# Filling the gaps of in-situ hourly $PM_{2.5}$ concentration data with the aid of empirical orthogonal function constrained by diurnal cycles

Kaixu Bai[1,2], Ke Li[2], Jianping Guo[3], Yuanjian Yang[4,5], Ni-Bin Chang[6]

[1]Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University, Shanghai 200241, China

[2]School of Geographic Sciences, East China Normal University, Shanghai 200241, China

[3]State Key Laboratory of Severe Weather, Chinese Academy of Meteorological Sciences, Beijing 100081, China

[4]School of Atmospheric Physics, Nanjing University of Information Science & Technology, Nanjing, China

[5]Institute of Environment, Energy and Sustainability, The Chinese University of Hong Kong, Hong Kong, China

[6]Department of Civil, Environmental, and Construction Engineering, University of Central Florida, Orlando, FL 32816, USA

*Correspondence to*: Dr./Prof. Jianping Guo (jpguocams@gmail.com)

**Abstract.** Data gaps frequently emerge in our retrieved in-situ hourly air quality data records. In this study, we propose a novel gap filling method called the diurnal cycle constrained empirical orthogonal function (DCCEOF) to fill in data gaps for the improvement of data completeness. The hourly $PM_{2.5}$ concentration data retrieved from the China national air quality monitoring network is used here as a demonstration. Generally, the DCCEOF method works in a principle of calibrating the diurnal cycle of $PM_{2.5}$ concentration that is reconstructed from discrete $PM_{2.5}$ neighborhood fields in space and time to the level of valid $PM_{2.5}$ concentration observed at adjacent times. Prior to gap filling, the data completeness and the impact of data gaps in hourly $PM_{2.5}$ concentration record on daily averages were examined. The statistical analysis indicates a high frequency of data gaps in our retrieved hourly $PM_{2.5}$ record, with $PM_{2.5}$ concentration measured on about 40% of days subject to data gaps. On the other hand, these data gaps could introduce significant bias to daily averages of $PM_{2.5}$ concentration, especially during clean episodes as larger biases would be introduced to $PM_{2.5}$ daily averages during clean days than polluted days even in the presence of same number of missingness. The cross-validation results indicate that the DCCEOF method has a good prediction accuracy, particularly in predicting daily peaks and/or minima that cannot be restored by the conventional spline interpolation approach, given the consideration of local diurnal variation pattern of $PM_{2.5}$ in our method. A practical application of the DCCEOF method to the retrieved hourly $PM_{2.5}$ record in China during 2014 to 2019 yields a significant improvement of the data completeness, with the frequency of days with data gaps reduced from 42.6% to 5.7%. In general, the results in this study have well demonstrated the performance and application potential of DCCEOF in handling data gaps in time series of geophysical parameters with significant diurnal variability, and this method can be easily applied to other data sets with similar barriers because of its self-consistent capability.

## 1 Introduction

A large variety of ground-based monitoring networks have been established worldwide to provide accurate measurements on various aspects of the atmospheric environment (Lolli and Di Girolamo, 2015). Many of these in-situ measurements, however, suffer from data losses due to various unexpected reasons, e.g., instrumental malfunction, interruption of power supply, and internet outage, thus resulting in salient data gaps in the archived data records. Undoubtedly, these gaps significantly impair the data qualities and the exploration of these valuable data sources. Therefore, filling data gaps present in such datasets is critical and of great value to facilitating the broad application of these in-situ measurements.

Confronted with frequently occurring severe haze pollution events, China started to establish the national ambient air quality monitoring network since 2012 by extending the range of the previous sparsely distributed monitoring network to cover most major Chinese cities. To date, more than 1,600 state-level stations are routinely operated to measure concentrations of six essential air pollutants (i.e., $PM_{10}$, $PM_{2.5}$, $O_3$, $NO_2$, $SO_2$, CO) on an hourly basis (Guo et al., 2017; Li et al., 2017a). These in-situ measurements are publicly released online via the China National Environment Monitoring Centre (CNEMC) in near real-time as of 2013 but without providing any direct data download interface. Consequently, users oftentimes use an automated software program (often known as a “web crawler”) to retrieve these valuable data sources from the CNEMC website. Such an endeavour helps users to acquire hourly air quality data more efficiently, and the retrieved data record, taken $PM_{2.5}$ mass concentration data as an example, have been widely used as a critical data source in many haze related studies (Gao et al., 2018; Miao et al., 2018; Bai et al., 2019a, 2019b; Zhang et al., 2019).

Although these $PM_{2.5}$ concentration data have been extensively used, how data gaps were treated in the data exploration process (e.g., data integration and data transformation), especially for those using daily or monthly averaged $PM_{2.5}$ data set (e.g., Guo et al., 2009; Miao et al., 2018; Ye et al., 2018; Zhang et al., 2018; Yang et al., 2019a), is oftentimes unclear. Since ignoring missing values would undoubtedly introduce biases into the final results (Bondon, 2005; Larose et al., 2019), some studies attempted to perform data analysis on a relatively long time scale to mitigate the impacts of data gaps by integrating hourly records into monthly resolution  (e.g., Bai et al., 2019b; Zhang et al., 2019). On the other hand, many previous studies preferred to exclude records on days subject to a certain degree of missing values (e.g., no more than 6 missing values within 24-h) from their analysis (e.g., van Donkelaar et al., 2016; Li et al., 2017; Huang et al., 2018; Manning et al., 2018; Shen et al., 2018; Bai et al., 2019a; Zhang et al., 2019). Nevertheless, such a treatment on data gaps (e.g., ignoring missing values or excluding records on days with missingness) would either introduce new bias to the aggregated data record or make the original $PM_{2.5}$ time series temporally discontinuous.

Since a non-gap $PM_{2.5}$ record is essential to $PM_{2.5}$ related haze control and environmental health risk assessment, filling data gaps presented in hourly $PM_{2.5}$ record is thus of critical importance. Although there exists versatile gap filling methods (e.g., Beckers and Rixen, 2003; Taylor et al., 2013; Chang et al., 2015; Dray and Josse, 2015; Gerber et al., 2018), most of them fail to properly restore missingness in $PM_{2.5}$ time series with high temporal resolution (e.g., hourly). In general, the conventional methods are oftentimes incapable of restoring $PM_{2.5}$ daily peaks and/or minima since a priori knowledge of the diurnal variation pattern of $PM_{2.5}$ is always required as $PM_{2.5}$ mass concentration varies significantly in space and time due to heterogeneous local emissions and atmospheric conditions (Guo et al., 2017; Lennartson et al., 2018; Shi et al., 2018). A similar barrier also applies for many other datasets which are sampled at high temporal resolution.

In this study, we propose a novel gap filling method termed as DCCEOF (that is, the diurnal cycle constrained empirical orthogonal function) to better handle data gaps present in time series with marked variability in space and time, by taking the diurnal variation pattern as a critical constraint in missing value prediction. To our knowledge, none of the existing gap filling methods have accounted for the diurnal variation pattern of the given data in their missing value reconstruction schemes, and hence the predicted values from these methods are prone to large bias. As an illustration, the hourly $PM_{2.5}$ concentration record retrieved from CNEMC during the time period of 2014 to 2019 is applied here to demonstrate the efficacy and accuracy of the proposed DCCEOF method. Science questions to be answered by this study include: (1) how about the data completeness of the Chinese in situ $PM_{2.5}$ record? (2) how much uncertainties can be introduced to $PM_{2.5}$ daily averages by missing values? (3) is it feasible to reconstruct the local diurnal variation pattern of $PM_{2.5}$ from discrete observations in the neighborhood? and (4) are missing values restored by DCCEOF reliable?

## 2 Overview of existing gap filling methods

Plenty of methods have been developed or adopted for gap filling with respect to various theoretical bases, ranging from simple replacement with surrogates (e.g., mean value) to spatiotemporal interpolation as well as complicated machine learning techniques. Generally, these methods can be classified into different groups according to different criteria. For instance, two major groups can be classified based on the number of variables (univariate versus multivariate) (Ottosen and Kumar, 2019) and theoretical basis (likelihood-based versus imputation-based) (Junger and Ponce de Leon, 2015). Table 1 summarizes a selection of popular gap filling methods to deal with missingness in geophysical data sets according to the domain specific data dependence (Gerber et al., 2018). Comparisons of the performance of these methods can also be found in other literatures, e.g., Kandasamy et al. (2013), Demirhan and Renwick (2018), Yadav and Roychoudhury (2018), and Julien and Sobrino (2019), to name a few.

Since each method is initially proposed to deal with missingness in one specific data set, adopting one method to another data set is often a challenge due to distinct features of missingness (e.g., missing at random versus missing not at random), in particular for data sets with salient spatiotemporal heterogeneity such as air pollutants time series (Junger and Ponce de Leon, 2015). $PM_{2.5}$ concentration often exhibits evident diurnal variation patterns, which are primarily governed by local air pollutants emissions and regional meteorological conditions such as boundary layer height (Guo et al., 2017; Li et al., 2017; Huang et al., 2018; Liu et al., 2018; Miao et al., 2018; Yang et al., 2018, 2019b). Consequently, conventional approaches like those listed in Table 1 may partially fail in accurately predicting missing values in hourly $PM_{2.5}$ time series.

In general, most available gap filling methods in Table 1 suffer from at least one of the following drawbacks: 1) partially fail for data sets with prominent gaps; 2) not self-consistent due to the requirement of supplementary data sets;3) computationally intensive (e.g., neural networks), and, most critically; 4) unable to fairly predict daily peaks and/or minima due to the lack of essential prior knowledge of diurnal variability of monitoring targets. Given the significant heterogeneity of $PM_{2.5}$ concentration in space and time (Guo et al., 2017; Manning et al., 2018), ignoring the diurnal phases of $PM_{2.5}$ would result in large bias to the gap filled $PM_{2.5}$ data set.

