Hindawi
Complexity
Volume 2024, Article ID 9806264, 1 page
https://doi.org/10.1155/2024/9806264

Retraction

Retracted: Application of Empirical Orthogonal Function Interpolation to Reconstruct Hourly Fine Particulate Matter Concentration Data in Tianjin, China

Complexity

Received 23 January 2024; Accepted 23 January 2024; Published 24 January 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] H. Zhou, H. Pan, S. Li, and X. Lv, "Application of Empirical Orthogonal Function Interpolation to Reconstruct Hourly Fine Particulate Matter Concentration Data in Tianjin, China," *Complexity*, vol. 2020, Article ID 9724367, 15 pages, 2020.

Hindawi
Complexity
Volume 2020, Article ID 9724367, 15 pages
https://doi.org/10.1155/2020/9724367

Research Article

Application of Empirical Orthogonal Function Interpolation to Reconstruct Hourly Fine Particulate Matter Concentration Data in Tianjin, China

Hongwu Zhou,[1,2] Haidong Pan,[1,2] Shuang Li,[3] and Xianqing Lv[1,2]

[1] *Physical Oceanography Laboratory, Qingdao Collaborative Innovation Center of Marine Science and Technology (CIMST), Ocean University of China, Qingdao, China*
[2] *Qingdao National Laboratory for Marine Science and Technology, Qingdao, China*
[3] *Ocean College, Zhejiang University, Zhoushan, China*

Correspondence should be addressed to Shuang Li; lshuang@zju.edu.cn and Xianqing Lv; xqinglv@ouc.edu.cn

Received 28 April 2020; Revised 18 August 2020; Accepted 28 September 2020; Published 28 October 2020

Academic Editor: Zhihan Lv

Fine particulate matter with diameters less than 2.5 μm (PM2.5) concentration monitoring is closely related to public health, outdoor activities, environmental protection, and other fields. However, the incomplete PM2.5 observation records provided by ground-based PM2.5 concentration monitoring stations pose a challenge to the study of PM2.5 propagation and evolution model. Consequently, PM2.5 concentration data imputation has been widely studied. Based on empirical orthogonal function (EOF), a new spatiotemporal interpolation method, EOF interpolation (EOFI) is introduced in this paper, and then, EOFI is applied to reconstruct the hourly PM2.5 concentration records of two stations in the first half of the year. The main steps of EOFI here are to firstly decompose the spatiotemporal data matrix of the original observation site into mutually orthogonal temporal and spatial modes with EOF method. Secondly, the spatial mode of the missing data station is estimated by inverse distance weighting interpolation of the spatial mode of the observation sites. After that, the records of the missing data station can be reconstructed by multiplying the estimated spatial mode and the corresponding temporal mode. The optimal mode number for EOFI is determined by minimizing the root mean square error (RMSE) between reconstructed records and corresponding valid records. Finally, six evaluation indices (mean absolute error (MAE), RMSE, correlation coefficient (Corr), deviation rate bias, Nash–Sutcliffe efficiency (NSE), and index of agreement (IA)) are calculated. The results show that EOFI performs better than the other three interpolation methods, namely, inverse distance weight interpolation, thin plate spline, and surface spline interpolation. The EOFI has the advantages of less computation, less parameter selection, and ease of implementation, it is an alternative method when the number of observation stations is rare, and the proportion of missing value at some stations is large. Moreover, it can also be applied to other spatiotemporal variables interpolation and imputation.

1. Introduction

Fine particulate matter (PM2.5) is particulate matter with aerodynamic diameter less than 2.5 μm in ambient air [1]. Hazy weather will form if PM2.5 concentration is too high, which has adverse impacts on human health, traffic, and outdoor activities [2], and it will also produce other indirect inestimable economic losses [3]. Therefore, many countries attach great importance to the monitoring and forecasting of PM2.5 concentration. A large number of ground-based monitoring stations have been established. For example, 1500 monitoring stations have been set up in the United States. In China, around 1500 stations have been set up in 454 cities by 2018, and a new national ambient air quality standard for PM2.5 was introduced in 2012 [1, 2]. Generally, it is believed that high PM2.5 concentration has become a prominent challenge for air pollution control in China, which is mainly caused by the industrial combustion of coal and gasoline, traffic emissions, and long-distance transport [4, 5]. The North China Plain, especially the Beijing-Tianjin-Hebei region (Figure 1(a)), is one of the regions most severely affected by the hazy weather [4, 6]. To monitor air

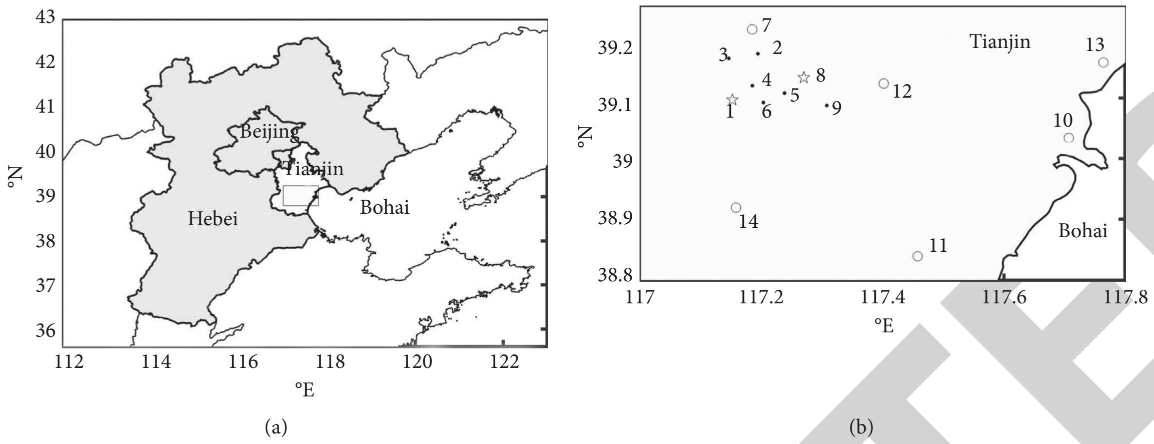

Figure 1: (a) Map of the Beijing-Tianjin-Hebei region. The rectangle in Tianjin is the study area of (b). (b) Location of 14 monitoring stations in Tianjin. From 1 to 14, they are located in the city testing center, Nankou Road, Qinjian Road, Nanjing Road, Dazhigu No. 8 Road, Qianjin Road, Beichen Science and Technology Park, Tianshan Road, Yuejin Road, Fourth Avenue, Yongming Road, Hangtian Road, Hanbei Road, and Tuanbowa. The stars represent the missing data stations (stations 1 and 8), the black dots represent the stations used for the interpolation (stations 2, 3, 4, 5, 6, and 9), and the circles represent the stations far from the missing data stations (stations 7, 10, 11, 12, 13, and 14).

pollution, many urban environmental stations have been built in this region, and many researchers have analyzed the causes and behavior of high PM2.5 concentration recently [3, 7].

There have been many studies on PM2.5 concentration data analysis methods, such as real-time data space interpolation of monitoring points, weighted regression models, and mixed models [1, 8]. The application of the preceding methods mostly depends on the complete and continuous monitoring data provided by local monitoring stations. However, problem arises when original spatiotemporal PM2.5 concentration data are incomplete, which hinders further analysis and modelling, such as aerosol-related haze control and environmental health risk assessment [9, 10].

In practice, missing values and data gaps always exist in the original spatiotemporal observation records due to various factors. For example, satellite-based remote sensing may be affected by clouds, rain, aerosols, or incomplete track coverage in atmospheric research [11, 12]; *in situ* observations from land-based stations, shipborne monitoring, offshore buyo stations, and other platforms may suffer unexpected factors such as instrumental malfunction, power supply failure, and Internet outage [10, 13]. Directly ignoring incomplete spatiotemporal observation data should be carefully considered. The reasons include that the some platforms of data acquisition are expensive and irreplaceable (e.g., ocean research vessels and buoy stations), the demanding requirements of data quality (e.g., coastal tidal gauge records), and ignoring missing values sometimes may lead to biased spatial patterns and invalid inferences [10, 13]. Thus, many temporal, spatial, and spatiotemporal data interpolation and imputation methods have been proposed to fill these gaps in records.

Simple methods commonly used to fill gaps in univariate time series include mean value substitution (or median value and mode value), polynomial interpolation (linear, piecewise polynomials, and spline interpolations), and last observation carried forward (locf), but they may result in large deviations when the time gaps are too large [14–17]. Based on a Markovian process, statistical parametric models include autoregressive (AR) models, moving average (MA) models, ARMA models, and linear weighted or exponential weighted MA. Complex machine learning techniques include gradient boosting and artificial neural networks (ANNs), which are computationally intensive [10, 18].

At present, there are also numerous spatial interpolation methods. Common simple methods include inverse distance weighting (IDW) interpolation [19], global polynomial interpolation (GPI), local polynomial interpolation (LPI) [20], surface spline (SS) interpolation [21], Cressman interpolation [22], and radial basis function (RBF). Using different basis functions, RBF includes thin plate spline (TPS), thin plate spline with tension, regularized spline, multiquadric spline, and inverse multiquadric spline. The TPS method does not need to set parameters, while other RBF needs to set parameters [23]. Some statistical-based methods (e.g., Kriging interpolation, optimal interpolation (OI), and Kalman filter) are conventional and classical methods in geoscience [12, 13, 24–27].

Numerous methods have been proposed to deal with spatiotemporal data containing missing values, and a considerable part of them are based on empirical orthogonal function (EOF) (e.g., [28–31]). Compared with other methods, EOF-based methods have the advantages of ease of implementation and less computation costs [32, 33].

EOF is based on the theory of matrix eigenvalue decomposition, and the core step of EOF is to decompose the spatiotemporal matrix into the sums of space-dependent spatial modes multiplied by corresponding time-dependent temporal modes. These EOF spatial and temporal modes can reveal data inherent characteristics or some phenomenon (e.g., ENSO) [13, 28]. EOF is usually used for spatiotemporal data analysis, but it can be also used to fill the missing data gaps.

One of the earliest applications of EOF interpolation is reconstruction of global-scale sea surface temperature (SST) [28]. Based on gridded data (1982–1993) processed by OI, EOF decomposition was performed to obtain spatial modes, and then, the temporal modes were expanded to longer time period (1950–1992) via least squares method when the data coverage was relatively poor; next, the longer time period spatiotemporal SST data were reconstructed. Their work can be considered as another form of optimal interpolation [13, 34]. In 2003, Data INterpolating Empirical Orthogonal Functions (DINEOF), an iterated EOF interpolation method, was proposed to fill the missing data gap [30]. Based on the principle of EOF, DINEOF was successfully used to reconstruct missing data and fill data gaps. Alvera-Azcárate et al. [32] reconstructed missing data of Adriatic sea surface temperature. Sirjacobs et al. [35] used DINEOF to show the reconstruction of complete space-time information for 4 years of surface chlorophyll-*a* (CHL), total suspended matter, and SST over the Southern North Sea and the English Channel. However, DINEOF may fail if the data gaps are too huge.

