Research Article

A Modified Spatiotemporal Mixed-Effects Model for Interpolating Missing Values in Spatiotemporal Observation Data Series

Qiang Shi,¹ Wujiao Dai¹ ,¹ Rock Santerre,² and Ning Liu¹

¹School of Geosciences and Info-Physics, Central South University, Changsha 410083, Hunan, China

²Center for Research in Geomatics, Laval University, Quebec G1V 0A6, Canada

Correspondence should be addressed to Wujiao Dai; wjdai@csu.edu.cn

Received 6 March 2020; Revised 8 June 2020; Accepted 15 June 2020; Published 10 August 2020

Academic Editor: Mijia Yang

Copyright © 2020 Qiang Shi et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Missing values in data series is a common problem in many research and applications. Most of existing interpolation methods are based on spatial or temporal interpolation, without considering the spatiotemporal correlation of observation data, resulting in poor interpolation effect. In this paper, a Modified Spatiotemporal Mixed-Effects (MSTME) model for interpolation of spatiotemporal data series is proposed. Experiments with simulated data and real SCIGN data are performed to assess the validity of the proposed model in comparison with Kriged Kalman Filter (KKF) model and Spatiotemporal Mixed-Effects (STME) model. The average improvements of simulated data and SCIGN data for observed stations are around 46% and 19% over the KKF model and 62% and 21% over the STME model, and those for unobserved stations are around 23% and 34% over the KKF model and 41% and 16% over the STME model, respectively, indicating that the proposed MSTME model can achieve better accuracy for interpolating missing values.

1. Introduction

Missing values is a problem encountered repeatedly in data acquisition and data analysis. The situation may be the result of data transmitting interruption, instrumentation failure, inaccurate observations, insufficient sampling, and other unintended observational problems. The existence of missing values will compromise the integrity of data series, and thus limits the use of the data in various applications and research works. For example, discontinuities pose a serious obstacle to time series analysis, which usually requires continuous data as the basic condition of time series analysis [1]. As said by Barzi [2], when there is too much data missing, the data loses its usefulness. Therefore, it is necessary to interpolate missing values to generate complete data needed for practical applications.

The interpolation of missing values has been studied widely and various methods have been explored. At present, there are mainly three types of interpolation methods: temporal interpolation, spatial interpolation, and

spatiotemporal interpolation. The first type is the temporal interpolation method including conventional interpolation methods, singular spectrum analysis (SSA)-based method, and empirical orthogonal function (EOF)-based method. The conventional interpolation methods, such as linear interpolation [3], cubic spline interpolation [4], and polynomial interpolation [5], are effective for random missing data, while those are poor for the continuous missing data. As discussed by Schoellhamer [6], singular spectrum analysis is a powerful time series analysis technology. It can extract as much reliable information as possible from observation data and construct a model based on this information, which can be used to recover missing data without any prior knowledge of the observation data [7]. Kondrashov and Ghil [8] first proposed to use the SSA model for analyzing uneven sampling data, and the SSA model was tested using sea surface temperature measurements. Wang et al. [9] used the SSA method to reconstruct a reliable model from unevenly sampled time series with missing data.

However, the interpolation results are too smooth and this method is less effective for the time series with “red” temporal spectra where noisy modes contribute significantly. The EOF method uses reconstructed time series to interpolate missing data. Lorenz [10] introduced it into the field of atmospheric science, and it has been widely used since then [11–13]. The EOF-based method can obtain better interpolation results for missing data with colored noise. However, the EOF-based method as well as conventional interpolation methods and SSA method only considers the temporal correlation of single station, without considering the spatial correlation between stations. The second type is the spatial interpolation method including inverse distance weight (IDW) and Kriging. The IDW method interpolates the missing data according to the principle that the closer the distance in space, the stronger the correlation will be [14]. The Kriging method uses the variogram structure of the sample data to interpolate the missing data [15]. The main disadvantage of Kriging interpolation and IDW interpolation is that they depend on the distribution of stations. The distribution density of stations is too sparse and the interpolation effect is poor. The temporal interpolation or spatial interpolation only considers the temporal or spatial correlation, without considering the spatiotemporal correlation of data, which affects the interpolation accuracy. The third type is the spatiotemporal interpolation method including Kriged Kalman Filter (KKF), Space Time Kalman Filter (STKF), and Spatiotemporal Mixed-Effects (STME). KKF is a spatiotemporal Kalman interpolation algorithm combined with Kriging interpolation and Kalman filter [16–18]. It mainly uses Kriging interpolation to construct spatial field to describe spatial correlation between stations and uses the Kalman filter to describe the temporal correlation of the data. For the large computation problem, the KKF model is suitable for the interpolation of spatial sparse stations. Wike [19] proposed the STKF model which has the dimension reduction function and a temporally dynamic and spatially descriptive statistical model. Cressie et al. [20–22] proposed a fixed rank spatiotemporal Kalman, STRE, or STME model to calculate massive spatiotemporal data. The STME model has better dimensionality reduction than the STKF model, and the selection of spatial basis functions of the STME model is easier than that of the STKF model. Therefore, the STME model is widely used in environmental pollution [23], climate [24], and deformation analysis [25, 26]. Both the STME model and STKF model reduce the dimension of spatiotemporal data by uniformly distributed spatial basis functions, while the positions and bandwidth of spatial bases need to be repeatedly adjusted to obtain better results. As mentioned earlier, the KKF model considers spatiotemporal correlation of stations and has high interpolation accuracy for sparse stations. However, the KKF model used in Liu et al. [27] relies too much on the initial values of the parameters and needs to adjust the initial parameters repeatedly to obtain the best accuracy, which limits the usefulness and reliability of the model.

In order to improve the interpolation accuracy of missing data for sparse stations by establishing a spatiotemporal interpolation model which can acquire the initial

parameters of the model adaptively with less manual intervention, we use the optimization algorithm combined with EM estimation to improve the existing spatiotemporal interpolation model and propose a high-precision adaptive spatiotemporal interpolation model called modified spatiotemporal mixed-effects (MSTME) model.