**Table 1.** Overview of several popular gap filling methods to impute missingness in geophysical data sets.

| | Method | Principle or core technique | Reference |
|---|---|---|---|
| Temporal | Weibull | Weibull frequency distribution mapping | Nosal et al. (2000) |
| | EM | Expectation-Maximization | Junger and Ponce de Leon (2015) |
| | Interpolation | Linear regression, Spline, NAR, ARIMA, ARCH | Stauch and Jarvis (2006); Neteler (2010); Demirhan and Renwick (2018) |
| | Machine learning | Gradient Boosting, neural networks | Körner et al. (2018); Şahin et al. (2011) |
| | SSA | Imputation using singular spectrum analysis | Mahmoudvand and Rodrigues (2016) |
| | DS | Conditional resampling of a temporal subset | Dembélé et al. (2019); Oriani et al. (2016) |
| | TIMESAT | Savitzky–Golay filter, harmonic and asymmetric Gaussian functions | Jönsson and Eklundh (2004) |
| | Hybrid method | Fuzzy c-means with support vector regression and genetic algorithm | Aydilek and Arslan (2013) |
| Spatial | IDW | Interpolate using inverse distance weighting | Shareef et al. (2016) |
| | Kriging | Interpolate neighborhoods using Kriging | Rossi et al. (1994); Zhu et al. (2015); Singh et al. (2017) |
| | NSPI / GNSPI | Replace or interpolate with adjacent similar pixels | Zhu et al. (2012); Chen et al. (2011) |
| Spatio-temporal | EOF / DINEOF | Iteratively decompose and reconstruct spatial and temporal subsets using empirical orthogonal function | Beckers and Rixen (2003); Taylor et al. (2013); Liu and Wang (2019) |
| | Mosaicing | Merge numerical outputs with satellite observations | Konik et al. (2019) |
| | gapfill | Quantile regression fitted to spatiotemporal subsets | Gerber et al. (2018) |
| | STWR | Spatially and temporally weighted regression | Chen et al. (2017) |
| | SMIR | Learning machine created from historical spatial and temporal subsets | Chang et al. (2015) |
| | RFRE | Learning from other information using random forest | Bi et al. (2018); Chen et al. (2019) |

* SSA: Singular Spectrum Analysis; DS: Direct Sampling; IDW: Inverse Distance Weighting; NSPI: Neighborhood Similar Pixel Interpolator; GNSPI: Geo-statistical Neighborhood Similar Pixel Interpolator; EOF: Empirical Orthogonal Function; DINEOF: Data Interpolating Empirical Orthogonal Function; STWR: Spatially and Temporally Weighted Regression; SMIR: SMart Information Reconstruction; RFRE: Random Forest Regression

## 3 The DCCEOF gap filling method

Given the significant heterogeneity of $PM_{2.5}$ diurnal variation pattern associated with local emissions of air pollutants and atmospheric conditions, we propose to apply the local diurnal variation pattern of $PM_{2.5}$ to constrain the reconstruction of missing values in the hourly time series of $PM_{2.5}$ concentration at each station. The goal is to better predict missing $PM_{2.5}$ values, especially for the daily peaks and/or minima, which are poorly predicted by conventional methods due to the absence of prior knowledge of local diurnal phases of $PM_{2.5}$. Figure 1 presents a schematic illustration of the proposed DCCEOF method. In general, the DCCEOF method consists of the following four primary procedures toward the filling of data gaps present in each 24-h $PM_{2.5}$ time series:

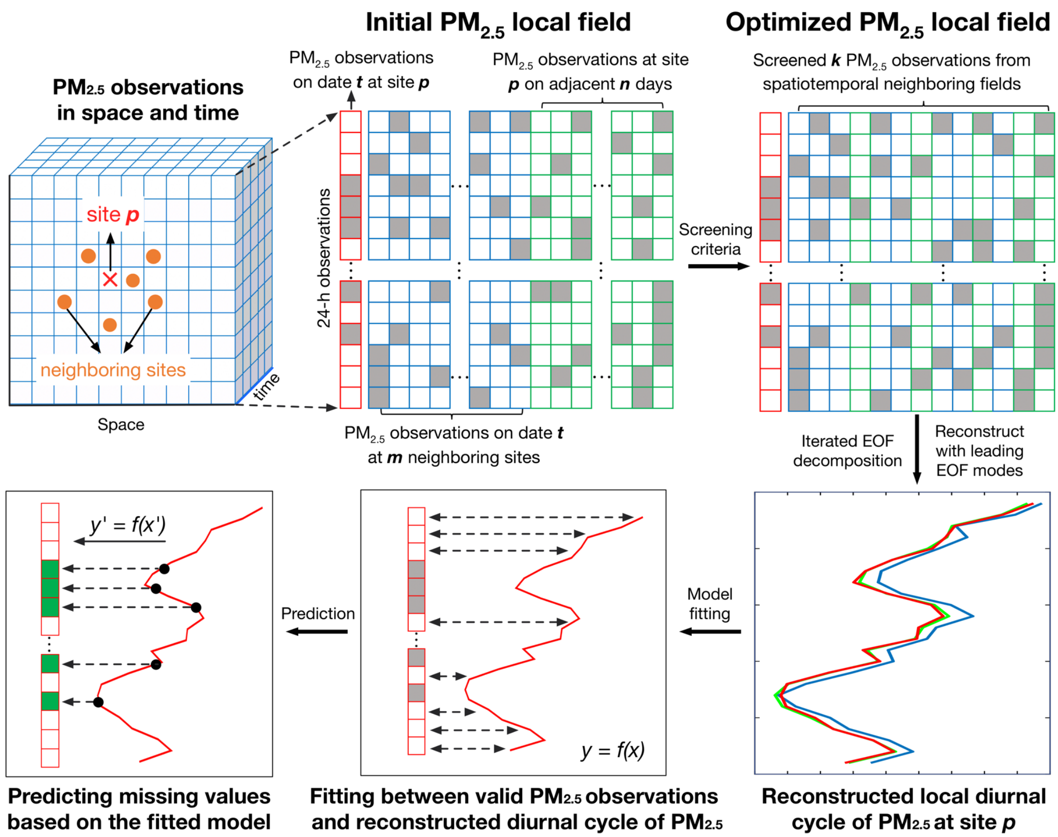


**Figure 1.** A schematic illustration of the proposed DCCEOF method to fill data gaps in hourly $PM_{2.5}$ records. The grey rectangles denote missing values while the green ones indicate reconstructed data values.

1) Initialize a local $PM_{2.5}$ neighborhood field: For any identified $PM_{2.5}$ missingness at site $\boldsymbol{p}$ on date $\boldsymbol{t}$, an initial $PM_{2.5}$ neighborhood field in space and time (denoted as $\boldsymbol{X_{p,t}^{m,n}}$) is first constructed using 24-h

$PM_{2.5}$ observations from nearby $\boldsymbol{m}$ stations on date $\boldsymbol{t}$ and adjacent $2\boldsymbol{n}$ days ($\boldsymbol{n}$ days before and after $\boldsymbol{t}$ respectively) at site $\boldsymbol{p}$. Mathematically, the neighborhood field $\boldsymbol{X_{p,t}^{m,n}}$ can be expressed as:

$$X_{p,t}^{m,n} = \left\{x_t^1, x_t^2, \dots, x_t^m, x_p^{t-n}, \dots, x_p^{t-2}, x_p^{t-1}, x_p^{t+1}, x_p^{t+2}, \dots, x_p^{t+n}\right\} \quad (1)$$

It is clear that $\boldsymbol{m}$ and $\boldsymbol{n}$ are two critical factors in determining the dimension of $\boldsymbol{X_{p,t}^{m,n}}$. Considering a too compact neighborhood field may be inadequate to reconstruct the local diurnal cycle of $PM_{2.5}$ fairly due to limited samples, because missingness may also present in each candidate 24–h $PM_{2.5}$ concentration time series, here $\boldsymbol{m}$ is defined as the number of stations within 100 km (spatial window size) to the target station while $\boldsymbol{n}$ is set to 7 (temporal window size) in our algorithm. The spatial and temporal window sizes used here is based on our recent results in which an optimal window size of 50 km and 3-day was found to attain a good autocorrelation of $PM_{2.5}$ concentration in space and time, respectively (Bai et al., 2019c). To have adequate samples for the construction of $\boldsymbol{X_{p,t}^{m,n}}$, here we enlarge the both window sizes by simply doubling the threshold found in our previous study. In general, these two window sizes would have little effect on the performance of the subsequent gap filling once they are large enough (at least greater than the identified optimal window sizes) to cover most similar observations nearby since a sorting scheme will be further applied to identify observations with similar diurnal variation pattern to that of the target station. In other words, the two window sizes used here is simply to include adequate samples while avoiding incorporating all available data for the subsequent data reconstruction, especially for those distant away.

2) Construct a compact $PM_{2.5}$ neighborhood field: Since the initial $PM_{2.5}$ neighborhood field $\boldsymbol{X_{p,t}^{m,n}}$ might include many irrelevant observations with distinct diurnal variation patterns given large spatial and temporal window sizes, a compact neighborhood field needs to be constructed by only retaining observations that are highly related to the target $PM_{2.5}$ time series $\boldsymbol{x_p^t}$ with respect to the diurnal variation pattern. Therefore, the covariance rather than correlation between the target time series $\boldsymbol{x_p^t}$ and every candidate $PM_{2.5}$ time series in $\boldsymbol{X_{p,t}^{m,n}}$ is first calculated (weighted by the number of valid data pairs within 24-h). Subsequently, the candidate $PM_{2.5}$ time series are sorted in terms of the magnitudes of covariances in a descending order. Finally, the first $\boldsymbol{k}$ time series are retained to construct the optimized $PM_{2.5}$ neighborhood field $\widehat{\boldsymbol{X^k}}$ by complying with the criterion that there are at least five valid observations at

each specific time from 00:00 to 23:00. The aim of this configuration is to avoid large bias in the subsequent diurnal cycle reconstruction using empirical orthogonal function (EOF), since large outliers may emerge at times without any valid observation. Mathematically, the process to construct $\widehat{\boldsymbol{X}^{\boldsymbol{k}}}$ can be formulated as follows:

$$C_{x'} = COV\left(x_p^t, x' \middle| X_{p,t}^{m,n}\right) \tag{2}$$

$$\widehat{X^k} = \left\{x'_1, x'_2, \dots, x'_k \middle| C_{x'_k} < C_{x'_{k-1}} < \cdots < C_{x'_1}\right\} \tag{3}$$

where $x'$ denotes the 24-h time series of candidate $PM_{2.5}$ in $\boldsymbol{X}_{\boldsymbol{p},\boldsymbol{t}}^{\boldsymbol{m},\boldsymbol{n}}$ and *COV* is the covariance function.