Similar to the principle of DINEOF, EOF interpolation (EOFI) was proposed to reconstruct spatially continuous water levels in the Columbia River Estuary using limited tide gauges along the river [36]. Their main steps are as follows: firstly, the spatial-temporal data matrix of the river existing observation stations was decomposed with EOF method. Then, Pan and Lv adopt one-dimensional linear interpolation and one-dimensional spline interpolation to estimate the missing data station's spatial modes, respectively; then, EOFI reconstruction sequence is obtained by the estimated spatial modes multiplied by corresponding temporal modes, and this reconstruction sequence was in good agreement with that of the NS_TIDE method. NS_TIDE is specially designed and applied to the analysis of river tidal water level, and river flow discharge data are needed [37].

Based on the research of Pan and Lv [36], this study attempts to extend the missing data station's EOFI spatial mode from one-dimensional spatial interpolation to two-dimensional spatial interpolation. The river upstream and downstream sites are nearly one-dimensional distributed, and there is a strong correlation between the upstream and downstream water level records (e.g., when the upstream of a river rises, the water level in the downstream generally rises). Therefore, it is reasonable to apply one-dimensional interpolation to establish the spatial mode's connection between the observation stations and the missing data station. Compared with the river water level reconstruction, the PM2.5 stations' correlation is not so strong and intuitive because the PM2.5 concentration stations are spatially distributed. To establish a connection between variables that two-dimensional distributed in space, a simple idea is using IDW, so EOFI here uses IDW to estimate the spatial modes of the missing data station. Of course, other spatial interpolation methods can also be applied to the establishment of spatial mode relationships, but we will not discuss them in this paper. We consider the simple case (IDW) to verify the usability of EOFI. To the best of our knowledge, our proposed EOFI has not been applied to PM2.5 concentration data reconstruction currently; therefore, we firstly introduce and use this method to fill the data gaps and compare the result with IDW interpolation, surface spline (SS), and TPS interpolation. The competing methods we choose here are all widely used and easy to implement [38].

Compared with widely used DINEOF- and other EOF-based methods, the novelty of our method is to deal with the case of sparsely distributed observation stations and a large proportion of missing values in some stations' records. In this case, the data of the station with too many missing values are not suitable for EOF decomposition (DINEOF fills these gaps with first guess values and then uses these data for EOF decomposition); otherwise, the accuracy of temporal and spatial modes will be affected. EOFI here only uses the observation data with a small proportion of missing values for decomposition; thus, the EOF decomposed temporal and spatial modes are more accurate and less affected. Then, spatial interpolation is applied to establish spatial modes' connection between observation stations and missing data station, and next, the reconstruction sequence with optimal mode number is determined by root mean square error (RMSE). The EOFI reconstruction sequence can be used as a reasonable first guess value of the missing data station for other methods further EOF decomposition (e.g., DINEOF). In this way, the spatial mode patterns are considered to some extent. Further comparison between DINEOF and EOFI will be explained in Discussion.

The paper is arranged as follows: Section 2.1 describes the study area and data. Then, we revisit the principle of EOF decomposition and introduce IDW, EOFI, TPS, and SS. The evaluation indices of these methods will also be mentioned in Section 2. Four methods (IDW, EOFI, TPS, and SS) are applied to reconstruct two stations' PM2.5 concentrations records, and then, the results are compared with corresponding valid observations in Section 3. EOFI inverse distance weighting power *P*, the impact of site number and data time length on the EOFI reconstruction, and comparison between DINEOF and EOFI will be discussed and analyzed in Section 4. Finally, we present the advantages and disadvantages of EOFI in Section 5.

2. Materials and Methods

2.1. Study Area and Data. There are 14 monitoring stations (Figure 1(b)) located in Tianjin. These stations are distributed in different regions of the city: some stations are located in the urban area (e.g., stations 1, 2, and 3), while other stations are near the Bohai Sea (e.g., stations 10, 11, and 13). The PM2.5 concentration data provided by these monitoring stations come from China National Environmental Monitoring Center (CNEMC). The CNEMC releases near real-time PM2.5 concentration data online, but there are no direct data download interface [10]. Bai et al. used web crawler technology to obtain many cities PM2.5 concentration data from 2014 to 2019. Here, our data sources and acquisition method are the same. In this study, some of the stations provided the hourly PM2.5 data throughout the year of 2015, except for the first 25 hours from January 1st 0:00 AM to January 2nd 0:00 AM. Thus, the total time length is

8735 hours (8760 hours in 2015). The reason for first 25 hours missing values may be web crawler technology failure, or CNEMC did not release the data for that time period. Figure 2 shows the original observation records of several stations used in this study. Among them, the first half year PM2.5 concentration data of station (sta) 1 and station (sta) 8 are reconstructed and compared with their corresponding valid records (Figure 2 (1 and 7)). There are no observed data from June 30th 23:00 PM to the end of the year (near six months) in sta 1 and sta 8. In addition, Bai et al. [10] mentioned that some monitoring stations across China have stopped releasing PM2.5 observations since the middle of 2015, and consequently, observations at these stations for the second half of 2015 are missing. This is the exact case at sta 1 and sta 8 in Tianjin. In sta 1, 10.70% of the data in the first half year are missing, and the percentage of missing data for the nearly whole year record is 55.86% (Figure 2 (1)). At sta 8, the proportions of missing data for the first half year and the nearly whole year are 9.59% and 55.31%, respectively (Figure 2 (7)). It shows that there are still nearly 400 missing values in the first half of the year for both sta 1 and sta 8.

2.2. Methods

2.2.1. EOF Decomposition. The EOF method was firstly proposed by the statistician Pearson in 1902, and meteorologist Lorenz firstly introduced the EOF method into meteorological and climatic research in 1956 [39]. We consider that there are N stations providing observation records with data length L, composing the $N \times L$ space-time matrix $\mathbf{X}$. The column $\mathbf{x}_i$ consists of N points records at time i ($i = 1, 2, \ldots$, L). The most important step of EOF is to solve the eigenvalues and eigenvectors of symmetric matrix $\mathbf{XX}^{\mathbf{T}}$; the results of this decomposition include eigenvalues λ_k and their corresponding eigenvectors $\mathbf{F}_k$ (normalized orthogonal spatial modes) [13]:

$$\mathbf{XX}^{\mathbf{T}}\mathbf{F}_k = \lambda_k \mathbf{F}_k, \quad k = 1, \ldots, N. \tag{1}$$

The column $\mathbf{F}_k$ of matrix $\mathbf{F}$ is arranged from left to right in the descending order of the corresponding eigenvalues λ_k ($k = 1, \ldots, N$), the elements of the diagonal matrix $\mathbf{D} = \text{diag}$ ($\lambda_1, \lambda_2, \ldots, \lambda_N$) are also arranged in this order, and thus, equation (1) can be written as follows:

$$\mathbf{XX}^{\mathbf{T}}\mathbf{F} = \mathbf{FD}. \tag{2}$$

The $N \times N$ matrix $\mathbf{F}$ is called spatial modes coefficient matrix, which is also orthogonal (i.e., $\mathbf{FF}^{\mathbf{T}} = \mathbf{F}^{\mathbf{T}}\mathbf{F} = \mathbf{I}$), corresponding to the temporal modes coefficient matrix $\mathbf{A}$ or principal component (PC). The $N \times L$ matrix $\mathbf{A}$ is calculated by the following equation:

$$\mathbf{A} = \mathbf{F}^{\mathbf{T}}\mathbf{X}. \tag{3}$$

The column vector $\mathbf{x}_i$, N points records at time i, is reconstructed as

$$\mathbf{x}_i = \mathbf{F}\mathbf{a}_i. \tag{4}$$

Here, $\mathbf{a}_i$ is the column of $\mathbf{A}$ at time i, and obviously, $\mathbf{X} = \mathbf{FA}$. The k-th row of the matrix $\mathbf{A}$ is called the temporal k-th mode, and the element of the i-th column is the temporal coefficient at time i. Correspondingly, the column $\mathbf{F}_k$ is called the spatial k-th mode, and the elements of the j-th row of $\mathbf{F}$ (i.e., $\mathbf{F}$ (j)) represent the coefficients of each spatial mode of the j-th station. Thus, matrix element $\mathbf{F}_{jk}$ is the k-th spatial mode of the j-th station. The temporal modes are time-dependent, while spatial modes are space-dependent [13]. In addition, different spatial modes and different temporal modes are, respectively, orthogonal (i.e., $\mathbf{FF}^{\mathbf{T}} = \mathbf{F}^{\mathbf{T}}\mathbf{F} = \mathbf{I}$ and $\mathbf{AA}^{\mathbf{T}} = \mathbf{D}$). Finally, the eigenvalue λ_j of the j-th mode can be used to calculate the cumulative variance contribution rate of the first k modes to the total variance:

$$G(k) = \frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{N} \lambda_j} \times 100\%, \quad (k \le N). \tag{5}$$

The closer the G (k) approaches 100%, the more information the first k modes reflect of the original signals [36]. In spatiotemporal data analysis, we often only care about the first k modes with large variance contribution and regard them as the dominant modes. However, many EOF-based interpolation methods do not only consider the dominant modes, and the less important modes should also be considered. The optimal number of modes for reconstruction is determined by the root mean square error between the reconstruction sequence and the corresponding valid observation record [40].

2.2.2. IDW and EOFI. The IDW formula is given as follows:

$$W_j = \frac{1/d_j^P}{\sum_{j=1}^{N} 1/d_j^P} \quad (j = 1, \ldots, N), \tag{6}$$

$$\widetilde{Z}_{\text{IDW}} = \sum_{j=1}^{N} Z_j \cdot W_j, \tag{7}$$

$$\widetilde{X}_{\text{IDW}} = \sum_{j=1}^{N} \mathbf{X}(j) \cdot W_j, \tag{8}$$

where d_j denotes the distance between the j-th station and the target station, P is the inverse distance power parameter, W_j is the corresponding normalized weight, $\mathbf{X}$ (j) denotes the observation records sequence at the j-th station (i.e., the j-th row of $\mathbf{X}$), and $\widetilde{Z}_{\text{IDW}}$ and $\widetilde{X}_{\text{IDW}}$ represent IDW estimated value and estimated reconstruction sequence, respectively. IDW is based on Tobler's First Law of Geography: "everything is related to everything else, but near things are more related than distant things" [41]. The feature of this method is to produce "bull's eyes" around the observation points in the nearby area when observation points are rare and distributed sparsely [20]. For IDW, the common values of P are 1 and 2 (also called inverse squared distance weighting), so we only discuss the influence of these two parameters on IDW and EOFI in the later experiments.

