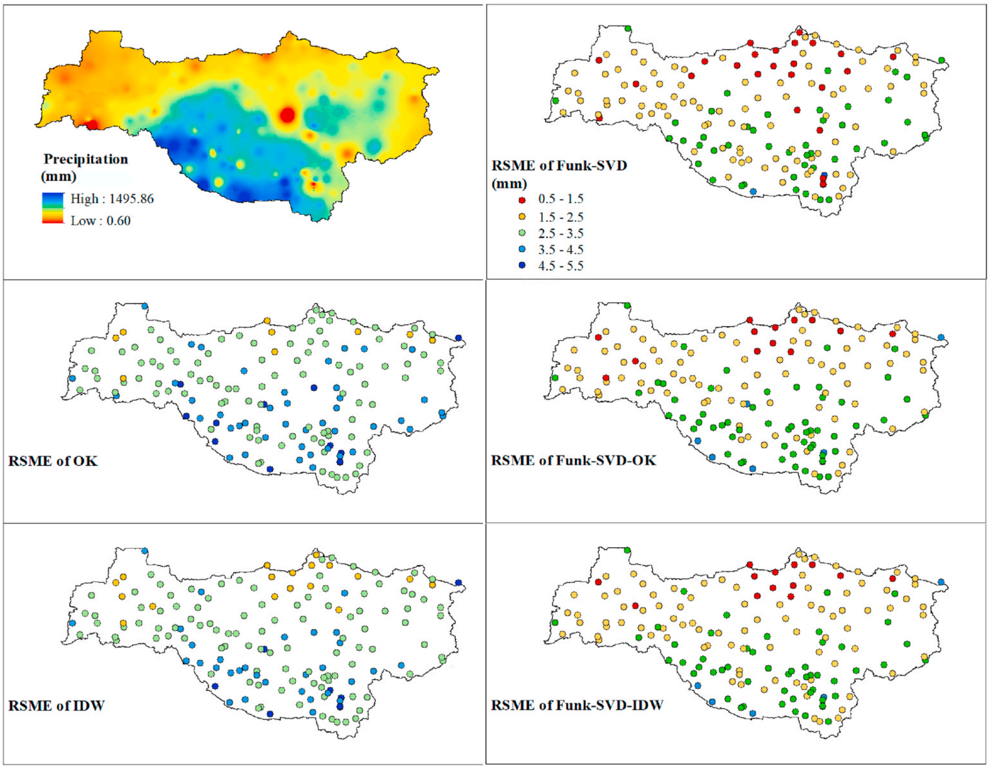


Fig. 7. Spatial distribution of cumulative rainfall events of 20 selected rainfalls over Hanjiang basin and average value of RSME using 5 interpolation methods.

Table 4
Information of basin rainfall characteristics of 20 selected events from the records and interpolation results.