3) Reconstruct the local diurnal cycle of $PM_{2.5}$: The diurnal cycle of $PM_{2.5}$ at site $\boldsymbol{p}$ on date $\boldsymbol{t}$ (denoted as $\boldsymbol{\beta}_{\boldsymbol{p}}^{\boldsymbol{t}}$) was then reconstructed from the optimized $PM_{2.5}$ neighborhood field $\widehat{\boldsymbol{X}^{\boldsymbol{k}}}$ using EOF in an iterative process similar to the DINEOF method (Beckers and Rixen, 2003). In our DCCEOF method, the target $PM_{2.5}$ time series at site $\boldsymbol{p}$ on date $\boldsymbol{t}$ (denoted as $\boldsymbol{x}_{\boldsymbol{p}}^{\boldsymbol{t}}$) were also included to constrain the reconstruction of $\boldsymbol{\beta}_{\boldsymbol{p}}^{\boldsymbol{t}}$, and the whole field was then denoted as $\widetilde{\boldsymbol{X}}$.

$$\tilde{X} = \left\{x_p^t, \widehat{X^k}\right\} \tag{4}$$

In general, the EOF-based gap filling process can be outlined as follows: a) 20% of valid $PM_{2.5}$ observations in $\widetilde{\boldsymbol{X}}$ were first held out for cross validation and then these data values were treated as gaps by replacing with nulls (i.e., missing value); b) given that a small amount of missing values would not significantly influence the leading EOF mode for the original data set, we may assign a first guess (here we used the mean value of valid data on each specific date) to the data points where missing values are identified to initialize the EOF analysis; c) EOF analysis was performed on the previously generated background field (that is, $\widetilde{\boldsymbol{X}}$ with gaps are filled with daily mean and denoted as $< \tilde{X} >$) in a form of singular value decomposition (SVD) and then data values at value-missing points were replaced by the reconstructed values using the first EOF mode. These processes can be expressed as:

$$[U, S, V] = svd\left(< \tilde{X} >\right) \tag{5}$$

$$X' = u_1 * s_1 * v_1 \tag{6}$$

where $< \tilde{X} >$ denotes the initial matrix in which the missing values were filled with daily means. *U*, *S*, and *V* are three matrices derived from SVD while $u_1$, $s_1$, and $v_1$ denote the SVD components in the first

EOF mode. $X'$ is the reconstructed matrix using the first EOF mode; e) iteratively decompose and reconstruct the matrix while updating data values at the value-missing points using the first EOF mode till the convergence is confirmed by the mean square error at each iteration; f) repeat the above iterative processes for the following EOF modes till the reach of the final convergence (i.e., error starts to increase as the new EOF mode is included). The $\boldsymbol{\beta}_{\boldsymbol{p}}^{\boldsymbol{t}}$ was finally derived by standardizing the identified leading EOF modes.

4) Missing value prediction: A linear relationship was finally established between valid $PM_{2.5}$ observations in $\boldsymbol{x}_{\boldsymbol{p}}^{\boldsymbol{t}}$ and the corresponding values in $\boldsymbol{\beta}_{\boldsymbol{p}}^{\boldsymbol{t}}$. Missing values in the time series of the original $PM_{2.5}$ ($\boldsymbol{x}_{\boldsymbol{p}}^{\boldsymbol{t}}$) were then predicted by mapping data values in the reconstructed diurnal cycle at missing time to the level of valid $PM_{2.5}$ observations based on the established linear relationship.

In short, the proposed DCCEOF method is a univariate and self-consistent gap filling method since no additional data record is required for missing value prediction. Rather, the method works relying primarily on the local diurnal cycle of $PM_{2.5}$ that can be reconstructed from discrete $PM_{2.5}$ neighborhood fields in space and time. In contrast to conventional gap filling methods that work on a purely statistical basis (e.g., spline interpolation), the unique feature and novelty of the proposed DCCEOF method lies in the accounting for the diurnal variation pattern in the prediction of missing values, making the predicted values with high accuracy and physically meaningful.

## 4 Demonstrative case study in China

### 4.1 China in-situ $PM_{2.5}$ concentration records

The near surface mass concentration of $PM_{2.5}$ across China are measured primarily using the tapered element oscillating microbalance analyzer and/or the beta-attenuation monitor at each monitoring station. The instruments' calibration, operation, maintenance, and quality control are all properly conducted by complying with the China Environmental Protection Standards of GB3095-2012 and HJ 618–2011. $PM_{2.5}$ concentration data are measured by these instruments with an accuracy of ±5 μg/m$^3$ for ten-minute averages and ±1.5 μg/m$^3$ for hourly averages (Guo et al., 2017; Miao et al., 2018). Although the hourly $PM_{2.5}$ observations in China have been publicly available since 2013, the $PM_{2.5}$ records used in the present study were retrieved since May 2014 via a web crawler program.

Figure 2 depicts the spatial distribution of monitors in the China national ambient air quality monitoring network as well as the start year for the first release of $PM_{2.5}$ measurements at each individual station. Given the fact that our data were retrieved following May 2014, stations deployed before that are hardly to be separated from those being built in 2014 and hence, they were all designated the same way in Figure 2. At present, this network consists of more than 1,600 stations, in which about 940 stations were established before 2015. The total number of stations was increased to 1,494 in June 2015, and then only four stations were newly deployed in the following one and half years till December 2016. In other words, the vast majority (92.4%) of air quality monitoring stations in the current network was deployed before the middle of 2015.

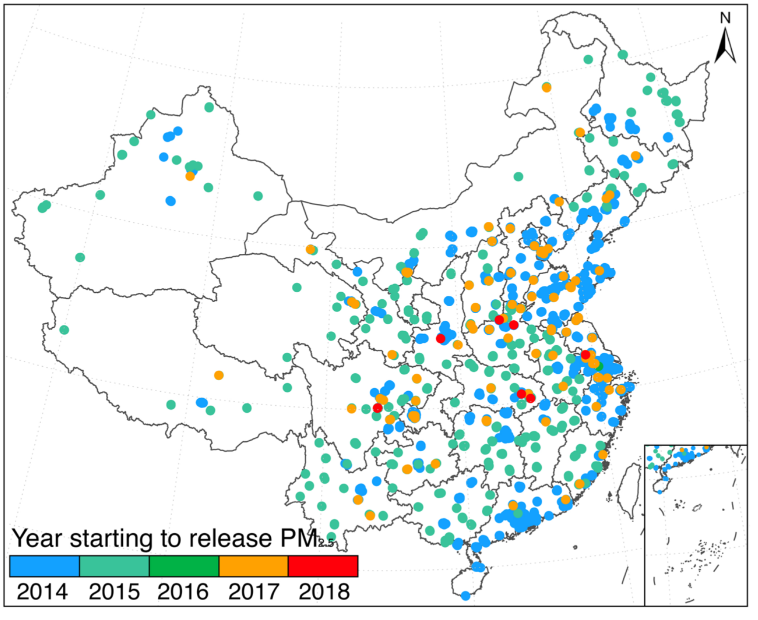


**Figure 2.** Spatial distribution of national ambient air quality monitoring stations in China during May 2014 to April 2019. Circles with distinct color indicate the year in which the first $PM_{2.5}$ observation was publicly released at each station in our retrieved data record.

## 4.2 Results

### 4.2.1 Data completeness of in-situ $PM_{2.5}$ records in China

Figures 3a–c present the daily averaged missing value ratio, the occurrence frequency of missingness (defined as the ratio of days with missing values in each 24-hour $PM_{2.5}$ observations divided by the total number of days), and the diurnal phases of the most frequently occurring missing values at each monitoring station since the first release of $PM_{2.5}$ observations to the public, while Figures 3d–f show the corresponding histograms, respectively. Although most of stations have a daily-averaged missing value ratio less than 10% (Figures 3a and 3d), significant data gaps are still observed at several monitoring stations (red dots in Figure 3a) with more than 70% of hourly $PM_{2.5}$ observations lost in daily 24-h measurements. After checking the retrieved $PM_{2.5}$ data records over these stations, we find that most of these stations stopped releasing $PM_{2.5}$ observations after the middle of 2015.

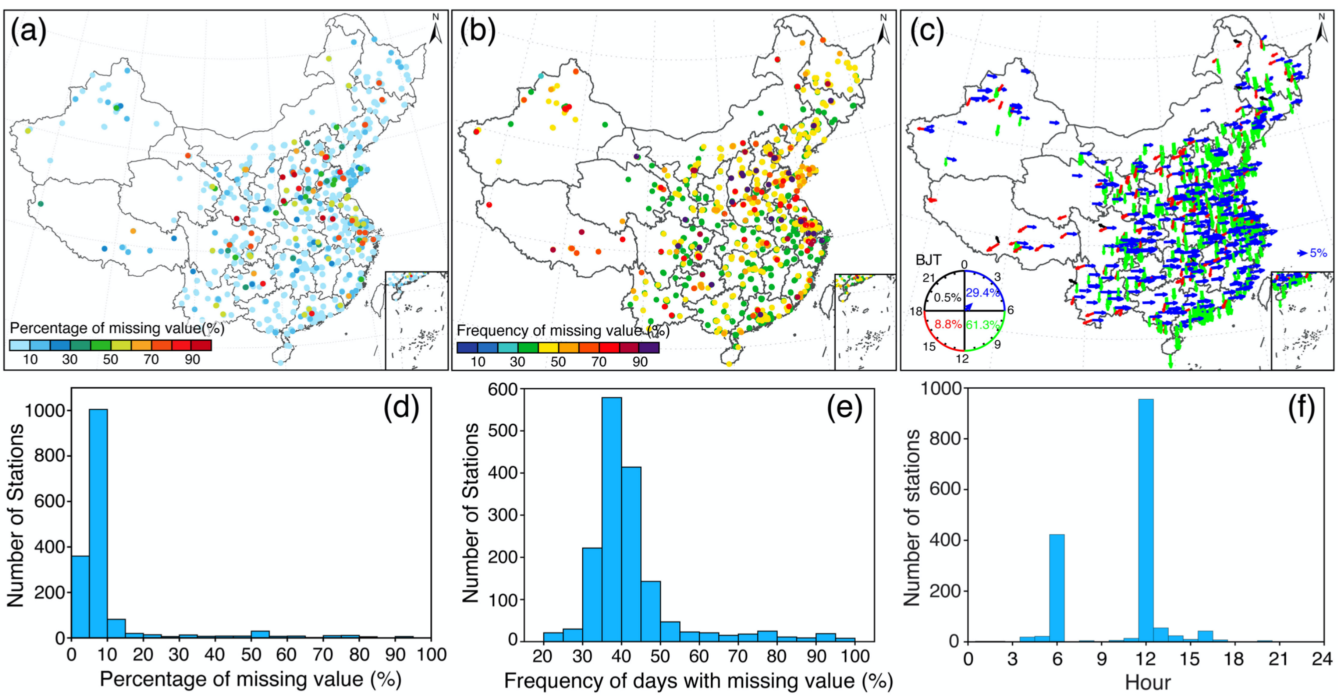


**Figure 3.** Statistics associated with missing values present in site specific hourly $PM_{2.5}$ record since the first release of $PM_{2.5}$ observations onward. (a) Percentage of missingness in each $PM_{2.5}$ record, (b) frequency of days with missing values, (c) diurnal phases of the maximum frequency of missing values occurred within 24-h, (d–f) histograms for (a–c), respectively. The arrow direction in (c) denotes the local

time (Beijing time, BJT) at which missing values occurred most frequently and the arrow length indicates the magnitude of frequency. The varying diurnal phases of missing values were represented by different color: blue (00~06 BJT), green (06~12 BJT), red (12~18 BJT), and black (18~24 BJT).