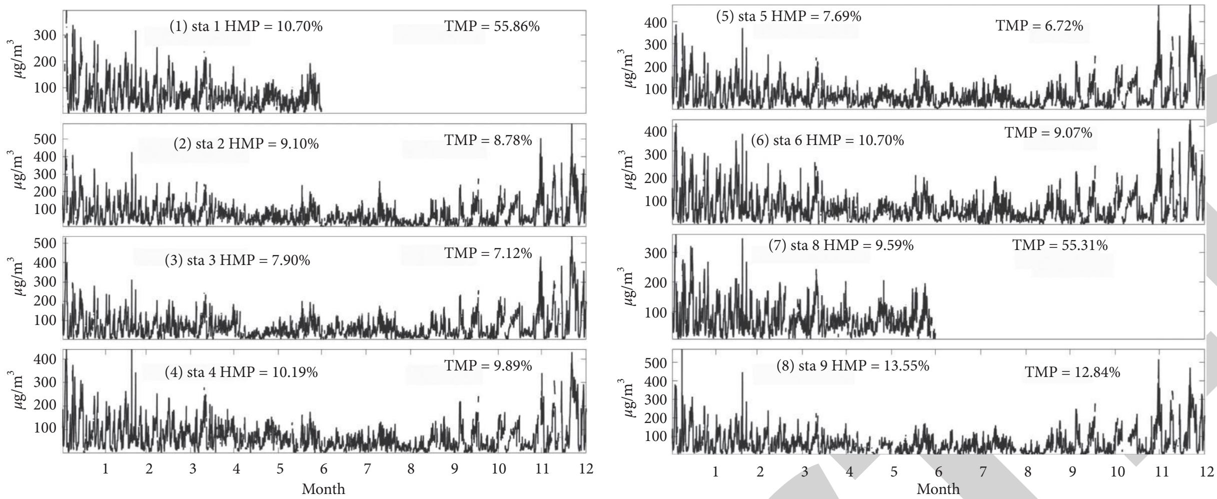

Figure 2: The original 12-month-long PM2.5 data records of 8 stations (sta 1, 2, 3, 4, 5, 6, 8, and 9) used in this study. The first half year missing percentage (HMP), i.e., the percentage of missing data in the first 4318 hours, and the total missing percentage (TMP), i.e., the percentage of missing data in 8735 hours, of each station are marked on the subgraphs.

In this study, the EOFI method steps are as follows: the missing data station shares the same temporal modes with observation stations, but the spatial modes $\widetilde{F}$ are estimated by the IDW interpolation of spatial modes of observation stations ($\mathbf{F}$ (j), $j = 1, \ldots, N$):

$$\widetilde{F} = \sum_{j=1}^{N} \mathbf{F}(j) \cdot W_j. \qquad (9)$$

Here, W_j is the same as the weight mentioned in IDW (equation (6)). Then, the $1 \times N$ row vector $\widetilde{F}$ and corresponding temporal mode $\mathbf{A}$ reconstruct the estimated value $\mathbf{x}_i^k$ at time i and estimated reconstruction sequence $\widetilde{X}_{\mathrm{EOF}}^k$ using the first k modes:

$$\begin{aligned} \mathbf{x}_i^k &= \widetilde{F}(1, 1{:}\ k)\mathbf{A}(1{:}\ k, i), \quad 1 \le k \le N, \\ \widetilde{X}_{\mathrm{EOF}}^k &= \widetilde{F}(1, 1{:}\ k)\mathbf{A}(1{:}\ k, 1{:}\ L), \quad 1 \le k \le N. \end{aligned} \qquad (10)$$

Using first k modes means that only the first k columns of $\widetilde{F}$ and the first k rows of $\mathbf{A}$ are considered. Finally, the optimal mode number for EOFI reconstruction is determined by the minimizing RMSE between the reconstructed sequence ($\mathbf{X}_{\mathrm{EOF}}^k$, $k = 1, \ldots, N$) and the corresponding valid observation sequence $\mathbf{X}$vid:

$$\widetilde{X}_{\mathrm{EOF}} = \min_{\mathrm{RMSE}(\mathbf{X}\mathrm{vid},\cdot)} \left\{ \widetilde{X}_{\mathrm{EOF}}^k, k = 1, \ldots, N \right\}. \qquad (11)$$

The spatial mode is deemed space-dependent and can reflect the spatial characteristics under the assumption of EOF decomposition. In this study, the estimated spatial mode $\widetilde{F}$ is closely related to the distance from the observation station. If the missing data station and the observation station are close in space, their spatial modes are also close to each other (larger weight, equation (6)); thus, the EOFI reconstruction sequence is also close to the observation sequence, which is consistent with our experience.

Prior to reconstruction, the raw data matrix $\mathbf{X}$ may contain missing values and cannot be directly EOF decomposed. Therefore, it is necessary to preprocess the raw data and get the data matrix without missing measured value before decomposition. Here, we first replace the missing values with observed values' space average at missing values time points and then apply linear interpolation to fill all the temporal intervals (i.e., spatial mean value substitution and temporal linear interpolation). Note that the temporal gaps should not be too large, so as to avoid that the interpolation affects the accuracy of dominant temporal and spatial modes [36]. In this study, the data used for EOF decomposition include the preprocessed records of stations 2, 3, 4, 5, 6, and 9 (near one year). Their temporal gaps of original records are short (Figure 2 (2–6, 8)), so we believe that the dominant modes are slightly affected and still reliable. The first half year records of sta 1 and sta 8 are both excluded from EOF decomposition.

2.2.3. *Thin Plate Spline Method and Surface Spline.*

The TPS method is a spatial interpolation method based on surface fitting, and it is one of the most frequently compared spatial interpolation methods [38], which was first proposed by Duchon [42]. It is often used to deal with uneven data in geoscience, such as generating continuous smooth elevation surface from discrete and sparse sample point elevation data. By simulating the bending of sheet metal, the TPS method generates a smooth surface with minimum bending energy through all observation points. Its form is as follows:

$$\widetilde{Z}_{\mathrm{TPS}} = \sum_{i=1}^{N} T_i d_i^2 \ln(d_i) + a + bx + cy. \qquad (12)$$

Among them, $d^2\log(d)$ term is the basic function and $a + bx + cy$ is the local trend function. The missing data station's horizontal coordinate (x, y) and its distances from the i-th ($i = 1, \ldots, N$) observation station are needed for TPS. In order to determine the $N + 3$ unknown parameter T_i ($i = 1, \ldots, N$), a, b, and c (equation (12)) are subject to the following relations:

$$
\begin{aligned}
&\sum_{j=1, j\neq i}^{N} T_j d_{ji}^2 \ln\left(d_{ji}\right) + a + bx_i + cy_i = Z_i,\\
&\sum_{i=1}^{N} T_i = 0,\\
&\sum_{i=1}^{N} T_i x_i\\
&\sum_{i=1}^{N} T_i y_i\\
&d_{ji}^2 = \left(x_j - x_i\right)^2 + \left(y_j - y_i\right)^2,
\end{aligned}
\tag{13}
$$

with the N observation points' horizontal coordinates (x_i, y_i, $i = 1, \ldots, N$), distances between each other (d_{ji}, $i, j = 1, \ldots, N$), and observation values (Z_i, $i = 1, \ldots, N$), a smooth surface ($N + 3$ linear equations and $N + 3$ unknown parameters) is generated, the value at the missing data station is also assumed to be on this surface, and then, the TPS estimated value $\widetilde{Z}_{\text{TPS}}$ is calculated by equation (12). The TPS matrix form was fully described in Bookstein [43], and the coefficient matrix of unknown parameter is only related to spatial attributes (coordinate and distance), but not to time attribute.

The surface spline (SS) method is also a good spatial interpolation method based on surface fitting. It generates smooth surfaces through discrete points too. However, the basic function of the SS method is different from TPS. It does not consider trend term, the fitting function is different, and the radius R is introduced. Guo et al. [44] used the SS method to interpolate the bottom friction coefficient of the selected independent points to obtain values for the entire Bohai Sea and combined the adjoint assimilation method to invert the bottom friction coefficient of the entire sea. The SS method is also used for the inversion of initial conditions and parameters estimation in the ocean pollutant transport model [21], which is a significant improvement over the Cressman interpolation. Its form is as follows:

$$
\widetilde{Z}_{\text{SS}} = \sum_{j=1}^{N} S_j \left(\frac{d_j^2}{R^2} \ln \frac{d_j^2}{R^2} + 1 - \frac{d_j^2}{R^2} \right), \tag{14}
$$

$$
d_j^2 = \left(x - x_j\right)^2 + \left(y - y_j\right)^2. \tag{15}
$$

Similar to TPS, the N observation points' spatial attributes and observation values sequences $\mathbf{z}$ generate a smooth surface, and then, the unknown parameter column vector $\mathbf{s}$ is solved by the matrix form:

$$
\begin{aligned}
&\mathbf{Ds} = \mathbf{z},\\
&\mathbf{D} = \left(\mathbf{D}_{ij}\right)_{N\times N},\\
&\mathbf{s} = \left(S_1, \ldots, S_N\right)^T,\\
&\mathbf{z} = \left(Z_1, \ldots, Z_N\right)^T,\\
&\mathbf{D}_{ij} = \begin{cases} \dfrac{d_{ij}^2}{R^2} \ln \dfrac{d_{ij}^2}{R^2} + 1 - \dfrac{d_{ij}^2}{R^2}, & i \neq j,\\ 1, & i = j. \end{cases}
\end{aligned}
\tag{16}
$$

Here, the elements of parameter matrix $\mathbf{D}$ are only related to the distance between observation points d_{ij} ($i, j = 1, \ldots, N$) and prescribed radius R. The radius R is set to 15 km because the distance between any two stations is within this radius. After solving the unknown sequence $\boldsymbol{s}$, SS estimated value $\widetilde{Z}_{\text{SS}}$ of missing data station is calculated with equations (14) and (15). Note that the value of $\mathbf{s}$ changes with radius R, but selecting R within the appropriate range will not have a great impact on the final interpolation result.