Event	Average (mm)						Standard deviation (mm)					
No.	record	OK	IDW	F-SVD-OK	F-SVD-IDW	F-SVD	record	OK	IDW	F-SVD-OK	F-SVD-IDW	F-SVD
1	0.68	0.55	0.62	0.65	0.74	0.68	1.08	0.75	0.79	1.00	1.12	1.07
2	2.77	2.32	2.45	2.62	2.80	2.83	2.65	2.20	2.12	2.23	2.53	2.70
3	1.80	1.62	1.68	1.73	1.84	1.81	2.08	1.84	1.85	2.07	2.07	2.09
4	1.81	1.59	1.61	1.76	1.86	1.84	2.16	1.80	1.86	2.20	2.19	2.22
5	1.22	1.13	1.13	1.19	1.24	1.21	0.87	0.73	0.77	0.85	0.83	0.86
6	1.06	0.85	0.89	1.02	1.02	1.04	1.55	1.08	1.05	1.46	1.47	1.57
7	1.57	1.40	1.44	1.53	1.54	1.52	1.44	1.18	1.21	1.40	1.40	1.42
8	1.45	1.33	1.32	1.34	1.43	1.42	1.04	0.95	0.94	0.97	1.01	1.06
9	1.86	1.17	1.49	1.64	2.04	1.94	2.01	1.20	1.34	2.29	2.22	2.10
10	1.71	1.60	1.57	1.63	1.67	1.67	0.96	0.85	0.82	0.82	0.93	0.94
11	1.22	1.05	1.13	1.10	1.21	1.22	1.38	1.18	1.17	1.22	1.36	1.38
12	1.22	1.09	1.10	1.16	1.28	1.22	1.15	0.94	1.03	0.96	1.22	1.16
13	2.05	1.77	1.84	1.98	2.14	2.09	1.30	1.06	1.14	1.12	1.32	1.35
14	1.02	0.93	0.97	1.00	1.03	1.02	1.24	1.03	1.12	1.11	1.23	1.26
15	1.49	1.28	1.38	1.36	1.47	1.49	1.30	1.12	1.09	1.11	1.24	1.31
16	1.78	1.68	1.73	1.76	1.83	1.80	1.57	1.45	1.49	1.47	1.58	1.59
17	1.55	1.32	1.37	1.48	1.61	1.57	2.02	1.64	1.73	1.91	2.10	2.10
18	1.87	1.75	1.75	1.84	1.92	1.89	1.61	1.49	1.48	1.54	1.67	1.65
19	1.04	0.86	0.96	1.03	1.04	1.04	1.87	1.43	1.44	1.68	1.88	1.87
20	2.85	2.58	2.71	2.76	2.93	2.86	2.16	1.97	2.09	2.08	2.24	2.22