Despite the small magnitudes (~10%) of daily-averaged missing value ratios (Figure 3d), data gaps in our retrieved hourly $PM_{2.5}$ record are still significant, which is evidenced by the occurrence frequency of missing values in daily $PM_{2.5}$ observations (Figure 3b). In contrast to the daily averaged missing value ratios (Figure 3a), the missing value frequency has a relatively larger magnitude of about 40%, indicating a high occurrence frequency of data gaps in the retrieved $PM_{2.5}$ record, as $PM_{2.5}$ data measured on four out of ten days were subject to missingness (Figure 3e). These results suggest an urgent need to fill in the data gaps present in our retrieved $PM_{2.5}$ record so as to facilitate the further exploration of this valuable data set.

Figure 3c presents the diurnal variation pattern of the occurrence of missingness in the retrieved $PM_{2.5}$ record in terms of the detailed time (represented by the arrow direction) and frequency (represented by the relative length of each arrow) of the most frequently occurring missing values, while Figure 3f shows the histogram of the local time at which missing values occurred most frequently at each monitoring station. It is interesting to note that the missing values occurred more frequently in the morning over most stations (90.7% of total population of stations), particularly at 0600 and 1200 of the Beijing time. However, detailed reason for this diurnal variation pattern remains unclear.

#### 4.2.2 Impacts of data gaps on $PM_{2.5}$ daily averages

Given the frequent usage of daily-averaged $PM_{2.5}$ concentration data in many studies, the possible impacts of data gaps on $PM_{2.5}$ daily averages were thus assessed here to examine how well the estimated $PM_{2.5}$ daily averages can be trusted in the presence of data gaps, especially in different pollution episodes. Toward such a goal, gap-free observations of hourly $PM_{2.5}$ within 24-h were first extracted. To make the computational workload manageable, we randomly sampled 1,000 days observations rather than using observations from all gap-free days. Moreover, days with $PM_{2.5}$ daily averages lower than that of the 10th quantile of all gap-free days were considered as clean scenario, while those greater than the 90th quantile were treated as polluted scenario. Subsequently, a varying number (range from 1 to 23) of data values were treated as gaps in every 24-h $PM_{2.5}$ observations randomly and then mean relative differences (MRDs) between $PM_{2.5}$ daily averages derived from hourly records with and without data gaps were calculated to evaluate the potential impacts of missingness.

Figure 4a shows the estimated MRDs at the 10th, 50th, and 90th quantiles associated with different numbers of missing values in each 24-h $PM_{2.5}$ observations, indicating that large biases could be introduced to the daily averages with the increase in the total number of missingness. Given the symmetrical behavior of MRDs around zero (like a Gaussian distribution) for each given number of missingness, we may infer that random biases could be introduced in $PM_{2.5}$ daily averages if missing values are ignored for the calculation of daily averages of $PM_{2.5}$. These random biases, in turn, could yield large uncertainties to the subsequent results such as trend estimations. To further evaluate the impacts of missingness on $PM_{2.5}$ daily averages, in particular at different pollution scenarios, MRDs were also calculated on 1,000 clean and polluted days, respectively (Figures 4b–d). On average, MRDs vary with larger deviations for a given number of missingness on clean days than on polluted days (Figure 4b). Regarding MRDs at 10th and 90th quantiles, we may deduce that missing values would result in larger bias to $PM_{2.5}$ daily averages on clean days than in polluted conditions given larger MRDs for clean scenarios (Figures 4c–d). This effect is in line with expectations since $PM_{2.5}$ concentration often exhibits relatively larger diurnal variations on cleaner days than during polluted episodes due to the possible boundary layer height effect (Li et al., 2017; Miao et al., 2018). Moreover, six missing values in 24-h observations would result in as large as approximately 5% of deviations (10% for 12 missing values) to $PM_{2.5}$ daily averages during clean days (Figures 4c–d).

In addition to the number of missing values, possible impacts of diurnal phases of missing values on $PM_{2.5}$ daily averages were also examined. It shows that different diurnal phases were observed for MRDs associated with missingness at different pollution levels (Figure 5). Specifically, missing values in the afternoon and evening would more likely result in overestimations to $PM_{2.5}$ daily averages, whereas an opposite effect (underestimations) was observed for missingness in the morning and night. Moreover, the missingness in the afternoon during clean days has a larger potential to overestimate $PM_{2.5}$ daily averages than during other times. This effect could be largely associated with the diurnal phases of $PM_{2.5}$ as daily peaks are oftentimes observed in the early morning (Wang and Christopher, 2003), though such a diurnal variation pattern may differ by regions (Lennartson et al., 2018). Also, the diurnal phases of $PM_{2.5}$ are largely dominated by the diurnal variation of regional emissions and boundary layer processes (Guo et al., 2016; Lennartson et al., 2018; Miao et al., 2018; Yang et al., 2019b). In contrast, the diurnal phases of MRDs are not evident during polluted days. All these findings collectively suggest the need to fill in data gaps present in hourly $PM_{2.5}$ observations, especially for those measured during clean days, since missing values would result in larger biases to $PM_{2.5}$ daily averages than those during polluted episodes.

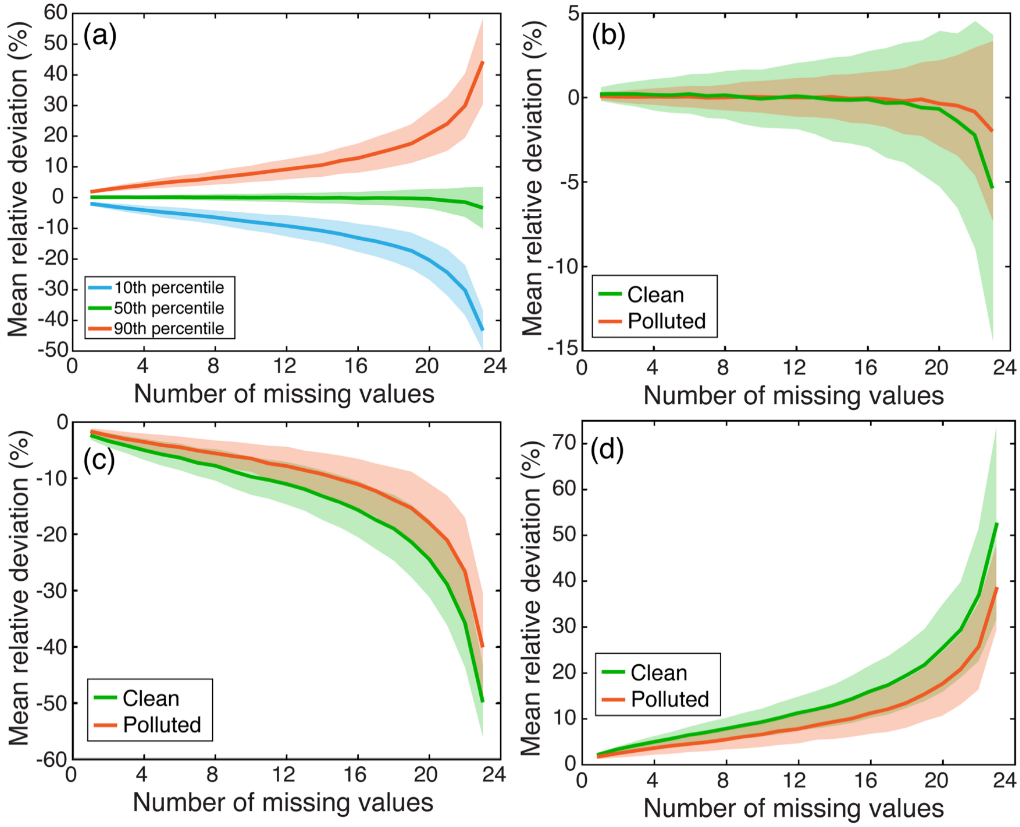


**Figure 4.** Impacts of the number of missing values on daily averages of $PM_{2.5}$. Mean relative deviations were calculated between $PM_{2.5}$ daily averages estimated from 1,000 hourly $PM_{2.5}$ records with a given number of missing values and the original one without missing values. (a) Deviations at different percentiles at all-sky conditions; (b) deviations at the 50th percentile under different pollution scenarios; (c) same as (b) but for the 10th percentile; (d) same as (b) but for the 90th percentile. Thick lines represent

mean deviations while shaded regions are uncertainties of one standard deviation from the mean at each side.

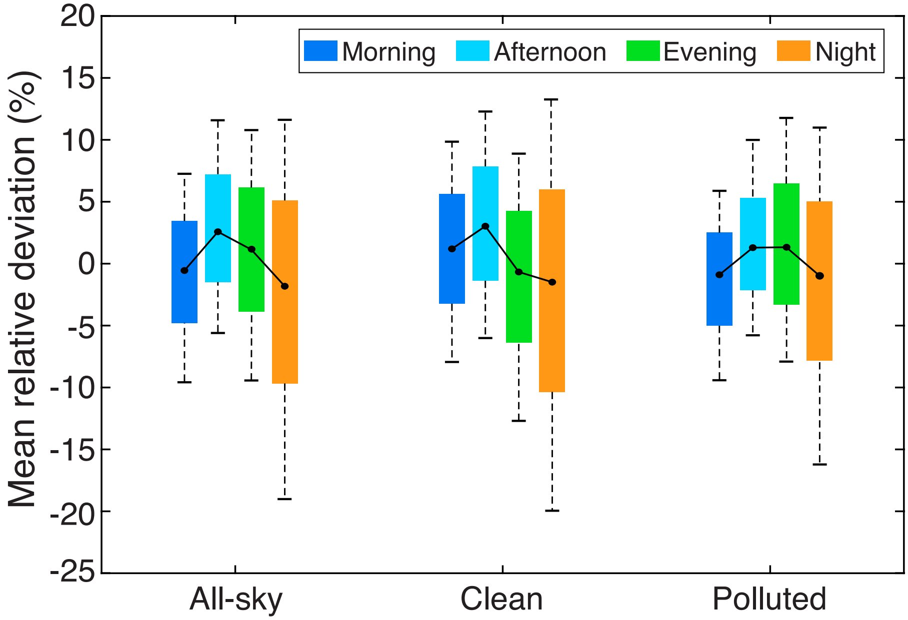


**Figure 5.** Impacts of diurnal phases of missing values on $PM_{2.5}$ daily averages. Hourly $PM_{2.5}$ values in the morning (07~11 BJT), afternoon (12~16 BJT), evening (17~21 BJT), and night (22~06 BJT) were removed from the original hourly $PM_{2.5}$ time series throughout the day to resemble missing values respectively. On each box, the black dots represent medians of mean relative deviations while the bottom and top edges of the box indicate the 25th and 75th percentiles and the whiskers extend to the 10th and 90th percentiles, respectively.