2.3. Evaluation Indices. At the end of Section 2.2.2, the preprocessing of the original data has been mentioned. We emphasize that the preprocessed data used for each interpolation method is the same. Therefore, the evaluation of different interpolation methods is persuasive and reliable. Table 1 summarizes their parameter settings. We will list a series of quantitative indices to evaluate these interpolation methods [38]. The evaluation indices listed in this study include mean absolute error (MAE), root mean square error (RMSE), correlation coefficient (Corr), and deviation rate bias, Nash–Sutcliffe efficiency (NSE) [45], and index of agreement (IA) (or Willmott's D) [46].

Among them, MAE (equation (17)) and RMSE (equation (18)) are often used as indicators of the performance of interpolation or models [38]. The smaller they are, the better the interpolation effect is. Corr (equation (19)) and bias (equation (20)) measure the correlation and deviation between simulation value sequence S and the observation series O, and $\overline{S}$ and $\overline{O}$ are their average values, respectively. Higher degree of correlation and smaller deviation both indicate the better interpolation effect. NSE (equation (21)) is a common index used to measure the performance or interpolation effect in meteorological, hydrological, and environmental models. Its value ranges from negative infinity to 1. The closer to 1, the simulation results are closer to observations; the closer to 0, the result are closer to the observation average values, but the process error is large, while negative NSE indicates that the performance of mean observed values is even better than simulated values and indicates this simulation unacceptable. IA (equation (22)) is referred as the potential error. IA is a nondimensional and bounded index with values closer to 1 indicating better agreement. The above six indices are defined as follows:

$$
\text{MAE} = \frac{\sum |S - O|}{n}, \tag{17}
$$

$$
\text{RMSE} = \sqrt{\frac{\sum (S - O)^2}{n}}, \tag{18}
$$

$$
\text{Corr} = \frac{\sum (S - \overline{S})(O - \overline{O})}{\sqrt{\sum (S - \overline{S})^2}\sqrt{\sum (O - \overline{O})^2}}, \tag{19}
$$

$$
\text{bias} = \frac{\sum |S - O|}{\sum O} \times 100\%, \tag{20}
$$

Table 1: Four interpolation methods parameter setting.

Method	Parameter setting
IDW	$P = 1$ and 2
EOFI	Spatial mode is dependent on IDW
TPS	—
SS	$R = 15$ km

$$\mathrm{NSE} = 1 - \frac{\sum (S - O)^2}{\sum (O - \overline{O})^2}, \tag{21}$$

$$\mathrm{IA} = 1 - \frac{\sum (S - O)^2}{\sum \left(|O - \overline{O}| + |S - \overline{O}|\right)^2}. \tag{22}$$

In Section 3.2, we calculated the above six evaluation indicators, which reflect the accuracy of these simulations, and the indicators for the EOFI first k modes ($k = 1, \ldots, N$) are also calculated. The results of EOFI with the optimal mode number will be compared with other three interpolation methods.

2.4. Site Selection. To pursue better interpolation performance, here we just choose the data of the five nearest stations for interpolation; that is, the imputation of sta 1 and sta 8 data is based on the data of stations 2, 3, 4, 5, and 6 and the data of stations 2, 4, 5, 6, and 9 (Figure 1(b)), respectively, while the data of other stations are not included. The near one-year records of sta 1 and sta 8 are reconstructed, respectively, by interpolating data of the five nearest stations with four interpolation methods, and then, the reconstructed sequences are compared with corresponding valid observation data in the first half year (Figure 2) to calculate the evaluation index. In Section 4.2 for further validation, multiple sets of experiments in different time periods are implemented, and the RMSE between four interpolation methods' reconstruction sequence and corresponding valid observation records are further compared.

3. Results

3.1. Interpolation Result of Four Methods. The distances between the observation stations and the target station and the corresponding normalized weight are presented in Table 2. The distance from sta 4 is the shortest, and the weight is the largest in the sta 1 group, while distance from sta 5 is the shortest, and the weight is the largest in the sta 8 group. With the increase in IDW and EOFI power parameter P (from 1 to 2), the normalized weights of the nearest stations (sta 4 and sta 5) increase, while the weights of other stations decrease. Therefore, the estimated spatial mode $\widetilde{F}$ of sta 1 and sta 8 calculated by equation (9) is more affected by those of sta 4 and sta 5, respectively.

The temporal modes or principal components (PCs) of sta 1 and sta 8 (Figure 3) and the corresponding spatial modes (Table 2) are obtained by EOF decomposition. It can be seen that the variance contribution rate of PC1 in sta 1 and sta 8 is both over 98%, and the spatial 1st modes are all around 0.44. Most of the other modes of PC change around 0 (Figure 3 (a2–a5 and b2–b5)), and the corresponding absolute value of spatial modes is also less than the first mode. Therefore, from the second PC to the fifth PC, these modes play a less important role in reconstructing data than the first mode, but the later indices show that ignoring these less important modes may lead to less perfect performance of EOFI reconstruction. In addition, Figure 3 (a1 and b1) illustrates that the amplitudes of PC1 in winter months (November, December, January, and February) were significantly greater than those in summer months (April, May, June, and July). It demonstrates that PM2.5 concentration in winter in North China Plain was significantly higher than that in summer [47].

Figures 4 and 5 depict the four interpolation reconstruction sequences and their residuals for sta 1 and 8, respectively. Both power parameters P (1 or 2) are adopted for IDW and EOFI reconstruction for sta 1 and sta 8, but the indices show that choosing $P = 1$ for IDW and EOFI is more accurate in sta 1, while $P = 2$ for IDW and EOFI is more accurate in sta 8. The optimal mode number for EOF reconstruction is both three in sta 1 and sta 8. In the part of Result Evaluation and Discussion, we try to explain the reasons for this. It can be seen that four methods can roughly reproduce the valid records in sta 1 and sta 8. In sta 1 (Figure 4), the residuals of the four interpolation methods all change near 0, but there are several errors which are quite different from the observed values. For example, they all show errors of more than 100 μg/m^3 around February 20th and mid-March. Regardless of the instrument failure and other factors, the large error at these times may indicate that the PM2.5 concentration varies greatly among different regions of the same city, and it is not accurate to rely on only the adjacent data in this case. In Figure 5 of sta 8, the situation is similar, but the fluctuation magnitude of the residual sequence is significantly larger than that of sta 1, and the large residuals are also more frequently occurred. The performance of the four methods in sta 8 is generally worse than that of sta 1.

3.2. Result Evaluation. In this section, we evaluate four interpolation methods with quantified indices. Figure 6 shows a comparison of 4 interpolation methods in terms of MAE, RMSE, and Corr, and Figure 7 shows bias, NSE, and IA. Because many indices of the TPS method are quite different from those of other methods, in order to see their differences clearly, the indicator values of TPS are directly marked on each subgraph. It can be seen that the EOFI interpolation performance of sta 1 and sta 8 varies with the number of modes, many indices show that the optimal mode number of EOFI is three (e.g., Figure 6 (a1 and b1)), and the performance of EOFI is sometimes worse than other interpolation methods when it is not the optimal mode number. We arrange all six indices of the best performing EOFI and other three interpolation methods in the descending order of performance. It can be seen that, in sta 1, all 6 indicators show that the performance of EOFI ($P = 1$) is the best (red lines) (1-EOFI>1-IDW > SS > TPS), while in sta 8, all 6 indicators show that EOFI ($P = 2$) is the best (green

Table 2: Distances from station 1 and 8 to other five observation stations and corresponding normalized inverse distance weights when power $P = 1$ and 2, respectively.

Target station	Station 1					Station 8				
Observation stations	2	3	4	5	6	2	4	5	6	9
Distance (km)	9.19	7.62	3.86	7.52	4.43	7.88	7.48	3.96	7.36	6.07
Weight ($P = 1$)	0.1268	0.1530	0.3018	0.1550	0.2634	0.1560	0.1643	0.3105	0.1669	0.2023
k-th mode ($P = 1$)	0.4471	0.1179	0.0579	−0.0574	0.0592	0.4449	−0.0714	0.0192	0.0076	0.1143
Weight ($P = 2$)	0.0718	0.1045	0.4066	0.1073	0.3098	0.1124	0.1247	0.4453	0.1287	0.1889
k-th mode ($P = 2$)	0.4474	0.2282	0.0871	−0.1222	0.1191	0.4426	−0.1291	0.0332	0.0227	0.2540
G (k) (%)	98	98.78	99.35	99.72	100	98.14	98.78	99.38	99.76	100

It is noteworthy that EOF estimated spatial 1st, 2nd, 3rd, 4th, and 5th modes are listed in the 5th and 7th rows (rather than observation station's coefficients). The contribution of the first k modes to the total variance G (k) is listed in the last row.

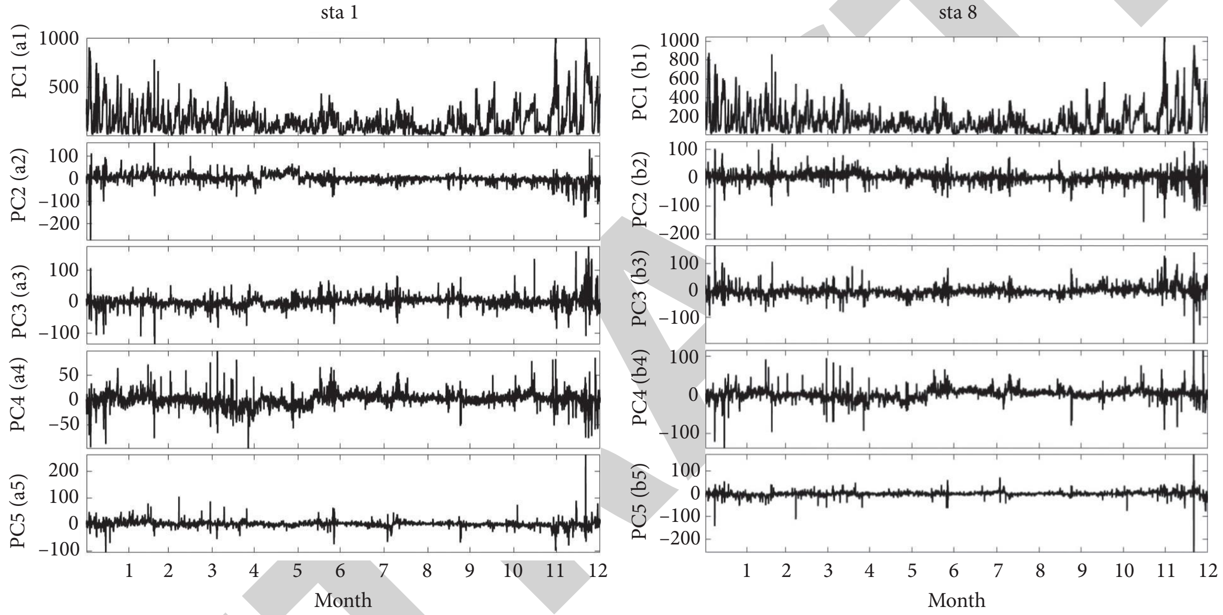

Figure 3: Station 1 (a1–a5) and station 8 (b1–b5) temporal variation in EOF modes.

lines) (2-EOFI > 2-IDW > SS > TPS). The IDW performance of many indices is similar; sta 1 prefers $P = 1$, while sta 8 prefers $P = 2$. In addition, the indices performance of sta 8 is generally worse than that of sta 1. In Section 4.1, we try to explain why different parameters are chosen at the two sites.