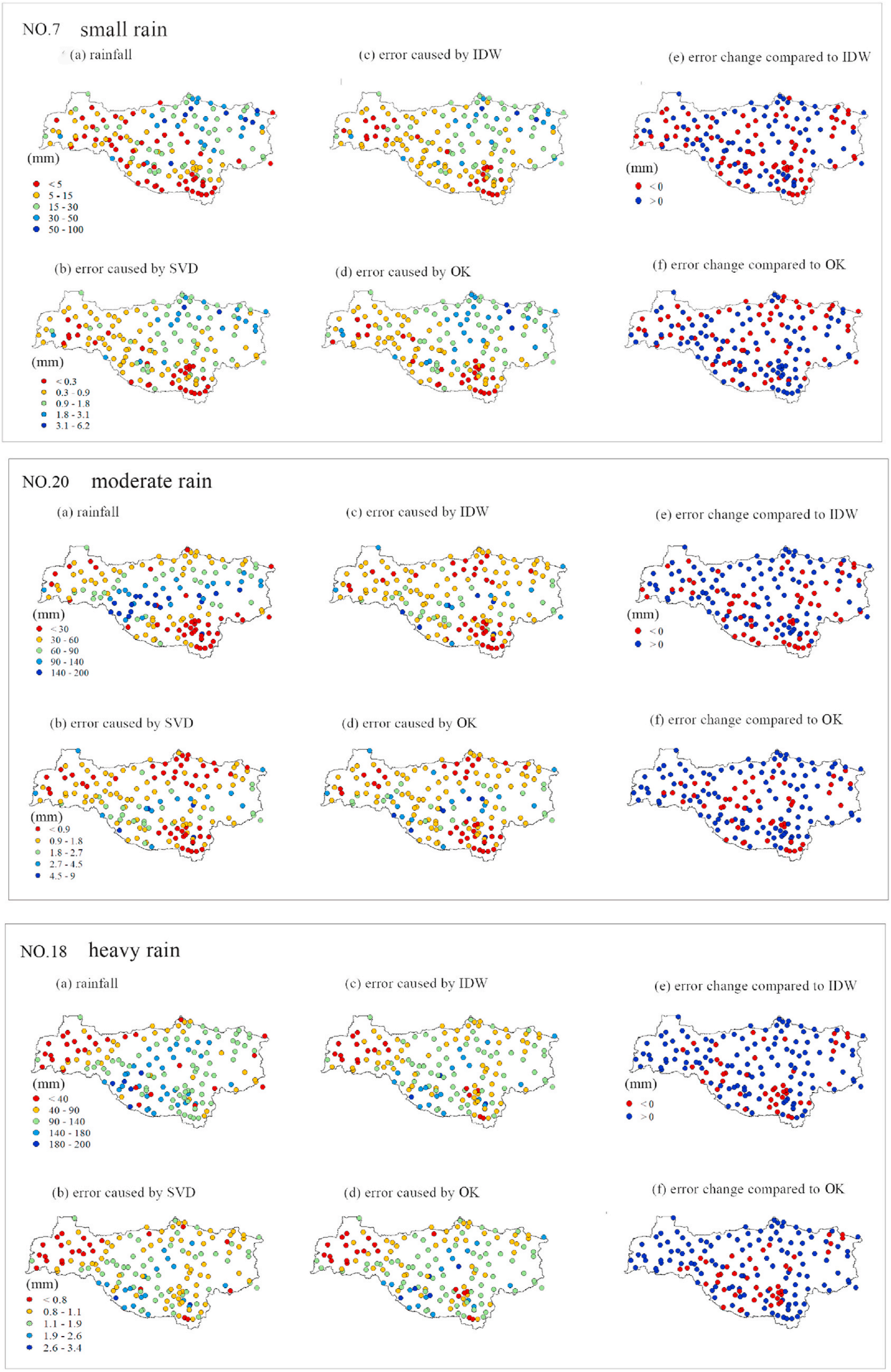


Fig. 8. For three representative rainfall events. (a) Recorded rainfall. (b) Error from interpolation using SVD. (c) Error from interpolation using IDW. (d) Error from interpolation using OK. (e) The change in error using result of IDW minus that of SVD. The blue dots represent an improvement, and the red dots represent a deterioration. (f) The change in error using result of OK minus that of SVD.

Table 5
Prediction Error for three representative rainfall events.

Event NO.	Scale	Methods	MAE (mm)	RSME (mm)	KS (−)	PERC (−)
7	small	OK	2.406	3.303	0.373	1.369
		IDW	2.334	3.098	0.363	1.315
		F-SVD-OK	1.077	1.704	0.279	0.865
		F-SVD-IDW	1.038	1.641	0.234	0.852
		F-SVD	0.965	1.507	0.238	0.851
20	moderate	OK	3.082	4.440	0.432	1.529
		IDW	3.107	4.469	0.372	1.599
		F-SVD-OK	1.699	2.615	0.324	0.928
		F-SVD-IDW	1.624	2.546	0.275	0.927
		F-SVD	1.406	2.227	0.226	0.926
18	heavy	OK	2.701	4.339	0.304	1.424
		IDW	2.665	4.266	0.159	1.414
		F-SVD-OK	1.355	2.365	0.173	1.016
		F-SVD-IDW	1.310	2.320	0.136	1.074
		F-SVD	1.188	2.076	0.133	0.976

interpolation error.

4.4. *Estimation error distribution of interpolation methods*

The average gauge error in different selected events using five interpolation methods is shown in Fig. 9. It is found that except for few events, most of them produce the largest error using OK and the smallest error using F-SVD. Through combination with F-SVD, the accuracy of both IDW and OK is improved. The evaluation results by RSME and MAE are similar since they are both greatly affected by the rainfall amount. The results of PERC and KS are not the same; some rainfall events (such as No. 11) have low RSME, MAE, and high KS and PERC. As the areal total rainfall in the No. 11 event is only 15.9 mm, the accumulated rainfall at each gauge is relatively low, so the RSME and MAE values are not too large. The calculation of PERC concerns more with the relative error than the absolute error, for gauges with little rainfall, although the interpolated rainfall is also very low, the relative error may be large, causing PERC to become large.