### 4.2.3 Performance of the DCCEOF method

To assess the efficacy and accuracy of the proposed DCCEOF method, cross validation experiments were conducted at two different monitoring stations. Specifically, three gap-free $PM_{2.5}$ records within 24-h during different pollution episodes were first extract randomly at each station and then six valid observations in each 24-h record were held out. Subsequently, the DCCEOF method was applied to reconstruct the diurnal cycle of $PM_{2.5}$ for each specific case. Figure 6 compares the reconstructed diurnal cycles of $PM_{2.5}$ with their actual $PM_{2.5}$ concentrations. The results indicate that the reconstructed diurnal cycles of $PM_{2.5}$ have a good fit with their actual observations, thus confirming the robustness of the

DCCEOF method in reconstructing the diurnal variation pattern of $PM_{2.5}$ from the discrete neighborhood field. In particular, the DCCEOF method also succeeded to restore the missing $PM_{2.5}$ information even at the inflection times, e.g., the peak value in Figure 6c and the minimum value in Figure 6e, which are oftentimes hardly to be recovered by statistical interpolation approaches. Nonetheless, compared with actual $PM_{2.5}$ observations, the reconstructed $PM_{2.5}$ diurnal cycle is still unable to totally restore all types of local variations (e.g., $PM_{2.5}$ observations between 0700 and 1100 shown in Figure 6f). This is consistent with our initial understanding that $PM_{2.5}$ concentrations vary significantly in space and time whereas the reconstructed $PM_{2.5}$ diurnal cycle is derived from a limited number of leading EOF modes and hence it only captures the dominant variation pattern of the neighborhood field while some local variations could be ignored. In spite of this potential defect, the proposed DCCEOF method still exhibits promising accuracy in restoring the local $PM_{2.5}$ diurnal cycle from a discrete neighborhood field.

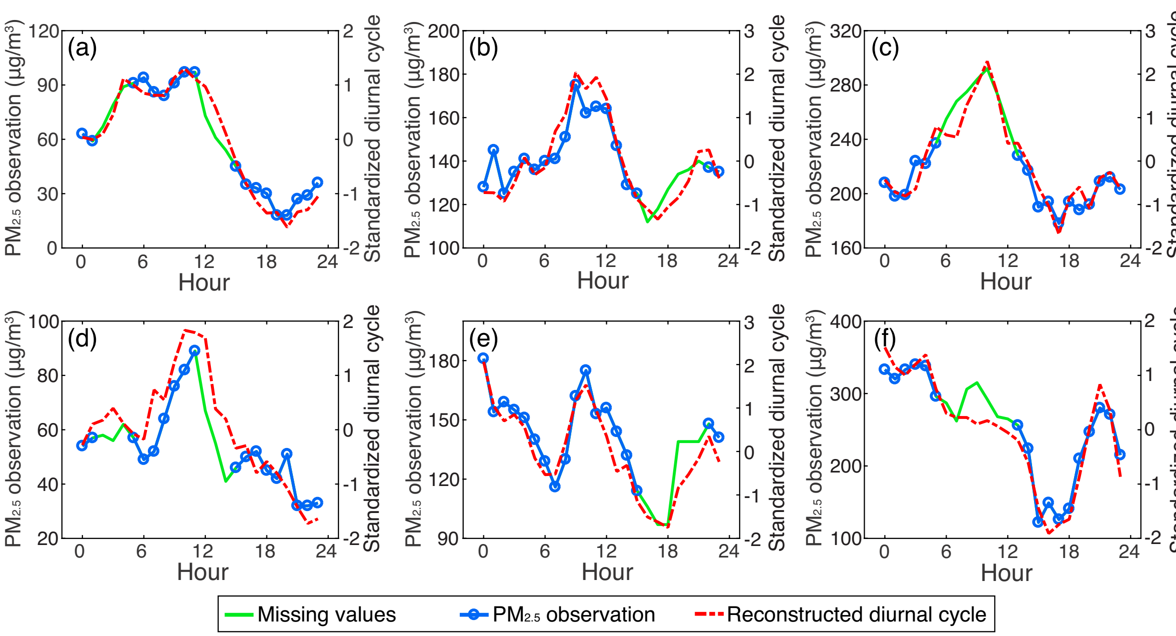


**Figure 6.** Comparisons between actual $PM_{2.5}$ concentrations and the reconstructed $PM_{2.5}$ diurnal cycles at different pollution levels. For each trial, 6 valid $PM_{2.5}$ observations were treated as missing values to simulate gapped $PM_{2.5}$ time series prior to diurnal cycle reconstruction for a given day. Note that the

number of neighboring stations differs between these two cases (58 for the top panel and 16 for the bottom).

To better assess the performance of the DCCEOF method, we retrieved the hourly $PM_{2.5}$ observations recorded at one monitoring station in Beijing during the time period August 1 to 7, 2014 and then some valid observations were treated as missing values for the subsequent gap filling practices. Both the DCCEOF method and a spline interpolation approach were then used to practically restore the retained missing values. The comparison results shown in Figure 7 indicate higher accuracy of the DCCEOF method than the spline interpolation approach in restoring the artificially masked missing values, especially for those at the inflection times at which spline interpolation failed to predict with good accuracy (e.g., peak values on August 3). However, both methods failed in predicting the minimum values on August 2. After checking the original data record, we found that the local variation of $PM_{2.5}$ at this station differed largely from that of all neighboring stations at that time. For such situation, the proposed DCCEOF method would fail to properly predict the missing values given the distinct diurnal variation pattern from that of neighbors in space and time.

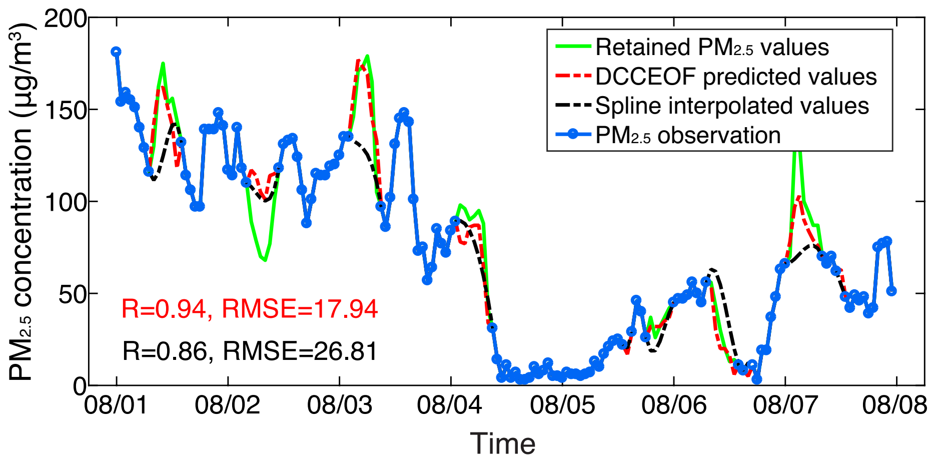


**Figure 7.** Comparison of gap filled hourly $PM_{2.5}$ time series reconstructed using spline interpolation and the proposed diurnal cycle prescribed gap filling method at the Wanshou Temple station in Beijing

between 1 and 7 August 2014. The green line shows the practical $PM_{2.5}$ observations that were treated as gaps while their original values were retained for cross validation.

Figure 8 presents a more general evaluation of the prediction accuracy of the proposed DCCEOF method, which compares the predicted values with the retained data values at different pollution levels. As indicated, there is a good fit between the predicted values and the actual observations, with a correlation coefficient of 0.82 on clean days (Figure 8a) and 0.95 during polluted episodes (Figure 8b), respectively. This is in line with our expectation as higher prediction accuracy would be reached by the DCCEOF method in filling data gaps on polluted days given smaller variability of $PM_{2.5}$ concentrations. This effect can also be evidenced by spread scatters shown in Figure 8a, which in turn reveals the large spatiotemporal heterogeneity of $PM_{2.5}$ concentrations during clean scenarios.

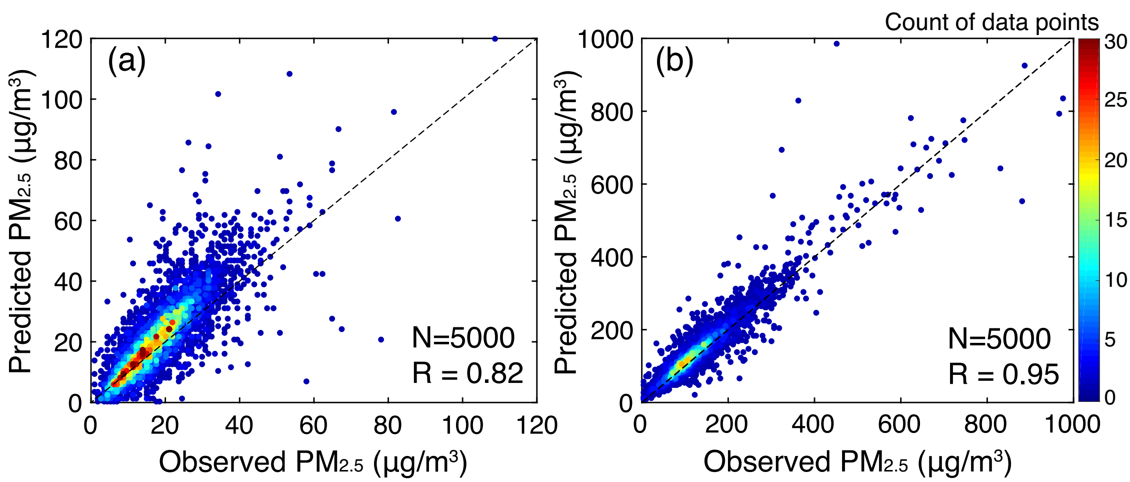


**Figure 8.** Comparisons of $PM_{2.5}$ observations with the reconstructed data values during clean (a) and polluted (b) phases. For each scenario, the results were derived from 1,000 days of gap-free $PM_{2.5}$ observations with 5 valid values being randomly retained from 24-h observations on each sampled date for cross validation.