4. Discussion

4.1. IDW Power P Choice and Sites Number Impact on EOFI. For the EOFI of this study, we did not take the data of sta 1 and sta 8 into EOF decomposition. The spatial modes of these two stations are calculated by the spatial modes of other 5 stations with IDW, and of course, their spatial modes estimates can also be obtained by other methods, such as Pan and Lv [36] using linear and spline interpolation, respectively, to calculate the spatial modes of river water level measurement points. Next, we try to explain why different P values are chosen in the two sites as mentioned in Section 3 and discuss the influence of the number of data sites on the EOFI reconstruction.

Firstly, the indices performance of sta 8 is obviously inferior to those of sta 1. There are four same stations (stations 2, 4, 5, and 6) data selected by both sta 1 and sta 8. But the number of missing values at sta 9 for sta 8 imputation is more than that of the sta 3 for sta 1 (the first half of the year missing percent of the sta 9 in Figure 2 reaches 13%), so the completeness of the original data may account for the worse results of sta 8. In addition, for sta 8, when P is increased from 1 to 2, the EOFI spatial modes and reconstruction sequence will be more dependent on the spatial modes (Table 2) and observation records of the closest station (sta 5), respectively. The adverse impact of the data of sta 9 is reduced, which may be an explanation of sta 8's preference for $P = 2$.

Furthermore, in previous experiment, data of sta 1 and sta 8 are reconstructed with the data of the other 5 adjacent stations, of which 4 stations (stations 2, 4, 5, and 6) are both used for reconstruction of sta 1 and sta 8. In order to further explore the influence of the remaining station on the interpolation results, another experiment is conducted where the data of sta 3 are not used for sta 1 reconstruction and the

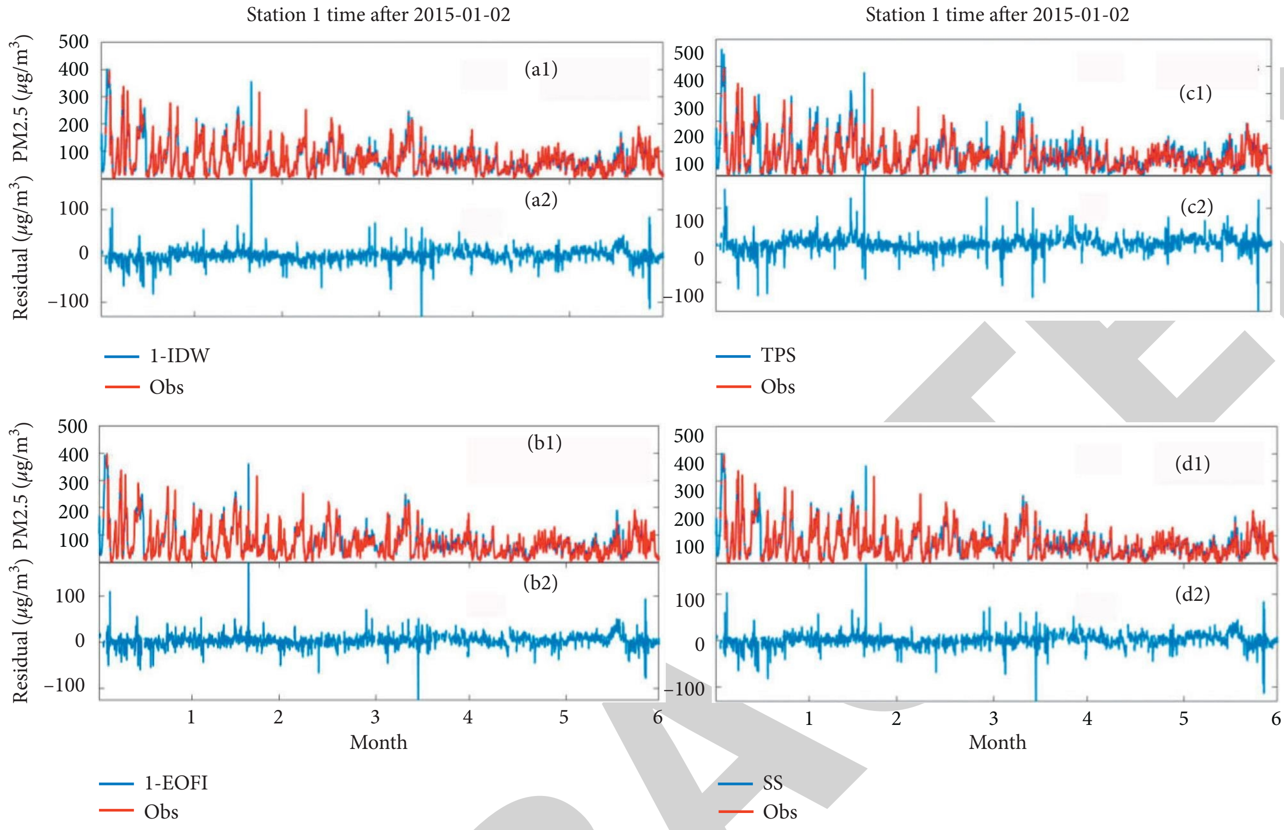

Figure 4: 1-IDW (a1),1-EOFI (b1), TPS (c1), and SS (d1) reconstruction sequence and observations in sta 1 and their corresponding residuals (a2, b2, c2, and d2), 4138 hours after 2015-01-02 1:00. The numbers in front of EOFI and IDW represent the value of their inverse distance weight P; for example, 1-EOFI represents the value of EOFI's inverse distance weight P is 1.

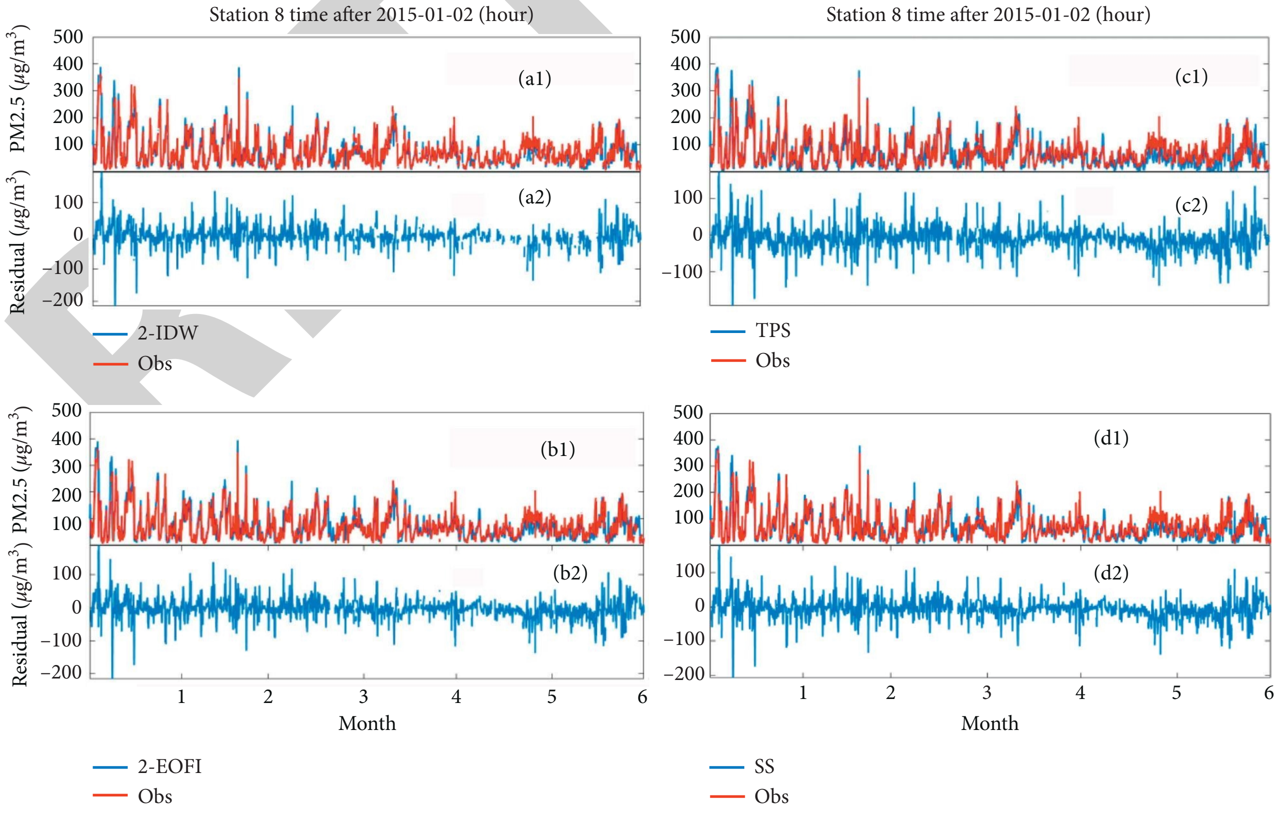

Figure 5: Same as Figure 4, but the IDW and EOFI use $P=2$ for sta 8.