2. Methodology

2.1. Kriged Kalman Filter Model. The KKF model is a particular type of General State Space model for spatiotemporal data, and the model is as follows [16]:

$$\begin{aligned} Z_t(s) &= \mathbf{H}(s)' \boldsymbol{\alpha}_t + \varepsilon_t(s), \\ \boldsymbol{\alpha}_t &= \Phi \boldsymbol{\alpha}_{t-1} + \mathbf{w}_t, \end{aligned} \quad (1)$$

where s and t represent the position of observation station and observation time, respectively, $Z_t(s)$ is the spatiotemporal observation, $\mathbf{H}(s)$ is the spatial field, $\boldsymbol{\alpha}_t$ is the state vector, $\varepsilon_t(s)$ is the observed noise with variance R , Φ is the state transition matrix, and $\mathbf{w}_t$ is the state innovation vector, and it is Gaussian random noise.

2.2. Spatiotemporal Mixed-Effects Model. The spatiotemporal mixed-effects model is a hierarchical model for massive spatiotemporal data, and the model is as follows [28]:

$$\begin{aligned} Z_t(s) &= \mu_t(s) + \mathbf{S}_t(s)' \boldsymbol{\alpha}_t + \xi_t(s) + \varepsilon_t(s), \\ \boldsymbol{\alpha}_t &= \phi \boldsymbol{\alpha}_{t-1} + \eta_t, \end{aligned} \quad (2)$$

where s and t represent the position of observation station and observation time, respectively, $Z_t(s)$ is the spatiotemporal observation, $\mu_t(s) = X_t(s)\beta_t$ is the global spatiotemporal trends, which can be removed by polynomial fitting, $X_t(s)$ is a known vector of covariates, β_t is a vector of fixed but unknown trend coefficients, $\xi_t(s)$ is the fine-scale spatial variation, $\varepsilon_t(s)$ is the observed noise with variance R , $\mathbf{S}_t(s)' \boldsymbol{\alpha}_t$ means local spatial variation, $\mathbf{S}_t(s) = [S_{1,t}(s), S_{2,t}(s), \dots, S_{r,t}(s)]'$ is the chosen r rank basis functions, $\boldsymbol{\alpha}_t$ is the state vector, Φ is the state transition matrix, and η_t is the state innovation vector, and it is Gaussian random noise.

The STME model is mainly used for spatiotemporal massive data, and it is not suitable for spatiotemporal data of sparse stations. Therefore, we use the idea of Fassò and Finazzi [29] to modify the STME model to interpolate missing values for sparse stations by spatiotemporal filtering.

2.3. Modified Spatiotemporal Mixed-Effects Model. As stated previously, $\mathbf{S}_t(s)$ in equation (2) is the chosen r rank basis functions, representing the spatial correlation and making dimension reduction for massive spatiotemporal data. In this paper, we make $\mathbf{S}_t(s)'$ as the n -dimensional unit matrix, and n is the number of observation stations. $\xi_t(s)$ in equation (2) is the fine-scale spatial variation caused by spatial dimension reduction, which is needed to be estimated by observations of massive stations. For the observations of sparse stations, we do not make dimension reduction by setting $\mathbf{S}_t(s)' = I$; therefore, the fine-scale spatial variation

can be ignored, and the modified spatiotemporal mixed-effects model is as follows:

$$\begin{aligned} Z_t(s) &= \mu_t(s) + \mathbf{V}_t(s)' \boldsymbol{\alpha}_t + \varepsilon_t(s), \\ \boldsymbol{\alpha}_t &= \phi \boldsymbol{\alpha}_{t-1} + \gamma_t, \end{aligned} \quad (3)$$

where $s, t, Z_t(s), \mu_t(s)$, and $\varepsilon_t(s)$ are the same as those in equation (2). The $\mathbf{V}_t(s)' = I$ is the basis functions. Since the spatial correlation cannot be described by $\mathbf{V}_t(s)'$, we modified the state vector $\boldsymbol{\alpha}_t$ to describe the spatiotemporal correlation. The STME model uses the basic functions $\mathbf{S}_t(s)$ to describe the spatial correlation, and the innovation vector, η_t in equation (2), is set to be spatially uncorrelated, which is estimated by the EM. However, the state vector $\boldsymbol{\alpha}_t$ in equation (2) only describes the temporal evolution of the data, and its spatial correlation is not described. Therefore, we set γ_t in equation (3) to be spatial correlation with matrix covariance function given by Γ rather than Gaussian random noise:

$$\Gamma(h, \theta) = \rho(h, \theta), \quad (4)$$

where $h = \|s - s'\|$ is the distance between two observation stations, $\rho(h, \theta)$ is a valid spatial correlation function, such as exponential function, spherical function, and Gaussian function, with range parameter θ . In particular, the widely used spherical spatial correlation function is used in both simulation experiments and real data experiments, that is, $\rho(h, \theta) = \text{sph}(h/\theta)$. Therefore, the parameter set Ψ of the MSTME model to be estimated is then given by $\{\beta, \phi, \theta, R\}$.

The KKF, STME, and MSTME models are all spatial extensions of Kalman, and they are all models that take into account the spatiotemporal correlation. Among them, KKF model and MSTME model are suitable for spatiotemporal modeling of sparse stations, and the STME model is suitable for spatiotemporal modeling of massive data. There are two main differences among the three models. In the description of spatial correlation, the KKF model describes the spatial correlation by spatial field, STME model describes the spatial correlation by spatial basis function, and MSTME model describes the spatial correlation by spatial correlation function. In the aspect of model implementation, the spatial field of the KKF model and the spatial basis function of the STME model need to be estimated empirically and then adjusted repeatedly, while the MSTME model can estimate the parameter of the spatial correlation function adaptively.