Given the underlying principle of utilizing discrete neighborhood field in space and time to reconstruct the local diurnal cycle of $PM_{2.5}$ for the subsequent missing value restoration, the performance of the DCCEOF method could be subject to the number of missing values and the total number of neighboring stations. To assess the possible dependence of prediction accuracy on these two factors, sensitivity experiments were also conducted. Figure 9a shows the response of prediction accuracy (in terms of correlation coefficient) of the DCCEOF method to the varying number of missing values in each sampled 24-h $PM_{2.5}$ time series. It clearly shows that the prediction accuracy generally decreases with the increase in the number of missing values. This effect can be ascribed to the fact that the target $PM_{2.5}$ time series is also applied as a critical constraint for the screening of similar $PM_{2.5}$ observations in space and time to construct the neighborhood field for the reconstruction of local diurnal cycle of $PM_{2.5}$. Consequently, more missingness would make the constructed neighborhood field have larger uncertainties due to less information for the selection of related time series of $PM_{2.5}$, which in turn undermines the overall accuracy of the final predictions.

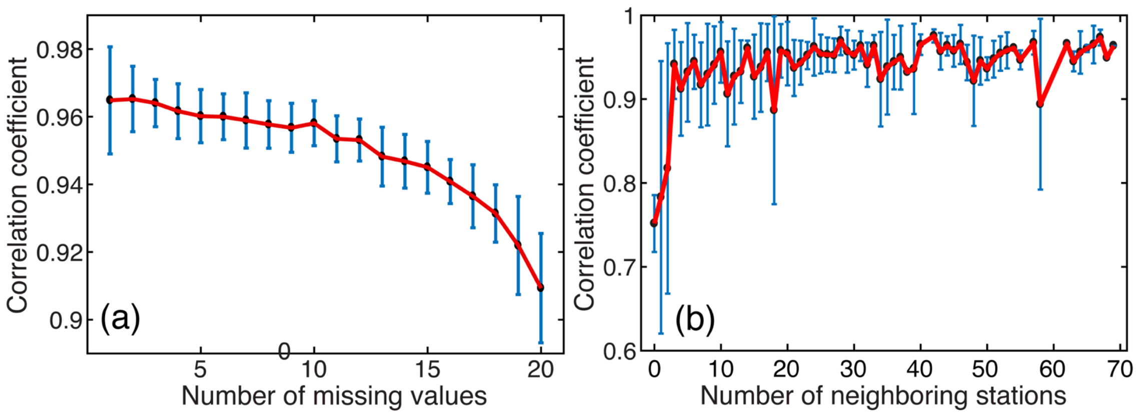


**Figure 9.** Impacts of the number of missing values present in hourly $PM_{2.5}$ records for every 24-h (a) and the total number of neighboring stations within 100 km (b) on the performance of the proposed gap filling method. The error bars denote one standard deviation of each value from the mean on each side.

Figure 9b shows the potential impacts of the total number of neighboring stations on the prediction accuracy at the target station. The total number of neighboring stations within a radius of 100 km to the target station was first calculated and then sensitivity experiments were performed for each specific number. Specifically, ten stations were randomly selected for each given number, and then 20 days gap-free $PM_{2.5}$ observations were sampled at each individual station. For each gap-free $PM_{2.5}$ observation within 24-h, six values were retained and then treated as gaps for cross validation while the DCCEOF method was finally applied to restore these values. It is indicative that the DCCEOF method would have high prediction accuracy with an adequate number of neighboring stations, as three neighboring stations suffice to yield promising prediction accuracy (Figure 9b). On the other hand, large biases could be introduced to the final predictions with a limited number of neighboring stations (<3) due to the lack of sufficient prior spatial information for the reconstruction of the diurnal cycle of $PM_{2.5}$. Nevertheless, good accuracy still can be guaranteed even in the absence of prior spatial information (that is, no neighboring station within 100 km), which in turn corroborates the beneficial effect of the inclusion of temporal neighborhood in gap filling. Although the prediction accuracy improves with the increase in the number of neighboring stations, the gains of accuracy is not prominent at stations with more than three neighboring stations. This is because we only use the similar observations rather than all available observations within 100 km to reconstruct the diurnal cycle of $PM_{2.5}$; otherwise, irrelevant observations would distort the reconstructed diurnal variation pattern and in turn the final predictions. On the other hand, the increase in the number of neighboring stations would reduce the uncertainties in the final predicted values, which is evidenced by smaller standard deviations of correlation coefficients for those with more neighboring stations (Figure 9b). Moreover, the diurnal cycle reconstructed from the neighborhood field in space is more accurate than using $PM_{2.5}$ observations from near-term days, which is evidenced by smaller correlation values with limited neighboring stations. Such effect is also in line with our recent results when comparing the beneficial effects of spatial and temporal neighboring terms in advancing the gridded $PM_{2.5}$ concentration mapping (Bai et al., 2019c).

Figure 10 shows the benefits of the DCCEOF method on our retrieved in–situ hourly $PM_{2.5}$ record at each individual monitoring station in terms of the improvement of data completeness ratio as well as the reduction of gap frequency. After applying the DCCEOF method, the data completeness ratio of hourly $PM_{2.5}$ concentration records in China has been improved by approximately 5% on average nationwide, with the overall data completeness ratio increasing from 89.2% to 94.3% (Figure 10a). Despite the small magnitude of data completeness improvement ratio, the occurrence frequency of missingness has been significantly reduced, with the averaged frequency of days with missingness declined from 42.6% to 5.7% nationwide (Figure 10b). In general, the gap-filled $PM_{2.5}$ record is temporally more complete given fewer data gaps and this data set can thus be used as a promising data source for $PM_{2.5}$-related studies in the future.

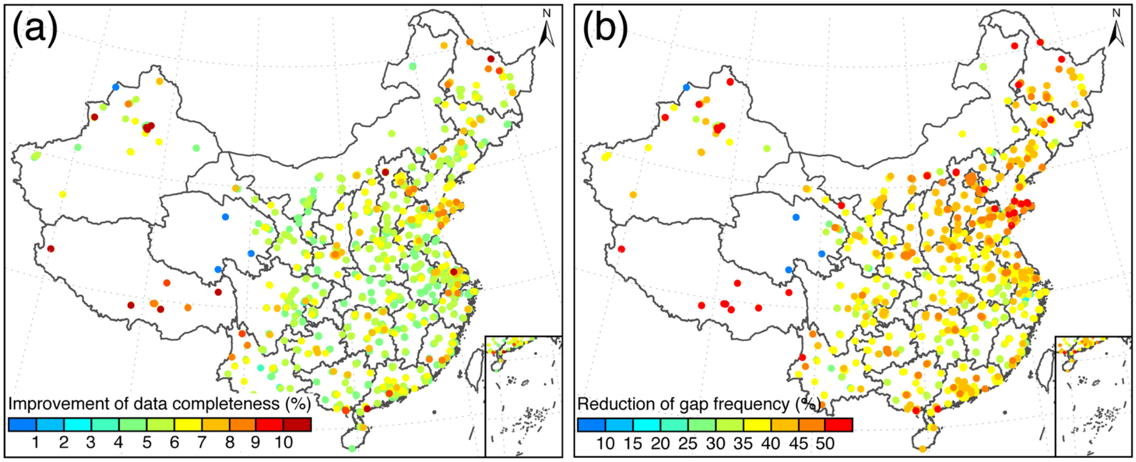


**Figure 10.** Benefits of the DCCEOF method on China in-situ hourly $PM_{2.5}$ records at each individual monitoring station. (a) Improvement of data completeness ratio, and (b) reduction of the percentage of days with missingness.

### 4.3 Discussion

Compared with conventional interpolation approaches, the proposed DCCEOF method has better accuracy in predicting missing values for the emerged data gaps in hourly $PM_{2.5}$ time series given the principle of accounting for the local diurnal variation pattern of $PM_{2.5}$ concentration. Specifically, site specific diurnal cycle of $PM_{2.5}$ is reconstructed from the discrete spatial and temporal neighborhood using EOF and is then used as a reference to predict the missing values. Such a scheme is able to capture the local variation pattern of $PM_{2.5}$ with good accuracy in regions with dense neighboring stations (e.g., eastern China) and less temporal dynamics of $PM_{2.5}$, in contrast, relatively poor accuracy could be attained in the western part of the country where stations are sparsely distributed given the lack of adequate neighboring information. In such context, the performance of DCCEOF could be further improved by using a general diurnal variation pattern of $PM_{2.5}$ that is prior determined though a typical classification. However, this endeavor needs us to have a clear prior information of diurnal variability of $PM_{2.5}$ in space and time. On the other hand, the diurnal variation pattern of other relevant factors that are highly associated with $PM_{2.5}$ variations, e.g., meteorological factors such as mixing layer height, might be also applied to better reconstruct the local diurnal cycle of $PM_{2.5}$.

Although the DCCEOF method has a promising accuracy in filling the data gaps present in hourly $PM_{2.5}$ concentration time series, the current method only works for days with at least several valid observations. In other words, the DCCEOF method is incapable of restoring values for days with all 24-h data missed. This is because the remnant valid observations within 24-h are used as a critical constraint not only to convolve with other neighboring observations in space and time to identify similar observations but also to modulate the data magnitude of predicted values for missingness. Moreover, the severity of data gaps in the initial neighborhood field is also associated with the final prediction accuracy because significant data gaps in the neighborhood field could introduce large bias to the reconstructed local diurnal variation pattern of $PM_{2.5}$. In such context, aforementioned proxy information such as the diurnal variation pattern of meteorological conditions might be applied as a good complementary.