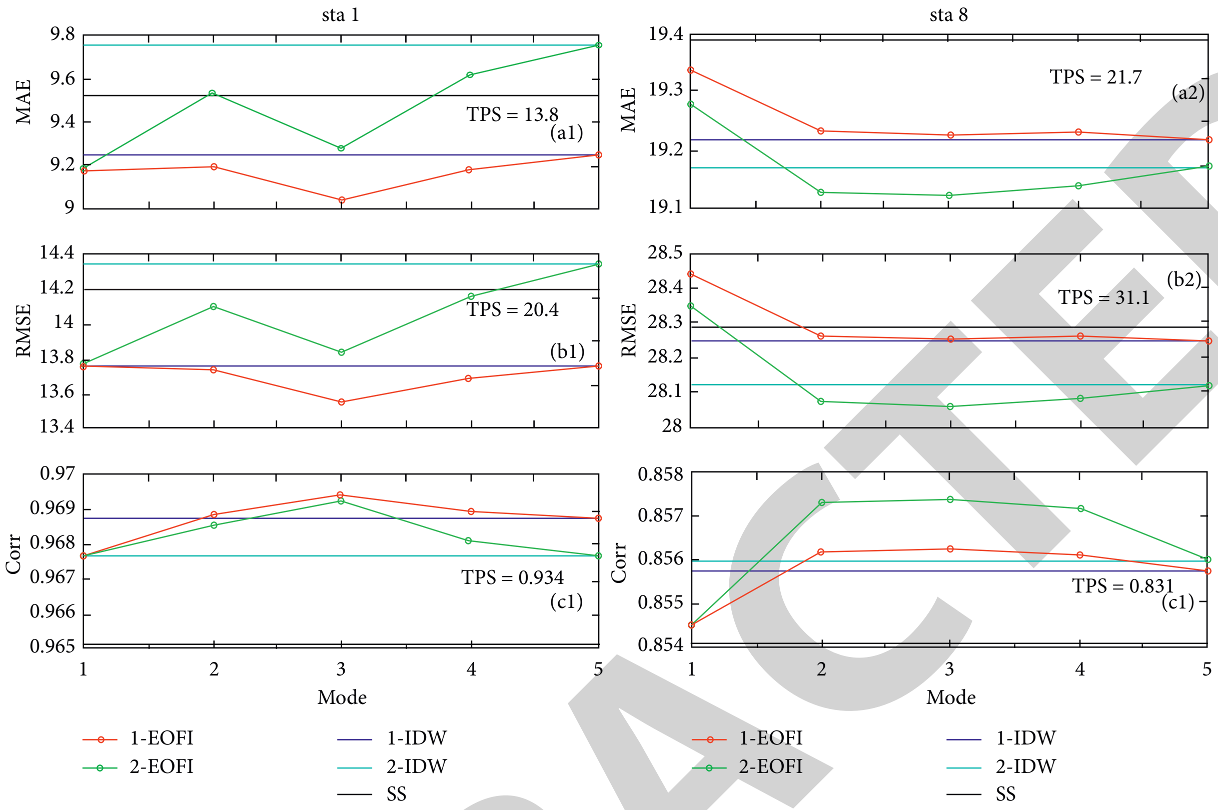

Figure 6: MAE, RMSE, and Corr of 4 interpolation methods in sta 1 and sta 8. The horizontal axis is the number of the first k modes selected in EOFI. The results of the other three interpolation methods are all straight lines and irrelevant to the modes.

data of sta 9 are not used for sta 8. The 4 sites and 5 sites EOFI reconstruction results are shown in Table 3.

It can be seen that, for both sta 1 and sta 8, the EOFI reconstruction with 5 sites is better than that with only 4 sites. In addition, inclusion of data from coastal stations such as sta 10 (Figure 1(b), far away from sta 1 and sta 8) in EOFI is not as good as interpolation with data from only five nearest sites. It is very vital to determine the appropriate number of stations for EOFI according to the feature and quality of the original data. As we can see the performance of using less sites data or adding costal sites data for EOFI, both of which are worse than that of only five nearest sites data.

4.2. Further Validation and Impact of Data Time Length on EOFI Results. In the previous experiment, EOFI selected PM2.5 data of nearly a full year from five adjacent stations data to perform EOF decomposition and obtained nearly a full year of PC and corresponding spatial modes. In this part, a number of experiments with different lengths of record are implemented to further evaluate and compare the four interpolation methods. Since there are only valid observation records in the first half of 2015 for both sta 1 and sta 8, the reconstruction sequence of four interpolation methods must be compared with valid observations during the same period. Divided by the calendar month, we divided the records in the first half of the year into six one-month sections (Jan, 1; Feb, 2; Mar, 3; Apr, 4; May, 5; and Jun, 6) in the experimental group E1 and five two-month sections (1-2, 2-3, 3-4, 4-5, and 5-6) in experimental group E2. Four three-month sections (1–3, 2–4, 3–5, and 4–6) are implemented in the experimental group E3. Similarly, E4, E5, and E6 represent the experimental groups with a duration of 4, 5, and 6 months, respectively. There are 21 experiments in total. Since the temporal mode of EOF decomposition is related to the continuity of record, experimental groups with continuous months are set to reduce the inaccuracy of the temporal and spatial modes of EOF decomposition. February in winter and June in summer represents different seasons, and the feature of PM2.5 concentration is significantly related to the seasons. For example, in winter, more fossil fuels may be consumed for heating; therefore, the PM2.5 concentration is significantly higher than other seasons.

Figure 8 depicts the main results of EOFI reconstruction sequence of sta 1 and sta 8. It can be seen that, although the spatial 2nd, 3rd, 4th, and 5th modes in different time periods are different, the spatial 1st mode always remains stable at around 0.44, and the corresponding variance contribution also accounts for more than 95% (c1 and c2), which is consistent with the previous results. The RMSE range of EOFI reconstruction for sta 1 is 10–16 $\mu g/m^3$ (b1), while the range for sta 8 is 22–36 $\mu g/m^3$ (b2). The range is also consistent with the previous results, which shows the stability of the EOFI method. In addition, the number of experiments with the optimal mode number 4 (i.e., using first 4 modes to reconstruct) for sta 1 and sta 8 are both largest,

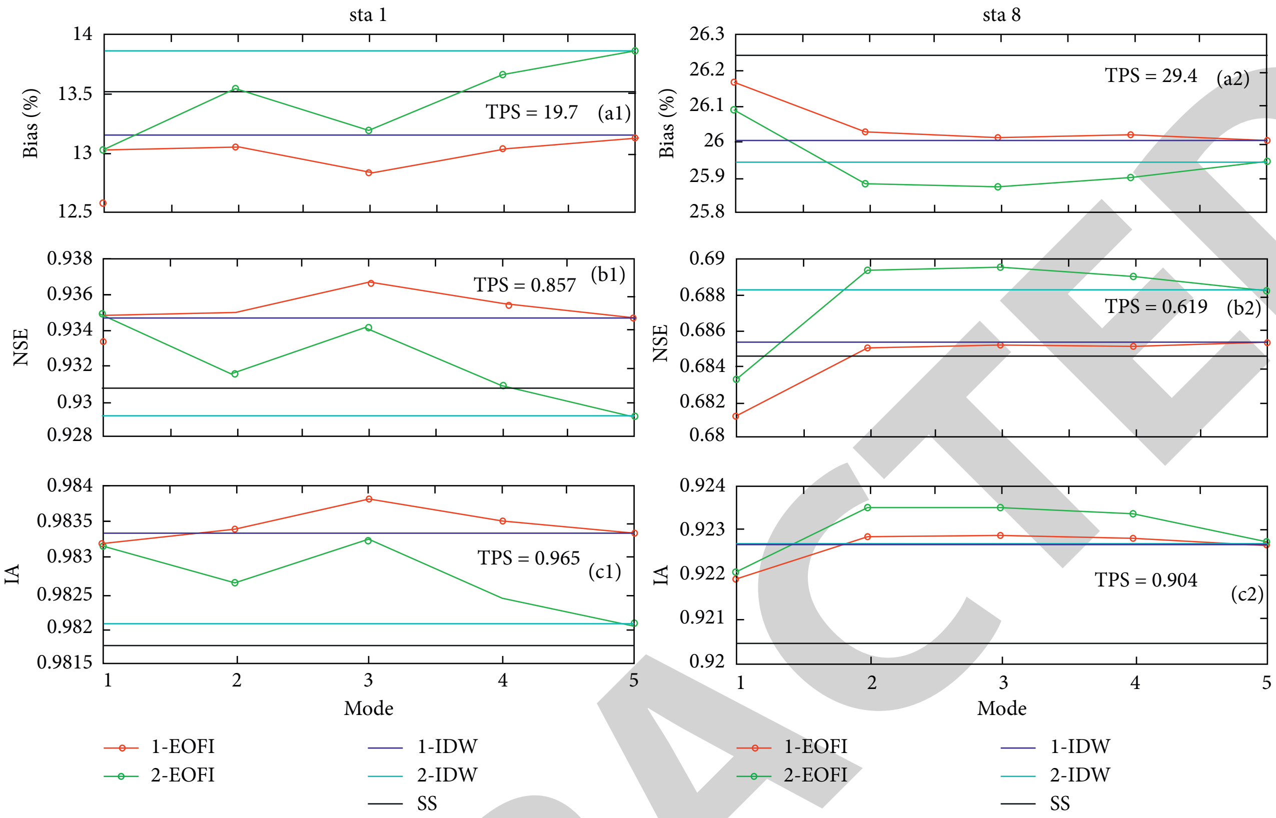

Figure 7: Same as Figure 6, but the indices are bias, NSE, and IA in sta 1 and sta 8.

Table 3: RMSE between the results of EOFI reconstruction and the valid observation records of 4 and 5 stations selected by sta 1 and sta 8, respectively.

RMSE (μg/m^3)	Stations	Modes				
		1	2	3	4	5
sta 1 ($P = 1$)	2, 3, 4, 5, 6	13.759	13.739	13.558	13.686	13.764
	2, 4, 5, 6	14.313	14.116	14.486	14.582	—
sta 8 ($P = 2$)	2, 4, 5, 6, 9	28.344	28.072	28.061	28.082	28.119
	2, 4, 5, 6	29.079	28.991	28.923	28.652	—

respectively, but there are still other optimal mode numbers. The optimal mode number can be determined by finding the smallest RMSE [40].

Table 4 compares the performance (in terms of RMSE) of four interpolation methods reconstruction sequence. Among the 21 experiments, there are 19 experiments in sta 1 and 13 experiments in sta 8 showing the RMSE of EOFI reconstruction is the smallest, respectively. There are also another 7 groups in sta 8 showing SS performed best in terms of RMSE, and these groups mainly include winter months January, February, and March. We infer that this is due to large PM2.5 concentration difference in different sites in winter, and the accuracy of spatial and temporal modes is not as good as those of other seasons.

4.3. Comparison between EOFI and DINEOF. There have been many EOF-based interpolation methods (e.g., DCCEOF in [10], EOFI in [36], and VE-DINEOF in [40]). One of the most widely utilized methods is the iterated EOF method, DINEOF [30]. Therefore, it is necessary to compare DINEOF and EOFI in this study.

First of all, two methods are both based on the matrix eigenvalue decomposition theory, and they all assume that the short missing value intervals of original spatiotemporal observation records will not affect the dominant temporal and spatial modes significantly. Moreover, the first guess values are given to the missing values to enable matrix decomposition. By calculating the RMSE and other indicators, the temporal and spatial modes of the optimal mode number will be used for final reconstruction.