2.4. Missing Data Interpolation Method for MSTME Model.

In order to interpolate the missing data, we need to firstly estimate the values of the parameter set Ψ using the observed data, and the MSTME observation equation can be written as follows:

$$\begin{bmatrix} Z_1(s, t) \\ Z_2(s, t) \end{bmatrix} = \begin{bmatrix} \mu_t^1(s) \\ \mu_t^2(s) \end{bmatrix} + \begin{bmatrix} \mathbf{V}_t^1(s)' \\ \mathbf{V}_t^2(s)' \end{bmatrix} \boldsymbol{\alpha}_t + \begin{bmatrix} \varepsilon_1(s, t) \\ \varepsilon_2(s, t) \end{bmatrix}, \quad (5)$$

where $Z_1(s, t), \mu_t^1(s), \mathbf{V}_t^1(s)'$, and $\varepsilon_1(s, t)$ are the observed data, the spatiotemporal trend, the basic functions, and the noise of the observed stations without missing values, respectively. $Z_2(s, t), \mu_t^2(s), \mathbf{V}_t^2(s)'$, and $\varepsilon_2(s, t)$ are the

observed data, the spatiotemporal trend, the basic functions, and the noise of the stations with missing values, respectively. The variance-covariance matrix of the noise is expressed by

$$\text{Var} \begin{pmatrix} \varepsilon_1(s, t) \\ \varepsilon_2(s, t) \end{pmatrix} = \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix}. \quad (6)$$

When we estimate the values of the parameter set Ψ , we set the values of $Z_2(s, t), \mu_t^2(s)$, and $\mathbf{V}_t^2(s)'$ to zero matrix to ensure that the missing data $Z_2(s, t)$ are not involved in the estimation of the parameter set Ψ . The ordinary least-squares method is used to estimate the parameter β by choosing a proper spatial polynomial function $X_t(s)$. The parameters ϕ and R are the same as those in equation (2), and they are estimated by the expectation maximization (EM) algorithm. The θ is a new parameter of the MSTME model, and it is same as the parameter of equation (5) in Fassò and Finazzi [29]. We use the numerical optimization to estimate θ , and the numerical optimization is as follows:

$$\{\theta\} = \arg \min T \log |\Sigma_r| + \text{tr} [\Sigma_r^{-1} (C - B\tilde{\theta}' - \tilde{\theta}B' + \tilde{\theta}A\tilde{\theta}')] , \quad (7)$$

where $\Sigma_r = \Gamma(h, \theta) = \rho(h, \theta)$ is the variance-covariance of γ_t , $\tilde{\theta}$ is the estimated value of θ obtained by EM algorithm. $A = \sum_{t=1}^T \alpha_{t-1}^T (\alpha_{t-1}^T)' + P_{t-1}^T$, $B = \sum_{t=1}^T \alpha_t^T (\alpha_{t-1}^T)' + P_{t,t-1}^T$, and $C = \sum_{t=1}^T \alpha_t^T (\alpha_t^T)' + P_t^T$, where α_t^T, P_t^T , and $P_{t,t-1}^T$ are the same as those in the STME model [28], denoting the output of the Kalman smoother.

Once the parameter set Ψ are estimated, the missing data $Z_2(s, t)$ can be interpolated by

$$Z_2(s, t) = X_t(s)\beta_t + \mathbf{V}_t^2(s)' \boldsymbol{\alpha}_t. \quad (8)$$

3. Simulation Experiment

The validity of the MSTME model to interpolate missing data for observed stations and unobserved stations under different time gaps and different station densities (139 stations and 209 stations) is tested in the simulation experiment by simulating the spatiotemporal data series of sparse stations.

3.1. Simulated Data. In order to validate the proposed MSTME model, the simulation study is implemented. We first simulate a series of data fields and then randomly select 139 stations and 209 stations from the data fields as observation stations to obtain the observations. The detailed simulation steps are as follows:

(1) Construct a series of true data fields.

- (a) We construct the spatial variations of the first day by taking Cholesky decomposition of a spatial covariance matrix K (e.g., $K = LL'$) with $v_1 = L\alpha_1$, where K obeying the spherical distribution with the sill value of 1 mm^2 and the range value of 250 km in the domain $D = \{(x, y) | 0 \text{ km} < x < 500 \text{ km}\}$, and $\alpha_1 \sim N(0 \text{ mm}, 1 \text{ mm}^2)$.

(b) In the random walk model, let $\alpha_t = \alpha_{t-1} + \eta_t, \eta_t \sim N(0 \text{ mm}, 9 \text{ mm}^2)$, with $t = 2, 3, \dots, 700$. The true data fields can be generated by $v_t = L\alpha_t$.

(2) Choose the real values of sparse stations.

We select 139 and 209 sparse stations from the constructed true data fields and take the reserve values in the temporal domain as the true values of the selected stations. The constructed true data fields and the selected stations are shown in Figure 1.

(3) Add noise.

The Gaussian white noise with a standard deviation of 5 mm is added to the simulated true value to simulate the actual observation values.

(4) Delete some data.

Nine of the 139 stations and 209 stations are randomly selected to delete part of the data to simulate missing data, and the selected 9 stations are shown in Figure 1. The missing gaps of the selected 9 stations are shown in Table 1 to simulate the missing situations of different time gaps, and the observations of the selected 9 stations are shown in Figure 2.

3.2. Simulated Results. The results of the MSTME model to interpolate the missing data of observed stations and unobserved stations are analyzed in Sections 3.2.1 and 3.2.2, respectively.