## 5 Conclusions

A practical and realistic gap filling method termed DCCEOF is proposed in this study to cope with emerged data gaps in geophysical time series, particularly for those with significant diurnal variability. Compared with the conventional interpolation methods, the proposed DCCEOF method is self–consistent, physically meaningful, and more accurate, given the accounting for the local diurnal variation pattern of the monitoring factor in missing value restorations. Such an endeavor enables the DCCEOF method to predict missing values even at inflection times, like daily peaks or minima that conventional methods always fail to predict properly, with promising accuracy.

A practical application of the DCCEOF method to the China in-situ hourly $PM_{2.5}$ concentration record reveals a good prediction accuracy of the DCCEOF method in restoring $PM_{2.5}$ missingness. The method performs even better in predicting missing values during polluted phases than on clean days given smaller variations of $PM_{2.5}$ concentration in space and time. Further sensitivity experiments suggest that the overall accuracy of the DCCEOF method would slightly decrease (from 0.96 to 0.9) with the increase in the amount of missingness in daily 24-h $PM_{2.5}$ observations. This effect is associated with larger uncertainties in the reconstruction of local $PM_{2.5}$ neighborhood fields since valid observations are required to convolve with other observations for the identification of observations with similar variation pattern. Also, an adequate number of neighboring stations in space is essential to the final prediction accuracy of missing value restoration. The experimental results suggest that three neighboring stations within 100 km to the target station would yield a promising prediction accuracy, and the more the neighboring stations, the less the uncertainties in the final predicted values.

In addition, we also assessed the severity of data gaps in our retrieved China in–situ hourly $PM_{2.5}$ records. In general, the missingness ratio was less than 10% over most stations across China while data gaps occurred more frequently at 0600 and 1200 BJT than during other times. After gap filling, the data completeness ratio of China in–situ hourly $PM_{2.5}$ record was improved to 94.3% while the frequency of days with missingness was markedly reduced from 42.6% to 5.7%. The gap-filled hourly $PM_{2.5}$ concentration record can thus be used as a promising data source for better $PM_{2.5}$ concentration mapping and exposure assessment.

Overall, the proposed DCCEOF method provides a realistic and promising way to deal with missingness emerged in hourly $PM_{2.5}$ concentration record which oftentimes exhibits significant diurnal variation patterns. Given the self-consistent nature, the DCCEOF method can thus be directly applied to $PM_{2.5}$ datasets measured in other regions and/or other geophysical time series with similar barriers. A more general comparison of this method with many other conventional gap filling methods will be conducted in the future to further examine the performance and accuracy of the DCCEOF method in handling various types of data gaps.

**Acknowledgments**

The authors thank two reviewers for their constructive and insightful suggestions in helping improve the quality of this manuscript. We also acknowledge the China National Environment Monitoring Centre (http://www.cnemc.cn) for making the essential hourly $PM_{2.5}$ concentration measurements publicly available. This study was supported by the National Natural Science Foundation of China under grants 41701413, 41771399 and 91544217, and the Shanghai Sailing Program under grant 17YF1404100.

## References


Aydilek, I.B. and Arslan, A.: A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. Inf. Sci., 233, 25–35, doi:10.1016/j.ins.2013.01.021, 2013.

Bai, K., Chang, N.-B., Zhou, J., Gao, W. and Guo, J.: Diagnosing atmospheric stability effects on the modeling accuracy of $PM_{2.5}$/AOD relationship in eastern China using radiosonde data. Environ. Pollut., 251, 380–389, doi:10.1016/j.envpol.2019.04.104, 2019a.

Bai, K., Ma, M., Chang, N.-B. and Gao, W.: Spatiotemporal trend analysis for fine particulate matter concentrations in China using high-resolution satellite-derived and ground-measured $PM_{2.5}$ data. J. Environ. Manage., 233, 530–542, doi:10.1016/j.jenvman.2018.12.071, 2019b.

Bai, K., Li, K., Chang, N.-B., Gao, W.: Advancing the prediction accuracy of satellite-based $PM_{2.5}$ concentration mapping: A perspective of data mining through in situ $PM_{2.5}$ measurements. Environ. Pollut., 254, 113047, doi:10.1016/j.envpol.2019.113047, 2019c.

Beckers, J.M. and Rixen, M.: EOF Calculations and Data Filling from Incomplete Oceanographic Datasets. J. Atmos. Ocean. Technol., 20, 1839–1856, doi: 10.1175/1520-0426(2003)020<1839:ECADFF>2.0.CO;2, 2003.

Bi, J., Belle, J.H., Wang, Y., Lyapustin, A.I., Wildani, A. and Liu, Y.: Impacts of snow and cloud covers on satellite-derived $PM_{2.5}$ levels. Remote Sens. Environ., 221, 665–674. doi:10.1016/j.rse.2018.12.002, 2018.

Bondon, P.: Influence of Missing Values on the Prediction of a Stationary Time Series. J. Time Ser. Anal., 26, 519–525, doi:10.1111/j.1467-9892.2005.00433.x, 2005.

Chang, N.-B., Bai, K. and Chen, C.-F.: Smart Information Reconstruction via Time-Space-Spectrum Continuum for Cloud Removal in Satellite Images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., 8, 1898–1912, doi:10.1109/JSTARS.2015.2400636, 2015.

Chen, B., Huang, B., Chen, L. and Xu, B.: Spatially and Temporally Weighted Regression: A Novel Method to Produce Continuous Cloud-Free Landsat Imagery. IEEE Trans. Geosci. Remote Sens., 55, 27–37, doi:10.1109/TGRS.2016.2580576, 2017.

Chen, J., Zhu, X., Vogelmann, J.E., Gao, F. and Jin, S.: A simple and effective method for filling gaps in Landsat ETM+ SLC-off images. Remote Sens. Environ., 115, 1053–1064, doi:10.1016/j.rse.2010.12.010, 2011.

Chen, S., Hu, C., Barnes, B.B., Xie, Y., Lin, G. and Qiu, Z.: Improving ocean color data coverage through machine learning. Remote Sens. Environ., 222, 286–302, doi:10.1016/j.rse.2018.12.023, 2019.

Dembélé, M., Oriani, F., Tumbulto, J., Mariéthoz, G. and Schaefli, B.: Gap-filling of daily streamflow time series using Direct Sampling in various hydroclimatic settings. J. Hydrol., 569, 573–586, doi:10.1016/j.jhydrol.2018.11.076, 2019.

Demirhan, H. and Renwick, Z.: Missing value imputation for short to mid-term horizontal solar irradiance data. Appl. Energy, 225, 998–1012, doi:10.1016/j.apenergy.2018.05.054, 2018.

Dray, S. and Josse, J.: Principal component analysis with missing values: a comparative survey of methods. Plant Ecol., 216, 657–667, doi:10.1007/s11258-014-0406-z, 2015.

Gao, S., Hu, H., Wang, Y., Zhang, X., Sun, L., Huang, F., Zhao, C., Wang, W., Liu, X., Wang, J., Zhou, Y. and Qu, W.: Effect of weakened diurnal evolution of atmospheric boundary layer to air pollution over eastern China associated to aerosol, cloud – ABL feedback. Atmos. Environ., 185, 168–179, doi:10.1016/j.atmosenv.2018.05.014, 2018.

Gerber, F., de Jong, R., Schaepman, M.E., Schaepman-Strub, G. and Furrer, R.: Predicting Missing Values in Spatio-Temporal Remote Sensing Data. IEEE Trans. Geosci. Remote Sens., 56, 2841–2853, doi:10.1109/TGRS.2017.2785240, 2018.

Guo, J., Zhang, X., Che, H., Gon, S., An, X., Cao, C., Guang, J., Zhang, H., Wang, Y., Zhang, X., Zhao, P. and Li, X.: Correlation between PM concentrations and aerosol optical depth in eastern China, Atmos. Environ., 43(37): 5876–5886, 2009.

Guo, J., Miao, Y., Zhang, Y., Liu, H., Li, Z., Zhang, W., He, J., Lou, M., Yan, Y., Bian, L. and Zhai, P.: The climatology of planetary boundary layer height in China derived from radiosonde and reanalysis data. Atmos. Chem. Phys., 16, 13309–13319, doi:10.5194/acp-16-13309-2016, 2016.

Guo, J., Xia, F., Zhang, Y., Liu, H., Li, J., Lou, M., He, J., Yan, Y., Wang, F., Min, M. and Zhai, P.: Impact of diurnal variability and meteorological factors on the $PM_{2.5}$-AOD relationship: Implications for $PM_{2.5}$ remote sensing. Environ. Pollut., 221, 94–104, doi:10.1016/j.envpol.2016.11.043, 2017.

Huang, K., Xiao, Q., Meng, X., Geng, G., Wang, Y., Lyapustin, A., Gu, D. and Liu, Y. Predicting monthly high-resolution $PM_{2.5}$ concentrations with random forest model in the North China Plain. Environ. Pollut., 242, 675-683, doi:10.1016/j.envpol.2018.07.016, 2018.

Huang, X., Wang, Z. and Ding, A.: Impact of Aerosol-PBL Interaction on Haze Pollution: Multiyear Observational Evidences in North China. Geophys. Res. Lett., 45, 8596–8603, doi:10.1029/2018GL079239, 2018.

Jönsson, P. and Eklundh, L.: TIMESAT—a program for analyzing time-series of satellite sensor data. Comput. Geosci., 30, 833–845, doi:10.1016/j.cageo.2004.05.006, 2004.

Julien, Y. and Sobrino, J.A.: Optimizing and comparing gap-filling techniques using simulated NDVI time series from remotely sensed global data. Int. J. Appl. Earth Obs. Geoinf., 76, 93–111, doi:10.1016/j.jag.2018.11.008, 2019.

Junger, W.L. and Ponce de Leon, A.: Imputation of missing data in time series for air pollutants. Atmos. Environ., 102, 96–104, doi:10.1016/j.atmosenv.2014.11.049, 2015.

Kandasamy, S., Baret, F., Verger, A., Neveux, P. and Weiss, M.: A comparison of methods for smoothing and gap filling time series of remote sensing observations–application to MODIS LAI products. Biogeosciences, 10, 4055–4071, doi:10.5194/bg-10-4055-2013, 2013.

Konik, M., Kowalewski, M., Bradtke, K. and Darecki, M.: The operational method of filling information gaps in satellite imagery using numerical models. Int. J. Appl. Earth Obs. Geoinf., 75, 68–82, doi:10.1016/j.jag.2018.09.002, 2019.

Körner, P., Kronenberg, R., Genzel, S.and Bernhofer, C.: Introducing Gradient Boosting as a universal gap filling tool for meteorological time series. Meteorol. Zeitschrift, 27, 369–376, doi:10.1127/metz/2018/0908, 2018.