However, the most significant difference between DINEOF and EOFI is the original data used for matrix decomposition. In EOFI, the data of sta 1 and sta 8 (the second half of the year data is missing) are not included in the decomposed matrix, but in DINEOF, the data of sta 1 and sta 8 are taken into EOF decomposition; firstly, the missing values are replaced with first guess values and then

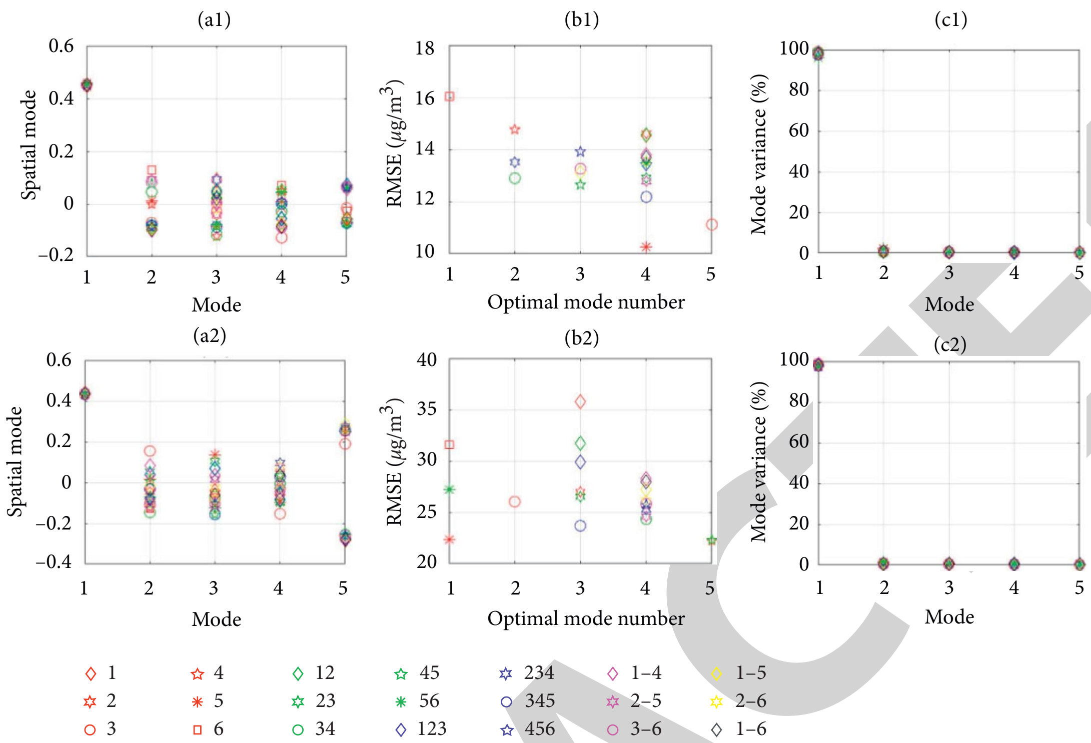

Figure 8: EOFI reconstruction sequence's spatial modes, corresponding variance contributions, and the RMSE with optimal mode number of sta 1 (a1, c1, and b1) and sta 8 (a2, c2, and b2) in the first half of the year. The legend color red, green, blue, magenta, yellow, and black represent the experimental groups with durations of 1, 2, 3, 4, 5, and 6 months, respectively.

Table 4: RMSE of four interpolation methods in sta 1 and sta 8.

RMSE	Station	1				8			
	Method	IDW	EOFI	TPS	SS	IDW	EOFI	TPS	SS
Groups	Month(s)								
E1	1	14.562	14.547*	23.675	15.878	36.031	35.751	38.931	35.458*
	2	14.629	14.613	20.924	13.780*	27.065	26.984	28.915	26.103*
	3	11.123*	11.123*	13.989	11.380	26.376	26.040	27.469	25.815*
	4	15.001	14.772*	22.446	15.689	22.207*	22.207*	24.905	22.820
	5	10.286	10.273*	16.905	10.968	22.399	22.313*	27.000	24.248
	6	16.229	16.029*	23.275	16.749	32.065	31.596*	36.881	33.100
E2	1-2	14.596	14.557*	22.308	14.840	31.889	31.688	34.317	31.159*
	2-3	12.980	12.908	17.769	12.627*	26.723	26.519	28.203	25.960*
	3-4	13.120	12.901*	18.518	13.610	24.371	24.339*	26.212	24.357
	4-5	12.732	12.650*	19.716	13.406	22.303*	22.303*	25.965	23.539
	5-6	13.545	13.478*	20.296	14.116	27.549	27.215*	32.208	28.913
E3	1–3	13.507	13.459*	19.855	13.753	30.175	29.934	32.211	29.497*
	2–4	13.650	13.508*	19.368	13.668	25.301	25.273	27.143	24.953*
	3–5	12.209	12.178*	17.973	12.752	23.737	23.692*	26.475	24.321
	4–6	14.016	13.913*	20.991	14.625	25.846	25.758*	29.911	26.984
E4	1–4	13.878	13.791*	20.503	14.239	28.390	28.262	30.545	27.974*
	2–5	12.858	12.791*	18.757	13.018	24.614	24.594*	27.108	24.780
	3–6	13.340	13.262*	19.450	13.872	25.981	25.837*	29.312	26.693
E5	1–5	13.204	13.114*	19.800	13.616	27.310	27.196*	29.878	27.278
	2–6	13.607	13.542*	19.754	13.853	26.204	26.142*	29.232	26.574
E6	1–6	13.764	13.691*	20.432	14.197	28.119	27.985*	31.096	28.283

The power P of IDW and EOFI is based on the analysis of Section 4.1 (i.e., $P = 1$ for sta 1 and $P = 2$ for sta 8), and the EOFI reconstruction with optimal mode number is considered. "*" represents the smallest RMSE of this experiment.

conducted matrix decomposition and iterative replacement until convergence. However, this step may be not suitable for the data processing of a small number of stations because the first guess values of these missing stations may greatly affect the accuracy of temporal and spatial modes in this case. Even if the final convergent temporal and spatial modes are obtained through iteration, the calculation resources consumed may be huge. Alvera-Azcárate et al. [32] mentioned that the data points with missing percent more than 95% are removed before data decomposition because they cannot provide effective information. The number of data points involved in their decomposition is huge; therefore, these less-informative points' removal has little impact on the final results. The DINEOF has been widely used for reconstruction of gap-free satellite images where densely sampled and numerous observations are obtained by remote sensing, while in other platforms (e.g., PM2.5 land-based stations in this study and offshore buoy stations array), where observations are relatively rare and sparse sampled, the temporal and spatial modes of iterated EOF methods may be not accurate when there is a large proportion of missing values in the few sites observation data matrix.

Therefore, for the observation records of finite number stations, if we want to make full use of the data of station with large proportion of missing values, EOFI may be more suitable for this kind of interpolation. The superiority of EOFI here is to obtain more reasonable spatial and temporal modes by excluding the records of large missing percent stations before EOF decomposition. All stations share the same time-dependent temporal mode, while the space-dependent spatial mode of the missing data station is estimated by spatial interpolation (IDW is used in this study), and the spatial mode features and patterns are considered. In addition, EOFI can provide more reasonable first guess values for the data of these missing stations, and next, DINEOF is used to iteratively calculate until convergence. For other differences, such as DINEOF iterative decomposition, EOFI can also use iterative decomposition in this study; DINEOF randomly selects a part of observation data as cross validation points, and EOFI here uses the first half year valid observation records and monthly records of sta 1 and sta 8 as check points, both of which can be unified in these aspects.

5. Conclusion

In this paper, two-dimensional EOFI is introduced and applied to reconstruct spatial-distributed PM2.5 data as an extension to one-dimensional EOFI in river water level reconstruction. The main step of EOFI here is to calculate the missing data station's estimated spatial modes $\widetilde{F}$ by IDW interpolation of spatial modes of the observation sites and then multiply $\widetilde{F}$ and the corresponding temporal modes to obtain the EOFI reconstruction sequence, and the optimal mode number of EOFI reconstruction is determined by minimizing RMSE. Compared with the other three interpolation methods (IDW, TPS, and SS), the quantitative indices show that EOFI can improve the interpolation effect. The conclusion is as follows.

TPS and SS have fixed function forms, and their coefficient matrices are space-dependent. The advantage of EOFI is that the spatiotemporal matrix is decomposed into time-dependent temporal modes and space-dependent spatial modes under EOF assumption. Observation stations and missing data stations share the same temporal modes, while the spatial modes of missing data station are estimated by the IDW of observation stations' spatial modes. The benefit of IDW is that when the distance between the missing station and the observation station is very close, the spatial mode estimated by IDW is very close to that of the observation station; thus, the EOFI reconstruction sequence of the missing station is also close to the data of the observation station, which is consistent with our cognition. More essentially, the IDW weights of neighboring points are generated by statistical estimate of covariance between the observation points. TPS and SS weights do not depend on the statistical features of interpolated fields. EOFI can reduce MAE and RMSE compared with other three methods, and other indices show that the performance of EOFI is better too. This shows that EOFI can improve the interpolation effect with optimal modes. The results of several experimental groups with different data lengths show that the dominant spatial modes of EOF decomposition almost do not change with the time length, which is consistent with the EOF assumption that the spatial modes are independent of time. At the same time, the RMSE of EOFI reconstruction with optimal mode number still shows the advantages over the other three methods.

The proposed method is suitable for interpolation when observations are rare and sparsely distributed, and there are large percent of missing values for some stations' original records. The EOFI reconstruction sequence of missing data station can be a reasonable first guess value for further DINEOF (or other iterated EOF-based method) steps.

EOFI has the advantages of less calculation, less parameter choices, and ease of implementation and can be extended to fill the missing data gaps of other two-dimensional spatial distribution physical variables. The limitation of EOFI is that the missing values' temporal and space gaps should not be too large; otherwise, it will affect the accuracy of spatial and temporal modes. At the same time, the quality of the original data has an impact on the reconstruction results. High quality and complete observation data can produce more accurate spatial and temporal modes, which is conducive to EOFI reconstruction.

Data Availability

The data (hourly PM2.5 concentration data of 8 stations in Tianjin and station locations) used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors would like to thank Professor Yang Gao for providing the PM2.5 concentration data. This work was supported by the National Natural Science Foundation of China (Grant no. 41876003) and the National Key Research and Development Program of China (Grant nos. 2017YFA0604101 and 2016YFC1401404).

References

[1] S. Zhai, D. J. Jacob, X. Wang et al., “Fine particulate matter (PM2.5) trends in China, 2013–2018: separating contributions from anthropogenic emissions and meteorology,” *Atmospheric Chemistry and Physics*, vol. 19, pp. 11031–11041, 2019.

[2] S. Gautam, A. K. Patra, and P. Kumar, “Status and chemical characteristics of ambient PM2.5 pollutions in China: a review,” *Environment, Development and Sustainability* vol. 21, pp. 1649–1674, 2018.