3.2.1. Interpolating Missing Data for Observed Stations.

The simulated data of 139 stations and 209 stations are used to establish the interpolation model, and the simulated missing data of the selected 9 stations are interpolated by the established MSTME model. In order to test the validity of the proposed MSTME model, we also use the KKF model and STME model to interpolate the missing data and compare the interpolation accuracy of the MSTME model with the KKF model and STME model. For the STME model, the spatial bases are set from different scales until the spatial residuals $V = (Z_t - \mu_t) - S((S'S)^{-1}S'(Z_t - \mu_t))$ fitted by least square are reliable, where Z_t is the observation, μ_t is the global spatiotemporal trend, S is the chosen r rank spatial basis functions, and $(*)'$ is transposition operator. 125 bisquare spatial bases for the STME model from two scales with 25 from the first scale and 100 from the second scale are selected, and the positions of spatial bases are uniformly distributed in the study area. The interpolation results of the MSTME model, KKF model, and STME model for the missing data are shown in Figure 3.

As shown in Figure 3, we can see that the interpolated results of the MSTME model and KKF model for missing data fit well with the original data, while the STME model has better interpolated results for station nos. 1, 2, 3, 4, and 5 and worse interpolation results for station nos. 6, 7, 8, and 9. In order to quantitatively analyze the interpolation accuracy of the MSTME model and KKF model, the statistic RMS values of residuals of interpolated results for missing data at the 9 selected stations are calculated in Table 2. Comparing

Figure 1 and Table 2, we can find that the interpolated results of the MSTME model are more consistent with the true series of the missing data, and the interpolated results of the MSTME model for the missing data are better than those of the KKF model and STME model. The average improvements for the KKF model and STME model are 43.4% and 58.6%, respectively.

It can be seen from the left panel of Figure 1 that there are more observation stations around the stations (nos. 1, 2, 3, 4, and 5), and there are fewer observation stations around the other stations (nos. 6, 7, 8, and 9). In theory, the space-related information of the surrounding stations that can be utilized for nos. 6, 7, 8, and 9 is less, and the interpolated results for nos. 6, 7, 8, and 9 will be poor than those for nos. 1, 2, 3, 4, and 5. As shown in Table 2, the interpolated results of the MSTME model, KKF model, and STME model for nos. 1, 2, 3, 4, and 5 are better than those for nos. 6, 7, 8, and 9, which is consistent with the theoretical analysis.

In order to compare the effect of station density on the interpolation of missing data, 209 stations are used to establish the MSTME model, KKF model, and STME model. The interpolation results are shown in Figure 4, and the statistic RMS values of residuals of interpolated results for missing data are calculated in Table 3.

As shown in Figure 4 and Table 3, the interpolation accuracy of the MSTME model for 209 stations is better than that for 139 stations, and the average improvements of the MSTME model are 48.3% and 64.8%, respectively, compared with the KKF model and STME model for 209 stations. Comparing the left panel and right panel in Figure 1, we can see that there are more stations around the selected 9 stations in right panel than those in the left panel. As mentioned earlier, the more observation stations around the missing data stations, the higher the interpolation accuracy. Therefore, the interpolation results of 209 stations once again verify the previous theoretical analysis and also prove the validity of the MSTME model. However, the interpolation results of nos. 1, 2, 3, and 4 for the KKF model and nos. 1, 2, 4, and 8 for the STME model are worse than those for 139 stations, which do not accord with the previous theoretical analysis.

From the analysis of the simulation data and the implementation of three models, there are two reasons for this phenomenon. The first reason is we can see from Figure 1 that the data field simulated in this paper varies greatly. Although there are more stations around the selected nine stations of the 209 stations than 139 stations, there are great differences between the data variation of some surrounding stations and those of the selected 9 stations. The second reason is the spatial field describing spatial correlation in the KKF model and the positions and bandwidth of spatial bases in the STME model cannot be adaptively estimated according to the data variation, which leads to the inaccuracy of both models in describing spatial correlation of data. For the data of 209 stations, the KKF model and STME model utilize not only the newly added station information with spatiotemporal correlation around the selected 9 stations but also the data of new stations without spatiotemporal correlation around the selected 9

FIGURE 1: The constructed true data fields and the locations of selected stations. (a) The locations of the selected 139 stations and (b) the locations of the selected 209 stations. The black points are the selected stations, and the selected 9 stations with missing data are marked with the red asterisk “*”.

TABLE 1: The data missing gaps of the selected 9 stations (unit: day).

Station No	1	2	3	4	5	6	7	8	9
Missing gaps	51	81	101	131	151	181	201	201	231

FIGURE 2: Observations of the selected 9 stations. The gray lines are the simulated observations with missing data. The red lines are the simulated true observations without noise. The “gap” in the bottom of each image is the data missing gaps.

FIGURE 3: The interpolation results of the MSTME model, KKF, model and STME model for 139 stations. The blue lines are the smooth results and the interpolation results of the MSTME model. The green lines are the smooth results and the interpolation results of the KKF model. The cyan lines are the smooth results and the interpolation results of the STME model. The red lines are the simulated true values without noise. The “gap” in the bottom of each image is the data missing gaps.

TABLE 2: The RMS values of residuals of interpolated results for missing data with 139 stations (unit: mm).

Station no.	MSTME	KKF	STME	Improvement 1 (%)	Improvement 2 (%)
1	12.0	14.6	14.8	17.8	18.9
2	8.8	17.6	21.1	50.0	58.3
3	8.7	20.7	14.1	58.0	38.3
4	15.4	25.6	19.9	39.8	22.6
5	7.0	13.6	32.7	48.5	78.6
6	15.9	40.9	76.8	61.1	79.3
7	17.3	21.2	33.9	18.4	49.0
8	18.1	37.5	33.2	51.7	45.5
9	20.1	26.4	51.3	23.9	60.8
Average	13.7	24.2	33.1	43.4	58.6

The “Improvement 1” is the improvement rate of the MSTME model over KKF model, and the “Improvement 2” is the improvement rate of the MSTME model over STME model.

stations to interpolate the missing data. On the contrary, the MSTME model can adaptively estimate the best values of the unknown parameters according to the characteristics of the data itself and can more accurately describe the spatio-temporal correlation of the data field.