To evaluate the uncertainty of the five methods, two indicators, MAE and RSME, are selected, and a box plot of the errors of all stations in 20 events is shown in Fig. 10. It can be seen from the figure that the variation trends of MAE and RSME are consistent. The median values of OK and IDW are higher than other methods, and the confidence intervals of them are wider. OK has a rather large uncertainty as its confidence interval varies in different events. For most rainfall events, the median value and the width of the confidence interval of F-SVD-OK and F-SVD-IDW are very close, and sometimes the median value of F-SVD-IDW is slightly lower than that of F-SVD-OK. Through combination, the width of the confidence interval of them is greatly shorter than the previous two methods, and the median is also lower. The error distribution of F-SVD is the best as it directly uses observed records for estimation. It shows that the F-SVD method not only improves the interpolation accuracy but also has higher stability. Among the 20 rainfall events interpolated by F-SVD, the relatively large errors mainly occur in No. 2, No. 4, No. 9, and No. 13 events and their durations are 30, 32, 16 and 38 h, respectively, thus the F-SVD method may perform less well in short duration rainfall events.

5. Conclusions

In this paper, a new method based on F-SVD to improve rainfall estimation is proposed. Unlike traditional interpolators that only focus on the spatial relationships of gauges, the proposed method incorporates rainfall information at the current and historical moments into the estimation process which results in a more accurate result. Through combination with traditional interpolators, it can be applied to interpolate rainfall at unknown points and obtain continuous spatial distribution of rainfall. Thus it is a practical method to process rainfall data for spatial pattern analysis and prepare input data for distributed hydrological models. Twenty rainfall events are selected from the hourly rainfall data of the rainfall gauges in the Hanjiang basin to verify this method by cross-validation using four indicators, and IDW and Kriging are included as benchmarks for accuracy comparison. The study concludes that:

(1) According to the interpolation results of 20 rainfalls, F-SVD has the highest accuracy and OK has the lowest accuracy. Through

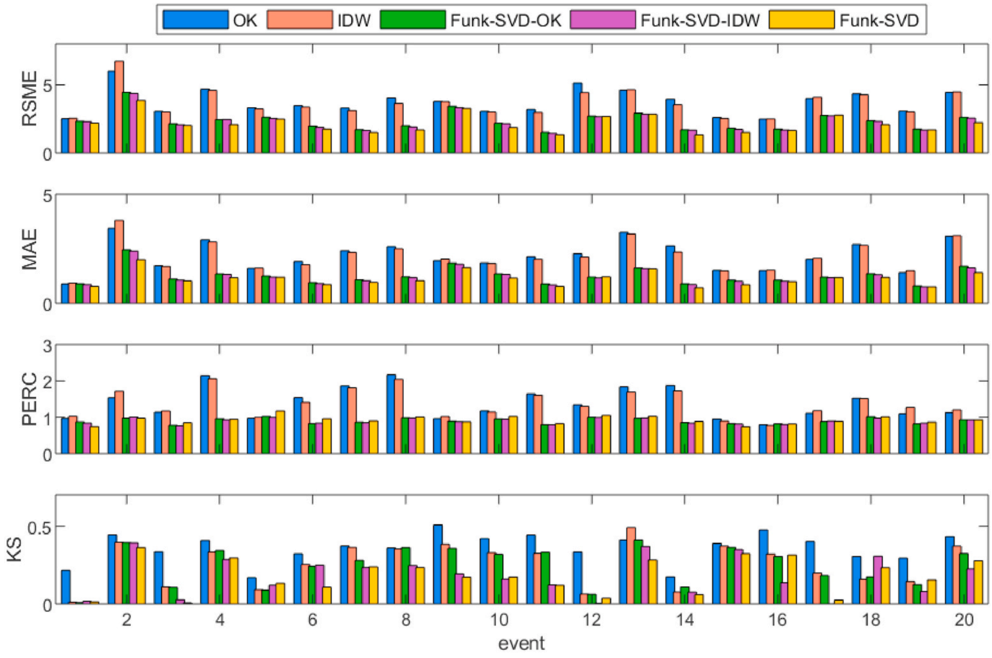


Fig. 9. Interpolation error of different methods for all selected events.