[3] H. Shi, S. Wang, J. Li, and L. Zhang, “Modeling the impacts of policy measures on resident’ s PM2.5 reduction behavior : an agent-based simulation analysis,” *Environmental Geochemistry and Health*, vol. 1, 2019.

[4] Y. Li, J. Wang, C. Chen, Y. Chen, and J. Li, “Estimating PM2.5 in the Beijing-tianjin-hebei region using modis aod products from 2014 to 2015,” *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 41, pp. 721–727, 2016.

[5] X. Liu and Coauthors, “Fine particulate matter pollution in North China: seasonal-spatial variations, source apportionment, sector and regional transport contributions,” *Environmental Research*, vol. 184, Article ID 109368, 2020.

[6] J. Feng, J. Quan, H. Liao, Y. Li, and X. Zhao, “An air stagnation index to qualify extreme haze events in northern China,” *Journal of the Atmospheric Sciences*, vol. 75, pp. 3489–3505, 2018.

[7] X. Wu, Y. Chen, J. Guo, G. Wang, and Y. Gong, “Spatial concentration, impact factors and prevention-control measures of PM2.5 pollution in China,” *Natural Hazards*, vol. 86, pp. 393–410, 2017.

[8] L. Zhou, C. Zhou, F. Yang, L. Che, B. Wang, and D. Sun, “Spatio-temporal evolution and the influencing factors of PM2.5 in China between 2000 and 2015,” *J. Geogr. Sci.*, vol. 29, pp. 253–270, 2019.

[9] P. Yin and Coauthors, “Higher risk of cardiovascular disease associated with smaller size-fractioned particulate matter,” *Environmental Science & Technology Letters*, vol. 7, pp. 95–101, 2020.

[10] K. Bai, K. Li, J. Guo, Y. Yang, and N.-B. Chang, “Filling the gaps of in situ hourly PM2.5 concentration data with the aid of empirical orthogonal function analysis constrained by diurnal cycles,” *Atmospheric Measurement Techniques*, vol. 13, pp. 1213–1226, 2020.

[11] A. Alvera-Azcárate, A. Barth, D. Sirjacobs, F. Lenartz, and J. M. Beckers, “Data interpolating empirical orthogonal functions (DINEOF): a tool for geophysical data analyses,” *Mediterranean Marine Science*, vol. 12, pp. 5–11, 2011.

[12] D. Kondrashov and M. Ghil, “Spatio-temporal filling of missing points in geophysical data sets,” *Nonlinear Processes in Geophysics*, vol. 13, pp. 151–159, 2006.

[13] J. Elken, M. Zujev, J. She, and P. Lagemaa, “Reconstruction of large-scale Sea surface temperature and salinity fields using sub-regional EOF patterns from models,” *Frontiers Earth Science*, vol. 7, pp. 1–20, 2019.

[14] L. Feng, G. Nowak, T. J. O. Neill, and A. H. Welsh, “CUTOFF : a spatio-temporal imputation method,” *Journal of Hydrology*, vol. 519, pp. 3591–3605, 2014.

[15] S. Moritz and T. Bartz-Beielstein, “ImputeTS: time series missing value imputation in R,” *The R Journal*, vol. 9, pp. 207–218, 2017.

[16] M. W. Beck, N. Bokde, G. Asencio-Cortés, and K. Kulat, “R package imputetestbench to compare imputation methods for Univariate time series,” *The R Journal*, vol. 10, pp. 218–233, 2018.

[17] N. Bokde, M. W. Beck, F. Martínez Álvarez, and K. Kulat, “A novel imputation methodology for time series based on pattern sequence forecasting,” *Pattern Recognition Letters*, vol. 116, pp. 88–96, 2018.

[18] M. Lepot, J. B. Aubin, and F. H. L. R. Clemens, “Interpolation in time series: an introductive overview of existing methods, their performance criteria and uncertainty assessment,” *Water (Switzerland)*, vol. 9, 2017.

[19] G. Y. Lu and D. W. Wong, “An adaptive inverse-distance weighting spatial interpolation technique,” *Computers & Geoscience*, vol. 34, pp. 1044–1055, 2008.

[20] Y. Chen, X. Shan, X. Jin, T. Yang, F. Dai, and D. Yang, “A comparative study of spatial interpolation methods for determining fishery resources density in the Yellow Sea,” *Acta Oceanologica Sinica*, vol. 35, no. 12, pp. 65–72, 2016.

[21] X. Zong, M. Xu, J. Xu, and X. Lv, “Improvement of the ocean pollutant transport model by using the surface spline interpolation,” *Tellus A: Dynamic Meteorology and Oceanography*, vol. 70, pp. 1–13, 2018.

[22] G. P. Cressman, “An operational objective analysis system,” *Monthly Weather Review*, vol. 87, pp. 367–374, 1959.

[23] S. Chen, C. F. N. Cowan, and P. M. Grant, “Orthogonal least squares learning algorithm for radial basis function networks,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 2, pp. 302–309, 1991.

[24] J. P. C. Kleijnen, “Kriging metamodeling in simulation : a review,” *European Journal of Operational Research*, vol. 192, pp. 707–716, 2009.

[25] Y. C. Fang, T. J. Weingartner, R. A. Potter, P. R. Winsor, and H. Statscewich, “Quality assessment of HF radar-derived surface currents using optimal interpolation,” *Journal of Atmospheric and Oceanic Technology*, vol. 32, pp. 282–296, 2015.

[26] Z. H. Liu, R. G. Huang, Y. M. Hu, S. D. Fan, and P. H. Feng, “Generating high spatiotemporal resolution LAI based on MODIS/GF-1 data and combined Kriging-Cressman interpolation,” *International Journal of Agricultural and Biological Engineering*, vol. 9, pp. 120–131, 2016.

[27] G. Burgers, P. J. Van Leeuwen, and G. Evensen, “Analysis scheme in the ensemble Kalman filter,” *Monthly Weather Review*, vol. 126, pp. 1719–1724, doi. 10.1175/1520-1998)126 2.0.CO;2 1998.

[28] T. M. Smith, R. W. Reynolds, R. E. Livezey, and D. C. Stokes, “Reconstruction of historical Sea surface temperatures using empirical orthogonal functions,” *Journal of Climate*, vol. 9, pp. 1403–1420, 1996.

[29] K. Y. Kim, “Statistical interpolation using cyclostationary EOFs,” *Journal of Climate*, vol. 10, pp. 2931–2942, 1997.

[30] J.-M. Beckers and M. Rixen, “EOF calculations and data filling from Incomplete Oceanographic Datasets,” *Journal of Atmospheric and Oceanic Technology*, vol. 20, pp. 1839–1856, 2003.

[31] C. Jayaram, N. Priyadarshi, J. Pavan Kumar, T. V. S. Udaya Bhaskar, D. Raju, and A. J. Kochuparampil, “Analysis of gap-free chlorophyll-a data from MODIS in Arabian Sea, reconstructed using DINEOF,” *International Journal of Remote Sensing*, vol. 39, pp. 7506–7522, 2018.

[32] A. Alvera-Azcárate, A. Barth, M. Rixen, and J. M. Beckers, “Reconstruction of incomplete oceanographic data sets using empirical orthogonal functions: Application to the Adriatic Sea surface temperature,” *Ocean Model*, vol. 9, pp. 325–346, 2005.

[33] Y. C. Liang, M. R. Mazloff, I. Rosso, S. W. Fang, and J. Y. Yu, “A multivariate empirical orthogonal function method to

construct nitrate maps in the Southern Ocean," *Journal of Atmospheric and Oceanic Technology*, vol. 35, pp. 1505–1519, 2018.

[34] Z. Zhang, X. Yang, H. Li, W. Li, H. Yan, and F. Shi, "Application of a novel hybrid method for spatiotemporal data imputation: a case study of the Minqin County groundwater level," *Journal of Hydrology*, vol. 553, pp. 384–397, 2017.

[35] D. Sirjacobs, A. Alvera-Azcárate, A. Barth et al., "Cloud filling of ocean colour and sea surface temperature remote sensing products over the Southern North Sea by the Data Interpolating Empirical Orthogonal Functions methodology," *Journal of Sea Research*, vol. 65, pp. 114–130, 2011.

[36] H. Pan and X. Lv, "Reconstruction of spatially continuous water levels in the Columbia River estuary: the method of empirical orthogonal function revisited," *Estuarine, Coastal and Shelf Science*, vol. 222, pp. 81–90, 2019.

[37] P. Matte, D. A. Jay, and E. D. Zaron, "Adaptation of classical tidal harmonic analysis to nonstationary tides, with application to river tides," *Journal of Atmospheric and Oceanic Technology*, vol. 30, no. 3, pp. 569–589, 2013.

[38] J. Li and A. D. Heap, "A review of comparative studies of spatial interpolation methods in environmental sciences : performance and impact factors," *Ecological Informatics*, vol. 6, pp. 228–241, 2011.

[39] E. N. Lorenz, *Empirical Orthogonal Functions and Statistical Weather Prediction*, Massachusetts Institute of Technology, Cambridge, MA, USA, 1956.

[40] B. Ping, F. Su, and Y. Meng, "An improved DINEOF algorithm for filling missing values in spatio-temporal sea surface temperature data," *PLoS One*, vol. 11, pp. 1–12, Article ID e0155928, 2016.

[41] W. R. Tobler, "A computer movie simulating urban growth in the Detroit region," *Journal of Economic Geography*, vol. 46, pp. 234–240, 1970.

[42] J. Duchon, "Splines minimizing rotation-invariant seminorms in Sobolev spaces," in *Constructive Theory of Functions of Several Variables*, pp. 85–100, Springer, Berlin, Germany, 1977.

[43] F. L. Bookstein, "Principal Warps : thin-plate splines and the decomposition of deformations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, pp. 567–585, 1989.

[44] Z. Guo, H. Pan, W. Fan, and X. Lv, "Application of surface spline interpolation in inversion of bottom friction coefficients," *Journal of Atmospheric and Oceanic Technology*, vol. 34, pp. 2021–2028, 2017.

[45] J. E. Nash and J. V. Sutcliffe, "River flow forecasting through conceptual models. Part 1 — a discussion of principles," *Journal of Hydrology*, vol. 10, pp. 282–290, 1970.

[46] C. J. Willmott, "On the validation of models," *Progress in Physical Geography*, vol. 2, pp. 184–194, 1981.

[47] X. Wang, R. R. E. Dickinson, L. Su, C. Zhou, and K. Wang, "PM 2.5 pollution in China and how it has been exacerbated by terrain and meteorological conditions," *Bulletin of the American Meteorological Society*, vol. 99, pp. 105–120, 2018.