3.2.2. Interpolating Missing Data for Unobserved Stations. The validity of the MSTME model on missing data for observed stations are analyzed in Section 3.2.1, and we will

analyze the validity of the MSTME model on missing data for 9 unobserved stations in this section. We delete all the data of the selected 9 stations and use the remaining 130 stations and 200 stations to interpolate the deleted data of the selected 9 stations using the MSTME model, KKF model, and STME model. The interpolation results of the selected 9 stations with 130 stations and 200 stations are shown in Figures 5 and 6, respectively, and the statistic RMS values of residuals of interpolated results are calculated in Tables 4 and 5.

FIGURE 4: The interpolation results of the MSTME model, KKF model, and STME model with 209 stations. The blue lines are the smooth results and the interpolation results of the MSTME model. The green lines are the smooth results and the interpolation results of the KKF model. The cyan lines are the smooth results and the interpolation results of the STME model. The red lines are the simulated true values without noise. The “gap” in the bottom of each image is the data missing gaps.

TABLE 3: The RMS values of residuals of interpolated results for missing data with 209 stations (unit: mm).

Station no.	MSTME	KKF	STME	Improvement 1 (%)	Improvement 2 (%)
1	11.4	17.5	20.7	34.9	44.9
2	8.2	19.1	34.5	57.1	76.2
3	8.6	24.4	16.1	64.8	46.6
4	14.9	27.3	34.9	45.4	57.3
5	6.9	13.1	28.3	47.3	75.6
6	14.6	31.9	44.8	54.2	67.4
7	13.1	17.3	17.7	24.3	26.0
8	17.4	37.4	95.3	53.5	81.7
9	15.5	26.0	21.4	40.4	27.6
Average	12.3	23.8	34.9	48.3	64.8

The “Improvement 1” is the improvement rate of the MSTME model over KKF model, and the “Improvement 2” is the improvement rate of the MSTME model over STME model.

As shown in Figures 5 and 6 and Tables 4 and 5, the interpolated results of the MSTME model, KKF model, and STME model for unobserved stations fit well with the original data. The MSTME model has better interpolated accuracy than the KKF model and STME model, and the average improvement rates are 15.4% and 48.0% for 130 stations and 30.7% and 34.2% for 200 stations, respectively. Comparing Tables 4 and 5, we can see that the interpolation

results of unobserved stations for the MSTME model with 200 stations are better than those with 130 stations, among which the interpolation results of nos. 3, 7, and 9 have great improvement rates, while those of nos. 1, 2, 4, 5, 6, and 8 with 200 stations are basically the same as those with 130 stations. As Figure 5 and Table 4 show, the interpolation results of the STME model for station no. 7 in Figure 5 have large error. The reasons maybe, it can be seen from the left panel of

FIGURE 5: The interpolation results of the deleted 9 stations with 130 stations. The blue lines are the interpolation results of the MSTME model. The green lines are the interpolation results of the KKF model. The cyan lines are the interpolation results of the STME model. The red lines are the simulated true values without noise.

Figure 7 that there are no stations in the same variation area as station no. 7 and the values of its spatial basis functions can only be estimated by stations in other variation areas. Among them, the data variations of some stations (such as stations under station no. 7) are similar to station no. 7, and those of other stations (such as stations on the right and upper side of station no. 7) are quite different from station no. 7. However, the spatial bases describing spatial correlation not only incorporates the data of known stations related to station no. 7 into the STME model but also incorporates the data unrelated to station no. 7 into the STME model. Therefore, the interpolation results of the STME model for station no. 7 are bad.

As mentioned earlier, the more observation stations around the missing data stations, the higher the interpolation accuracy for the MSTME model. As shown in Figure 1, there are more stations around the unobserved stations (nos. 3, 7, and 9) in the right panel than those in the left panel, so that there are more spatiotemporal correlation data available for the MSTME model. However, the interpolation results of nos. 1 and 6 for the KKF model and nos. 2, 4, and 6 for the STME model with 200 stations are worse than those with 130 stations, which do not accord with the previous theoretical analysis, and the reasons are the same as those analyzed in Section 3.2.1.

In summary, the simulation experiment shows that the MSTME model can consider the spatiotemporal correlation between observation stations. It has better accuracy for interpolating the missing data of observed stations and unobserved stations and is better than the KKF model and STME model.

4. Real Experiment for GPS Deformation Data

Here, we shall use the real GPS deformation data to demonstrate the effectiveness of the proposed MSTME model for missing data and compare with the KKF model used in Liu et al. [27] and the STME model.

4.1. SCIGN Data. In order to compare the effectiveness of the proposed MSTME model with the KKF model and STME model, we use the same real experiment data used in Liu et al. [27]. The GPS daily coordinates of 209 sites from January 1, 2005, to January 1, 2007, for a total of 731 days, in the Southern California Integrated GPS Network, are downloaded from the SCIGN website (<http://www.scign.org/>), and the selected GPS sites are shown in Figure 8.

FIGURE 6: The interpolation results of the deleted 9 stations with 200 stations. The blue lines are the interpolation results of the MSTME model. The green lines are the interpolation results of the KKF model. The cyan lines are the interpolation results of the STME model. The red lines are the simulated true values without noise.

TABLE 4: The RMS values of residuals of interpolated results for the deleted 9 stations with 130 stations (unit: mm).

Station no.	MSTME	KKF	STME	Improvement 1 (%)	Improvement 2 (%)
1	12.4	15.7	13.9	21.0	10.8
2	35.9	40.4	37.6	11.1	4.5
3	27.7	29.6	51.1	6.4	45.8
4	20.0	21.1	43.9	5.2	54.4
5	18.4	29.1	30.0	36.8	38.7
6	34.1	39.1	36.3	12.8	6.1
7	57.3	59.1	225.9	3.0	74.6
8	22.8	26.9	36.0	15.2	36.7
9	43.1	60.6	48.2	28.9	10.6
Average	30.2	35.7	58.1	15.4	48.0

The “Improvement 1” is the improvement rate of the MSTME model over KKF model, and the “Improvement 2” is the improvement rate of the MSTME model over STME model.