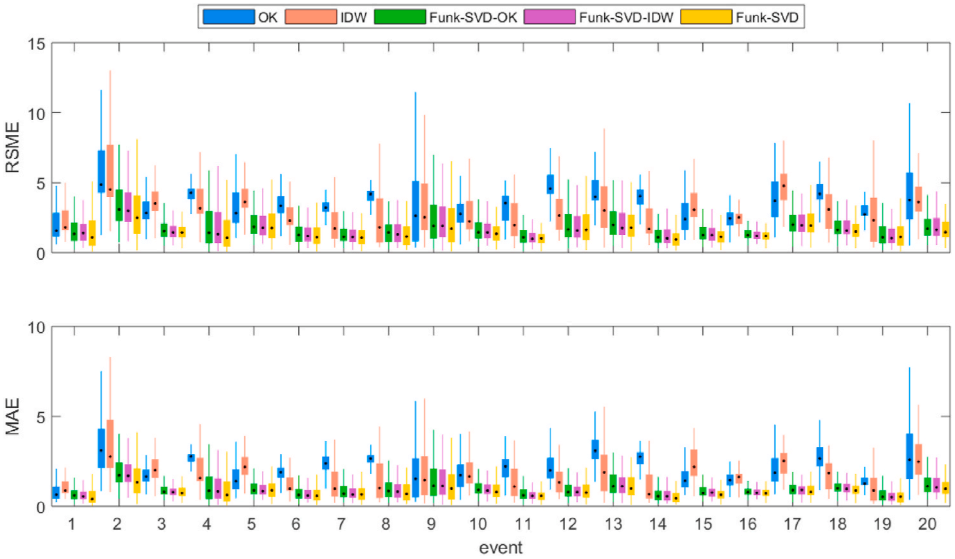


Fig. 10. Boxplot of interpolation error of all gauges in different event using SVD method.

combination of OK and IDW with F-SVD, the accuracy of OK and IDW can be greatly improved. Due to the different emphasis of the indicators, the gap between five methods in RSME and MAE is larger than that in PERC and KS.

(2) In cross-validation, the rainfall magnitude has an effect on the interpolation accuracy. For different rainfall events, the larger the rainfall magnitude, the more gauges of which the accuracy can be improved by the F-SVD method.

(3) According to the error distribution of all stations in each rainfall event, F-SVD not only improves the interpolation accuracy but also reduces the uncertainty of the error, so F-SVD has better stability.

However, there are also some limitations in this study. F-SVD is a latent factor model and its algorithmic meaning is to build relationships between time and space through latent factors, which cannot correspond to physical concepts in reality, thus it has a poor interpretability. Besides, only two widely used interpolation methods (OK and IDW) are used for comparison, and one basin, Hanjiang basin, is considered in this study, the conclusions may not be generalized. Therefore, more basins with different distribution of gauges and more methods for comparison will be helpful to validate the proposed interpolation method.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study is financially supported by the National Key Research and Development Program (2018YFC0407904, 2017YFA0603702).

References

Ahrens, B., 2006. Distance in spatial interpolation of daily rain gauge data. Hydrol. Earth Syst. Sci. 10 (2), 197–208.

Bargaoui, Z., Chebbi, A., 2009. Comparison of two kriging interpolation methods applied to spatiotemporal rainfall. J. Hydrol. 365 (1), 56–73.

Bell, R.M., Koren, Y., 2007. In: Ramakrishnan, N., Zaiane, O.R., Shi, Y., Clifton, C.W., Wu, X.D. (Eds.), Icdm 2007: Proceedings of the Seventh Ieee International Conference on Data Mining, pp. 43–52. Ieee Computer Soc, Los Alamitos.

Cai, X., Wang, X., Jain, P., Flannigan, M.D., 2018. Evaluation of gridded precipitation data and interpolation methods for forest fire danger rating in alberta, Canada. J. Geophys. Res. 124 (1), 3–17.