Larose, C., Dey, D. and Harel, O.: The Impact of Missing Values on Different Measures of Uncertainty. Stat. Sin., 29:511-566, doi:10.5705/ss.202016.0073, 2019.

Lennartson, E.M., Wang, J., Gu, J., Castro Garcia, L., Ge, C., Gao, M., Choi, M., Saide, P.E., Carmichael, G.R., Kim, J., Janz, S.J.: Diurnal variation of aerosol optical depth and $PM_{2.5}$ in South Korea: a synthesis from AERONET, satellite (GOCI), KORUS-AQ observation, and the WRF-Chem model. Atmos. Chem. Phys., 18:15125–15144, doi:10.5194/acp-18-15125-2018, 2018.

Li, L., Zhang, J., Qiu, W., Wang, J. and Fang, Y.: An ensemble spatiotemporal model for predicting $PM_{2.5}$ concentrations. Int. J. Environ. Res. Public Health, 14, 549, doi:10.3390/ijerph14050549, 2017a.

Li, T., Shen, H., Zeng, C., Yuan, Q. and Zhang, L.: Point-surface fusion of station measurements and satellite observations for mapping $PM_{2.5}$ distribution in China: Methods and assessment. Atmos. Environ., 152, 477-489, doi:1 0.1016/j.atmosenv.2017.01.004, 2017.

Li, Z., Guo, J., Ding, A., Liao, H., Liu, J., Sun, Y., Wang, T., Xue, H., Zhang, H. and Zhu, B.: Aerosol and boundary-layer interactions and impact on air quality. Natl. Sci. Rev., 4, 810–833, doi:10.1093/nsr/nwx117, 2017b.

Liu, L., Guo, J., Miao, Y., Liu, L., Li, J., Chen, D., He, J. and Cui, C.: Elucidating the relationship between aerosol concentration and summertime boundary layer structure in central China. Environ. Pollut., 241, 646–653, doi:10.1016/j.envpol.2018.06.008, 2018.

Liu, X. and Wang, M.: Filling the Gaps of Missing Data in the Merged VIIRS SNPP/NOAA-20 Ocean Color Product Using the DINEOF Method. Remote Sens, 11, 178, doi:10.3390/rs11020178, 2019.

Lolli, S., Di Girolamo, P.: Principal Component Analysis Approach to Evaluate Instrument Performances in Developing a Cost-Effective Reliable Instrument Network for Atmospheric Measurements. J. Atmos. Ocean. Technol., 32, 1642–1649. doi:10.1175/JTECH-D-15-0085.1, 2015.

Mahmoudvand, R. and Rodrigues, P.C.: Missing value imputation in time series using Singular Spectrum Analysis. Int. J. Energy Stat., 04, 1650005, doi:10.1142/S2335680416500058, 2016.

Manning, M.I., Martin, R. V., Hasenkopf, C., Flasher, J. and Li, C.: Diurnal Patterns in Global Fine Particulate Matter Concentration. Environ. Sci. Technol. Lett., 5, 687–691, doi:10.1021/acs.estlett.8b00573, 2018.

Miao, Y., Liu, S., Guo, J., Huang, S., Yan, Y. and Lou, M.: Unraveling the relationships between boundary layer height and $PM_{2.5}$ pollution in China based on four-year radiosonde measurements. Environ. Pollut., 243, 1186–1195, doi:10.1016/j.envpol.2018.09.070, 2018.

Miller, L., Xu, X., Wheeler, A., Zhang, T., Hamadani, M. and Ejaz, U.: Evaluation of missing value methods for predicting ambient BTEX concentrations in two neighbouring cities in Southwestern Ontario Canada. Atmos. Environ., 181, 126–134, doi:10.1016/j.atmosenv.2018.02.042, 2018.

Neteler, M.: Estimating daily land surface temperatures in mountainous environments by reconstructed MODIS LST data. Remote Sens., 2, 333–351, doi:10.3390/rs1020333, 2010.

Nosal, M., Legge, A.H. and Krupa, S. V.: Application of a stochastic, Weibull probability generator for replacing missing data on ambient concentrations of gaseous pollutants. Environ. Pollut., 108, 439–446, doi:10.1016/S0269-7491(99)00220-1, 2000.

Oriani, F., Borghi, A., Straubhaar, J., Mariethoz, G. and Renard, P.: Missing data simulation inside flow rate time-series using multiple-point statistics. Environ. Model. Softw., 86, 264–276, doi:10.1016/j.envsoft.2016.10.002, 2016.

Ottosen, T.-B. and Kumar, P.: Outlier detection and gap filling methodologies for low-cost air quality measurements. Environ. Sci. Process. Impacts, 21, 701–713, doi:10.1039/C8EM00593A, 2019.

Rossi, R.E., Dungan, J.L. and Beck, L.R.: Kriging in the shadows: Geostatistical interpolation for remote sensing. Remote Sens. Environ., 49, 32–40, doi:10.1016/0034-4257(94)90057-4, 1994.

Şahin, Ü.A., Bayat, C. and Uçan, O.N.: Application of cellular neural network (CNN) to the prediction of missing air pollutant data. Atmos. Res., 101, 314–326, doi:10.1016/j.atmosres.2011.03.005, 2011.

Shareef, M.M., Husain, T. and Alharbi, B.: Optimization of Air Quality Monitoring Network Using GIS Based Interpolation Techniques. J. Environ. Prot., 07, 895–911, doi:10.4236/jep.2016.76080, 2016.

Shen, H., Li, T., Yuan, Q. and Zhang, L.: Estimating regional ground-level $PM_{2.5}$ directly from satellite top-of-atmosphere reflectance using deep belief networks. J. Geophys. Res. Atmos., 123, 13,875-13,886, doi:10.1029/2018JD028759, 2018.

Shi, X., Zhao, C., Jiang, J.H., Wang, C., Yang, X. and Yung, Y.L.: Spatial Representativeness of $PM_{2.5}$ Concentrations Obtained Using Observations From Network Stations. J. Geophys. Res. Atmos., 123, 3145–3158, doi:10.1002/2017JD027913, 2018.

Singh, M.K., Venkatachalam, P. and Gautam, R.: Geostatistical methods for filling gaps in level-3 monthly-mean aerosol optical depth data from multi-angle imaging spectroradiometer. Aerosol Air Qual. Res., 17, 1963–1974, doi:10.4209/aaqr.2016.02.0084, 2017.

Stauch, V.J. and Jarvis, A.J.: A semi-parametric gap-filling model for eddy covariance $CO_2$ flux time series data. Glob. Chang. Biol., 12, 1707–1716, doi:10.1111/j.1365-2486.2006.01227.x, 2006.

Taylor, M.H., Losch, M., Wenzel, M. and Schröter, J.: On the sensitivity of field reconstruction and prediction using empirical orthogonal functions derived from Gappy data. J. Clim., 26, 9194–9205, doi:10.1175/JCLI-D-13-00089.1, 2013.

van Donkelaar, A., Martin, R. V., Brauer, M., Hsu, N.C., Kahn, R.A., Levy, R.C., Lyapustin, A., Sayer, A.M. and Winker, D.M.: Global Estimates of Fine Particulate Matter using a Combined Geophysical-Statistical Method with Information from Satellites, Models, and Monitors. Environ. Sci. Technol., 50, 3762–3772, doi:10.1021/acs.est.5b05833, 2016.

Wang, J. and Christopher, S.A.: Intercomparison between satellite-derived aerosol optical thickness and PM 2.5 mass: Implications for air quality studies. Geophys. Res. Lett., 30: 2095, doi:10.1029/2003GL018174, 2003.

Yadav, M.L. and Roychoudhury, B.: Handling missing values: A study of popular imputation packages in R. Knowledge-Based Syst., 160, 104–118, doi:10.1016/j.knosys.2018.06.012, 2018.

Yang, Q., Yuan, Q., Yue, L., Li, T., Shen, H. and Zhang, L.: The relationships between $PM_{2.5}$ and aerosol optical depth (AOD) in mainland China: About and behind the spatio-temporal variations. Environ. Pollut., 248, 526–535, doi:10.1016/j.envpol.2019.02.071, 2019a.

Yang Y., X. Zheng, Z. Gao, H. Wang, T. Wang, Y. Li, G.N.C. Lau, S.H.L. Yim,(2018) LongTerm Trends of Persistent Synoptic Circulation Events in Planetary Boundary Layer and Their Relationships with Haze Pollution in Winter HalfYear over Eastern China, *Journal of Geophysical Research - Atmospheres*, doi: 10.1029/2018JD028982

Yang Y., S. H.L. Yim, J. Haywood, M. Osborne, J. C.S. Chan, Z. Zeng, J. C.H. Cheng (2019b), Characteristics of heavy particulate matter pollution events over Hong Kong and their relationships with vertical wind profiles using high-time-resolution Doppler Lidar measurements, Journal of Geophysical Research -Atmospheres, doi: 10.1029/2019JD031140

Ye, W.F., Ma, Z.Y. and Ha, X.Z. Spatial-temporal patterns of $PM_{2.5}$ concentrations for 338 Chinese cities. Sci. Total Environ., 631–632, 524–533, doi:10.1016/j.scitotenv.2018.03.057, 2018.

Zhang, D., Bai, K., Zhou, Y., Shi, R. and Ren, H.: Estimating Ground-Level Concentrations of Multiple Air Pollutants and Their Health Impacts in the Huaihe River Basin in China. Int. J. Environ. Res. Public Health, 16, 579, doi:10.3390/ijerph16040579, 2019.

Zhang, T., Zhu, Zhongmin, Gong, W., Zhu, Zerun, Sun, K., Wang, L., Huang, Y., Mao, F., Shen, H., Li, Z. and Xu, K.: Estimation of ultrahigh resolution $PM_{2.5}$ concentrations in urban areas using 160 m Gaofen-1 AOD retrievals. Remote Sens. Environ., 216, 91–104, doi:10.1016/j.rse.2018.06.030, 2018.

Zhu, X., Liu, D. and Chen, J.: A new geostatistical approach for filling gaps in Landsat ETM+ SLC-off images. Remote Sens. Environ., 124, 49–60, doi:10.1016/j.rse.2012.04.019, 2012.

Zhu, Y., Kang, E., Bo, Y., Tang, Q., Cheng, J. and He, Y.: A robust fixed rank kriging method for improving the spatial completeness and accuracy of satellite SST products. IEEE Trans. Geosci. Remote Sens., 53, 5021–5035, doi:10.1109/TGRS.2015.2416351, 2015.