4.2. Data Processing. Before the experiment, the GPS coordinate series are normalized to the deformation relative to the first day. In order to prevent the data noise of the first day from affecting the data of other days, the mean values of the first 10 days data are taken as the denoising values of the first day data and then made the deformation relative to these mean values. Furthermore, in order to demonstrate the

effectiveness of the proposed MSTME model for missing data, we select 4 GPS stations with less data missing and delete part of the data to simulate large missing GPS data, and the selected 4 GPS stations (7ODM, BKMS, BSRY, and HOGS) are marked in red in Figure 8. Finally, the MSTME model, the KKF MATLAB software by Liu et al. [27], and STME model are used to interpolate these artificial gaps.

TABLE 5: The RMS values of residuals of interpolated results for the deleted 9 stations with 200 stations (unit: mm).

Station no.	MSTME	KKF	STME	Improvement 1 (%)	Improvement 2 (%)
1	12.7	26.5	13.4	52.1	5.2
2	35.7	40.0	39.4	10.8	9.4
3	10.7	19.3	28.7	44.6	62.7
4	19.8	20.8	52.0	4.8	61.9
5	17.7	21.3	22.1	16.9	19.9
6	33.1	47.7	46.0	30.6	28.0
7	12.8	34.0	14.2	62.4	9.9
8	19.7	23.7	24.7	16.9	20.2
9	14.5	21.8	27.5	33.5	47.3
Average	19.6	28.3	29.8	30.7	34.2

The “Improvement 1” is the improvement rate of the MSTME model over KKF model, and the “Improvement 2” is the improvement rate of the MSTME model over STME model.

FIGURE 7: The distribution of the spatial basis for the STME model. The black hollow circles represent the bisquare spatial bases of the first scale, and the black “+” represents the bisquare spatial bases of the second scale.

FIGURE 8: The GPS stations of SCIGN network.

4.3. Results and Analysis

4.3.1. Interpolating Missing Data for Observed Stations. Since the KKF model is only performed on the U (Up) direction of SCIGN data in Liu et al. [27], in order to compare the effectiveness of the proposed method with Liu’s work, we also only perform the MSTME model, KKF model, and STME model on the U direction of the SCIGN data in this paper. For the STME model, the method to select spatial basis is similar to that in Section 3.2.1, and 89 bisquare spatial bases from two scales with 25 from the first scale and 64 from the second scale are selected. The positions of spatial basis are uniformly distributed in the study area. The interpolation results of the selected 4 GPS stations with the missing gaps for 100 days are shown in Figure 9.

FIGURE 9: The interpolation results of the MSTME model, KKF model, and STME model for the artificial gaps of 100 days with 209 stations. The black lines are the original GPS data. The blue lines are the smooth results and the interpolation results of the MSTME model. The green lines are the smooth results and the interpolation results of the KKF model. The cyan lines are the smooth results and the interpolation results of the STME model. The pink dotted lines are the artificial gaps.

TABLE 6: The RMS values of residuals of the interpolated results for the artificial gaps of 100 days with 209 stations (unit: mm).

Site	MSTME	KKF	STME	Improvement 1 (%)	Improvement 2 (%)
7ODM-U	8.3	10.4	9.1	20.2	8.8
BKMS-U	2.1	2.5	2.9	16.0	27.6
BSRY-U	2.3	2.9	3.8	20.7	39.5
HOGS-U	2.0	2.6	2.7	23.1	25.9
Average	3.7	4.6	4.6	19.6	19.6

The "Improvement 1" is the improvement rate of the MSTME model over KKF model, and the "Improvement 2" is the improvement rate of the MSTME model over STME model.

As shown in Figure 9, the interpolated results of the MSTME model for missing GPS data fit well with the original data, while the KKF model and STME model have relatively poor smooth and interpolation results. The MSTME models can better capture the deformation characteristics of the deformation field and reflect the actual deformation for the missing data. Similar to the simulated experiment, we can see from Figure 8 that there are more GPS stations around the HOGS station, and the spatio-temporal interpolated results for HOGS will be better than other stations in theory. The experimental results of the MSTME model, KKF model, and STME model are consistent with the theoretical analysis that the interpolated

results of HOGS-U data are better than those of other stations, which can be seen from the RMS values of residuals of the interpolated results for missing data in Table 6.

Table 6 also shows that the interpolation accuracy of the MSTME model is slightly better than those of the KKF model and STME model for the missing data with the missing gaps of 100 days, and the improvements of the MSTME model compared with the KKF model and STME model in the SCIGN experiment is not as obvious as that in the simulation experiment.

To test the validity of the MSTME model, we perform another experiment with a longer continuous missing gaps. We set the data missing gaps of the selected 4 GPS

FIGURE 10: The interpolation results of the MSTME model, KKF model, and STME model for the artificial gaps of 200 days with 209 stations. The black lines are the original GPS data. The blue lines are the smooth results and the interpolation results of the MSTME model. The green lines are the smooth results and the interpolation results of the KKF model. The cyan lines are the smooth results and the interpolation results of the STME model. The pink dotted lines are the artificial gaps.

TABLE 7: The RMS values of residuals of the interpolated results for the artificial gaps of 200 days with 209 stations (unit: mm).

Site	MSTME	KKF	STME	Improvement 1 (%)	Improvement 2 (%)
7ODM-U	10.0	10.5	10.9	4.8	8.3
BKMS-U	2.2	4.2	3.9	47.6	43.6
BSRY-U	2.0	2.7	3.9	25.9	48.7
HOGS-U	2.6	3.3	2.9	21.2	10.3
Average	4.2	5.2	5.4	19.2	22.2

The “Improvement 1” is the improvement rate of the MSTME model over KKF model, and the “Improvement 2” is the improvement rate of the MSTME model over STME model.

stations to 200 days and delete the data of these 4 GPS stations for 200 consecutive days. The interpolation results of the selected 4 GPS stations with the missing gaps of 200 days are shown in Figure 10, and the RMS values of residuals of the interpolated results for missing data are shown in Table 7.