Carrera-Hernández, J.J., Gaskin, S.J., 2007. Spatio temporal analysis of daily precipitation and temperature in the Basin of Mexico. J. Hydrol. 336 (3), 231–249.

Chen, H., Guo, S., Xu, C.-y., Singh, V.P., 2007. Historical temporal trends of hydro-climatic variables and runoff response to climate variability and their relevance in water resource management in the Hanjiang basin. J. Hydrol. 344 (3), 171–184.

Cliff, A., Ord, J.K., 1975. Space-time modelling with an application to regional forecasting. Trans. Inst. Br. Geogr. 119–128.

Dai, Q., Zhu, X., Zhuo, L., Han, D., Liu, Z., Zhang, S., 2020. A hazard-human coupled model (HazardCM) to assess city dynamic exposure to rainfall-triggered natural hazards. Environ. Model. Software 127, 104684.

Dalezios, N.R., Adamowski, K., 1995. Spatio-temporal precipitation modelling in rural watersheds. Hydrol. Sci. J. 40 (5), 553–568.

Delhomme, J.P., 1978. Kriging in the hydrosciences. Adv. Water Resour. 1 (5), 251–266.

Diez-Sierra, J., del Jesus, M., 2017. A rainfall analysis and forecasting tool. Environ. Model. Software 97, 243–258.

Engmann, S., Cousineau, D., 2011. Comparing distributions: the two-sample Anderson-Darling test as an alternative to the Kolmogorov-Smirnoff test. Journal of applied quantitative methods 6 (3), 1–17.

Foehn, A., Hernandez, J.G., Schaefli, B., De Cesare, G., 2018. Spatial interpolation of precipitation from multiple rain gauge networks and weather radar data for operational applications in Alpine catchments. J. Hydrol. 563, 1092–1110.

Funk, S., 2006. Netflix update: try this at home (December 2006). URL. http://sifter.org/~simon/journal/20061211.html.

Garcia, M., Peterslidard, C.D., Goodrich, D.C., 2008. Spatial interpolation of precipitation in a dense gauge network for monsoon storm events in the southwestern United States. Water Resour. Res. 44 (5).

González-Macías, C., Sánchez-Reyna, G., Salazar-Coria, L., Schifter, I., 2014. Application of the positive matrix factorization approach to identify heavy metal sources in sediments. A case study on the Mexican Pacific Coast. Environ. Monit. Assess. 186 (1), 307–324.

Goovaerts, P., 2000. Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. J. Hydrol. 228 (1), 113–129.

Hadi, S.J., Tombul, M., 2018. Comparison of spatial interpolation methods of precipitation and temperature using multiple integration periods. Journal of the Indian Society of Remote Sensing 46 (7), 1187–1199.

Hu, Y.F., Koren, Y., Volinsky, C., 2008. Proceedings. In: Gunopulos, D., Turini, F., Zaniolo, C., Ramakrishnan, N., Wu, X.D. (Eds.), Icdm 2008: Eighth Ieee International Conference on Data Mining, p. 263. Ieee Computer Soc, Los Alamitos.

Koren, Y., Bell, R., Volinsky, C., 2009. Matrix factorization techniques for recommender systems. Computer 42 (8), 30–37.

Kumari, M., Basistha, A., Bakimchandra, O., Singh, C.K., 2016. In: Raju, N.J. (Ed.), Comparison of Spatial Interpolation Methods for Mapping Rainfall in Indian Himalayas of Uttarakhand Region. Springer International Publishing, Cham, pp. 159–168.

Lee, C.M., Mudaliar, M., Haggart, D.R., Wolf, C.R., Miele, G., Vass, J.K., Higham, D.J., Crowther, D.J.P.O., 2012. Simultaneous non-negative matrix factorization for multiple large scale gene expression datasets in toxicology, 7, 12.