As Figure 10 shows, the interpolated results of the MSTME model for missing GPS data fit well with the original data, and the quantitative analysis of the residuals of the interpolated results for missing data in Table 7 also proves that the interpolation accuracy of the MSTME model is still better than those of the KKF model and STME model for the missing data with the missing gaps of 200 days.

4.3.2. Interpolating Missing Data for Unobserved Stations.

Similar to the simulation experiment, we delete all the data of the selected 4 stations (7ODM, BKMS, BSRY, and HOGS) and use the remaining 205 stations to interpolate the deleted data of the selected 4 stations using the MSTME model, KKF model, and STME model, respectively. The results are shown in Figure 11, and the statistic RMS values are calculated in Table 8.

We can see from Figure 11 and Table 8 that the interpolation results of the MSTME model are better than those of the KKF model and STME model, and the average improvements are 33.8% and 16.4%, respectively. As shown in Figure 11, the interpolation results of the KKF model for

FIGURE 11: The interpolation results of the MSTME model, KKF model, and STME model for the deleted stations. The black lines are the original GPS data. The blue lines are the interpolation results of the MSTME model. The green lines are the interpolation results of the KKF model. The cyan lines are the interpolation results of the STME model.

TABLE 8: The RMS values of residuals of interpolated results for the deleted 4 stations (unit: mm).

Site	MSTME	KKF	STME	Improvement 1 (%)	Improvement 2 (%)
7ODM-U	10.0	11.9	12.0	16.0	16.7
BKMS-U	4.4	11.6	5.5	62.1	20.0
BSRY-U	2.6	3.1	3.3	16.1	21.2
HOGS-U	3.2	4.1	3.6	22.0	11.1
Average	5.1	7.7	6.1	33.8	16.4

The “Improvement 1” is the improvement rate of the MSTME model over KKF model, and the “Improvement 2” is the improvement rate of the MSTME model over STME model.

BKMS stations have great errors. The reason is that the spatial field describing spatial correlation in the KKF model cannot be adaptively estimated according to the data variation, which leads to the inaccuracy of the KKF models in describing spatial correlation of data and results in the inaccurate interpolation results of BKMS station.

The real experimental results once again prove that the MSTME model has high accuracy for interpolating the missing data of observed stations and unobserved stations, which is better and stable than the KKF model and STME model.

5. Conclusion

This paper proposes a MSTME model for interpolating the missing values. The model uses EM algorithm and numerical optimization algorithm to obtain unknown parameters, which reduces manual intervention and the dependence of model accuracy on the initial values. Both simulation experiment and SCIGN experiment show that the proposed MSTME model is effective. The proposed MSTME model can not only interpolate the missing data of observed stations but also interpolate the data of unobserved stations

using the spatiotemporal correlation data of the stations around the unobserved stations. It can be used for interpolation of various spatiotemporal data, such as atmospheric data, hydrological data, and deformation data.

In this paper, the results of experiments with simulated data and SCIGN data show that the interpolation accuracy of the MSTME model is higher than that of the KKF model and STME model, and the stability of the MSTME model is stronger. The main reasons could be that (1) the spatial field of the KKF model is based on the variogram function, and the selection of the variogram in Liu et al. [27] needs to repeatedly adjust the initial value of unknown parameters to obtain the optimal value. Different parameter values will lead to different interpolation results. (2) The STME model use the basic functions to describe the spatial correlation, and the positions and bandwidth of spatial bases in the STME model need to be determined by experience, which cannot be adaptively estimated according to the data variation. The empirical selection of the variogram used to describe the spatial field for the KKF model, as well as the positions and bandwidth of spatial bases describing the spatial correlation for the STME model, not only requires a lot of processing time to repeatedly adjust the model parameters but also leads to the inaccuracy of the KKF model and STME model in describing spatial correlation. On the contrary, the MSTME model proposed in this paper uses EM algorithm and numerical optimization algorithm to adaptively obtain the optimal parameters based on the characteristics of data itself, which does not need to adjust the initial values of parameters repeatedly.

Data Availability

The “.mat” data used to support the findings of this study are included within the article. The experimental data in this paper includes three files: “simulate_data_139.mat,” “simulate_data_209.mat,” and “scign.mat.” The file format is matlab’s .mat file, where “simulate_data_139.mat” is the simulated data for 139 stations, “simulate_data_209.mat” is the simulated data for 209 stations, and “scign.mat” is the SCIGN data. All experimental data in this article are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant no. 41674011), Fundamental Research Funds for the Central Universities of Central South University (Grant no. 2018zzts033), and Innovation Foundation for Postgraduate of Hunan Province, China (Grant no. CX2018B104).

References

- [1] E. B. Brooks, V. A. Thomas, R. H. Wynne et al., “Fitting the multitemporal curve: a fourier series approach to the missing data problem in remote sensing analysis,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 9, pp. 3340–3353, 2012.
- [2] F. Barzi, “Imputations of missing values in practice: results from imputations of serum cholesterol in 28 cohort studies,” *American Journal of Epidemiology*, vol. 160, no. 1, pp. 34–45, 2004.
- [3] X. Dong, W. S. Lu, and A. C. K. Soong, “Linear interpolation in pilot symbol assisted channel estimation for OFDM,” *Wireless Communications IEEE Transactions on*, vol. 6, no. 5, pp. 1910–1920, 1910.
- [4] Y. Feng and Y. Xiang, “Mitigation of spectral mis-registration effects in imaging spectrometers via cubic spline interpolation,” *Optics Express*, vol. 16, no. 20, pp. 15366–15374, 2008.
- [5] A. Sinap and W. V. Assche, “Polynomial interpolation and Gaussian quadrature for matrix valued functions,” *Linear Algebra & Its Applications*, vol. 207, no. 94, pp. 71–114, 2012.
- [6] D. H. Schoellhamer, “Singular spectrum analysis for time series with missing data,” *Geophysical Research Letters*, vol. 28, no. 16, pp. 1499–1512, 2001.
- [7] F. J. Alonso, J. M. D. Castillo, and P. Pintado, “Application of singular spectrum analysis to the smoothing of raw kinematic signals,” *Journal of Biomechanics*, vol. 38, no. 5, pp. 1085–1092, 2005.
- [8] D. Kondrashov and M. Ghil, “Spatio-temporal filling of missing points in geophysical data sets,” *Nonlinear Processes in Geophysics*, vol. 13, no. 2, pp. 151–159, 2006.
- [9] X. Wang, Y. Cheng, S. Wu, and K. Zhang, “An effective toolkit for the interpolation and gross error detection of GPS time series,” *Survey Review*, vol. 48, no. 348, pp. 202–211, 2016.
- [10] E. N. Lorenz, “The essence of chaos,” *Physics Today*, vol. 48, no. 2, 1993.
- [11] A. Baldacci, G. Corsini, R. Grasso et al., “A study of the Alboran sea mesoscale system by means of empirical orthogonal function decomposition of satellite data,” *Journal of Marine Systems*, vol. 29, no. 1–4, pp. 293–311, 2001.
- [12] M. S. Pritchard and R. C. J. Somerville, “Empirical orthogonal function analysis of the diurnal cycle of precipitation in a multi-scale climate model,” *Geophysical Research Letters*, vol. 36, no. 36, pp. 126–127, 2009.
- [13] R. M. Shore, M. P. Freeman, and J. W. Gjerloev, “An empirical orthogonal function reanalysis of the northern polar external and induced magnetic field during solar cycle 23,” *Journal of Geophysical Research: Space Physics*, vol. 123, no. 1, pp. 781–795, 2018.
- [14] S. M. Indirawati, S. Pandia, H. Mawengkang, and W. Hasan, “Inverse distance weighted method and environmental health risks of plumbum pollution in drinking water in belawan coastal area,” *Advanced Science Letters*, vol. 23, no. 4, pp. 3339–3343, 2017.
- [15] P. Tziachris, E. Metaxa, F. Papadopoulos, and M. Papadopoulou, “Spatial modelling and prediction assessment of soil iron using kriging interpolation with pH as auxiliary information,” *ISPRS International Journal of Geo-Information*, vol. 6, no. 9, p. 283, 2017.
- [16] K. V. Mardia, C. Goodall, E. J. Redfern, and F. J. Alonso, “The kriged kalman filter,” *Test*, vol. 7, no. 2, pp. 217–282, 1998.
- [17] J. Cortés, “Distributed kriged kalman filter for spatial estimation,” *IEEE Transactions on Automatic Control*, vol. 54, no. 12, pp. 2816–2827, 2009.
- [18] V. N. Ioannidis, D. Romero, and G. B. Giannakis, “Inference of spatio-temporal functions over graphs via multikernel kriged kalman filtering,” *IEEE Transactions on Signal Processing*, vol. 66, no. 12, pp. 3228–3239, 2018.

- [19] C. Wikle, "A dimension-reduced approach to space-time Kalman filtering," *Biometrika*, vol. 86, no. 4, pp. 815–829, 1999.
- [20] N. Cressie, T. Shi, and E. L. Kang, "Fixed rank filtering for spatio-temporal data," *Journal of Computational and Graphical Statistics*, vol. 19, no. 3, pp. 724–745, 2010.
- [21] E. L. Kang, N. Cressie, and T. Shi, "Using temporal variability to improve spatial mapping with application to satellite data," *Canadian Journal of Statistics*, vol. 38, no. 2, pp. 271–289, 2010.
- [22] H. Nguyen, M. Katzfuss, N. Cressie, and A. Braverman, "Spatio-temporal data fusion for very large remote sensing datasets," *Technometrics*, vol. 56, no. 2, pp. 174–185, 2014.
- [23] S. K. Sahu and K. V. Mardia, "A Bayesian kriged Kalman model for short-term forecasting of air pollution levels," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 54, no. 1, pp. 223–244, 2005.
- [24] M. Katzfuss and N. Cressie, "Bayesian hierarchical spatio-temporal smoothing for very large datasets," *Environmetrics*, vol. 23, no. 1, pp. 94–107, 2012.
- [25] N. Liu, W. Dai, R. Santerre, J. Hu, Q. Shi, and C. Yang, "High spatio-temporal resolution deformation time series with the fusion of InSAR and GNSS data using spatio-temporal random effect model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 1, pp. 364–380, 2019.
- [26] Q. Shi, W. Dai, R. Santerre, Z. Li, and N. Liu, "Spatially heterogeneous land surface deformation data fusion method based on an enhanced spatio-temporal random effect model," *Remote Sensing*, vol. 11, no. 9, p. 1084, 2019.
- [27] N. Liu, W. Dai, R. Santerre, and C. Kuang, "A MATLAB-based Kriged Kalman Filter software for interpolating missing data in GNSS coordinate time series," *GPS Solutions*, vol. 22, no. 1, p. 25, 2018.
- [28] M. Katzfuss and N. Cressie, "Spatio-temporal smoothing and EM estimation for massive remote-sensing data sets," *Journal of Time Series Analysis*, vol. 32, no. 4, pp. 430–446, 2011.
- [29] A. Fassò and F. Finazzi, "Maximum likelihood estimation of the dynamic coregionalization model with heterotopic data," *Environmetrics*, vol. 22, no. 6, pp. 735–748, 2011.