Morris, F., Toucher, M.L.W., Clulow, A.D., Kusangaya, S., Morris, C., Bulcock, H., 2016. Improving the understanding of rainfall distribution and characterisation in the Cathedral Peak catchments using a geo-statistical technique. WaterSA 42 (4), 684–693.

Pfeifer, P.E., Deutrch, S.J., 1980. A three-stage iterative procedure for space-time modeling phillip. Technometrics 22 (1), 35–47.

Plouffe, C.C.F., Robertson, C., Chandrapala, L., 2015. Comparing interpolation techniques for monthly rainfall mapping using multiple evaluation criteria and auxiliary data sources: a case study of Sri Lanka. Environ. Model. Software 67, 57–71.

Ryu, S., Song, J.J., Kim, Y., Jung, S.-H., Do, Y., Lee, G., 2021. Spatial interpolation of gauge measured rainfall using compressed sensing. Asia-Pacific Journal of Atmospheric Sciences 57 (2), 331–345.

Shepard, D., 1968. A Two-Dimensional Interpolation Function for Irregularly-Spaced Data. Association for Computing Machinery, pp. 517–524.

Sivakumar, B., Woldemeskel, F.M., 2015. A network-based analysis of spatial rainfall connections. Environ. Model. Software 69, 55–62.

Spadavecchia, L., Williams, M.J.A., 2009. Can spatio-temporal geostatistical methods improve high resolution regionalisation of meteorological variables? Agric. For. Meteorol. 149 (6), 1105–1117.

Takacs, G., Pilaszy, I., Nemeth, B., Tikk, D., 2009. Scalable collaborative filtering approaches for large recommender systems. J. Mach. Learn. Res. 10, 623–656.

Thiessen, A.H., 1911. Precipitation averages for large areas. Mon. Weather Rev. 39 (7), 1082–1089.

Tobler, W.R., 1970. A computer movie simulating urban growth in the detroit region. Econ. Geogr. 46, 234–240.

Wasko, C., Sharma, A., Rasmussen, P.F., 2013. Improved spatial prediction: a combinatorial approach. Water Resour. Res. 49 (7), 3927–3935.

Xie, Y., Berkowitz, C.M., 2006. The use of positive matrix factorization with conditional probability functions in air quality studies: an application to hydrocarbon emissions in Houston, Texas. Atmos. Environ. 40 (17), 3070–3091.

Xu, M., Yang, Y., Han, M., Qiu, T., Lin, H., 2019. Spatio-temporal interpolated echo state network for meteorological series prediction. IEEE Trans. Neural Network. 30 (6), 1621–1634.

Xue, J.-l., Zhi, Y.-y., Yang, L.-p., Shi, J.-c., Zeng, L.-z, Wu, L.-s., 2014. Positive matrix factorization as source apportionment of soil lead and cadmium around a battery plant (Changxing County, China). Environ. Sci. Pollut. Control Ser. 21 (12), 7698–7707.

Yang, X., Xie, X., Liu, D.L., Ji, F., Wang, L., 2015. Spatial interpolation of daily rainfall data for local climate impact assessment over greater sydney region. Advances in Meteorology 2015, 1–12.

Yeh, C.-H., Lin, C.-Y., Muchtar, K., Liu, P.-H., 2018. Rain streak removal based on non-negative matrix factorization. Multimed. Tool. Appl. 77 (15), 20001–20020.

Yu, X., Fu, Y., Xu, L.W., Liu, G.Z., 2018. A Cross-Domain Recommendation Algorithm for D2D Multimedia Application Systems, pp. 62574–62583. Ieee Access 6.

Zhang, K., Shang, X., Herrmann, H., Meng, F., Mo, Z., Chen, J., Lv, W., 2019. Approaches for identifying PM2.5 source types and source areas at a remote background site of South China in spring. Sci. Total Environ. 691, 1320–1327.

Zhang, M., Leon, C.d., Migliaccio, K., 2018. Evaluation and comparison of interpolated gauge rainfall data and gridded rainfall data in Florida. USA. Hydrological Sciences Journal 63 (4), 561–582.