

Optimal Planning of Air Quality-Monitoring Sites for Better Depiction of PM_{2.5} Pollution across China

Chenhong Zhou, Meng Gao,* Jianjun Li, Kaixu Bai, Xiao Tang, Xiao Lu, Cheng Liu, Zifa Wang, and Yike Guo*



Cite This: *ACS Environ. Au* 2022, 2, 314–323



Read Online

ACCESS |



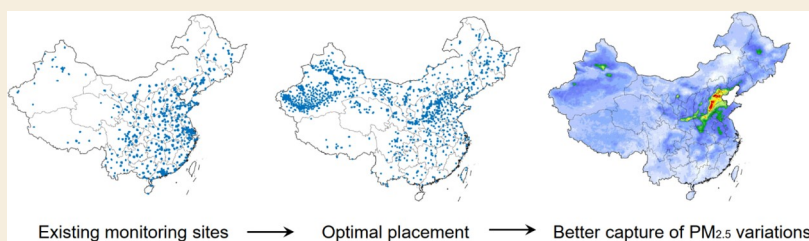
Metrics & More



Article Recommendations



Supporting Information



ABSTRACT: A myriad of studies have attempted to use ground-level observations to obtain gap-free spatiotemporal variations of PM_{2.5}, in support of air quality management and impact studies. Statistical methods (machine learning, etc.) or numerical methods by combining chemical transport modeling and observations with data assimilation techniques have been typically applied, yet the significance of site placement has not been well recognized. In this study, we apply five proper orthogonal decomposition (POD)-based sensor placement algorithms to identify optimal site locations and systematically evaluate their reconstruction ability. We demonstrate that the QR pivot is relatively more reliable in deciding optimal monitoring site locations. When the number of planned sites (sensors) is limited, using a lower number of modes would yield lower estimation errors. However, the dimension of POD modes has little impact on reconstruction quality when sufficient sensors are available. The locations of sites guided by the QR pivot algorithm are mainly located in regions where PM_{2.5} pollution is severe. We compare reconstructed PM_{2.5} pollution based on QR pivot-guided sites and existing China National Environmental Monitoring Center (CNEMC) sites and find that the QR pivot-guided sites are superior to existing sites with respect to reconstruction accuracy. The current planning of monitoring stations is likely to miss sources of pollution in less-populated regions, while our QR pivot-guided sites are planned based on the severity of PM_{2.5} pollution. This planning methodology has additional potentials in chemical data assimilation studies as duplicate information from current CNEMC-concentrated stations is not likely to boost performance.

KEYWORDS: PM_{2.5} pollution, sensor placement, ground-level PM_{2.5} estimation, aerosol data assimilation

1. INTRODUCTION

China's rapid and energy-intensive development over recent decades has caused a series of environmental issues, among which PM_{2.5} (particulate matter with an aerodynamic diameter of less than 2.5 μm) pollution received worldwide attention.^{1–4} It has been well recognized that both acute and chronic exposures to high PM_{2.5} concentrations are associated with multiple health issues, including heart disease, lung cancer, respiratory infection, etc.^{5–8} Monitoring air pollution is critical to understanding the formation mechanism of PM_{2.5}, to supporting air quality management, and to reducing human exposure.^{9,10} Since 2013, Chinese government has established a ground-based monitoring network across China to measure hourly concentrations of six air pollutants, including PM_{2.5}.¹¹ These monitoring stations were empirically established, mainly concentrated in urban areas.^{12,13} Activities in rural or suburb regions including agricultural burning also exert influences on air quality, weather, and climate, yet are commonly missed by the current monitoring network.

To fill these gaps, numerous studies^{12,14–16} have proposed to integrate satellite aerosol optical depth (AOD) with ground-measured PM_{2.5} using machine learning algorithms to obtain a better depiction of spatiotemporal variations of PM_{2.5}. Due to existences of clouds and occurrences of snow/extreme haze, AOD-retrieving algorithms usually fail to offer valid values.¹⁷ Moreover, a large range of missing and abnormal data could happen when some satellites stay in service beyond their design life.¹³ These issues could be addressed with advances in machine learning and growing volume of observations, and a wide range of algorithms or satellite retrievals (low-Earth orbit,

Received: November 16, 2021

Revised: February 27, 2022

Accepted: February 28, 2022

Published: March 10, 2022



geostationary, etc.) were explored previously.^{12,14–16} Aerosol data assimilation methods were also developed to improve the spatiotemporal representation of PM_{2.5}.^{18,19}

It has been generally believed in the data assimilation community that not every observation influences prediction identically.¹⁹ Similarly, reconstruction of PM_{2.5} based on machine learning algorithms is also largely influenced by locations of monitoring sites. Currently, the planning of locations of air quality monitoring has not dealt with the issue of optimal placement of sites with respect to data assimilation or data-driven reconstruction of air pollution. In the weather forecasting community, Ancell and Hakim²⁰ demonstrated that observation targeting inferred with sensitivity analysis could help improve the accuracy of forecasts. However, the significance of air pollution observation targeting has not been well recognized.

These problems fall under the umbrella of the optimal sensor placement (OSP), which answers how to use sparse sensor networks to infer a spatiotemporal varying process. The most intuitive solution to design optimal sensor locations is a brute-force search or an exhaustive search, which generates $\frac{n!}{(n-p)!p!}$ combinatorial possibilities for p observing locations out of n points in the whole field. However, the complexity of this method exponentially increases as p and n grow. It seems intractable for more complex problems with larger search spaces, especially for a high-dimensional and possibly nonlinear dynamical system.²¹ With the generation of big data and advances in computing power, many modern techniques, such as machine learning and sparse sampling, have been leveraged to determine optimal sensor locations for reconstruction. More specifically, dimensionality reduction techniques, such as proper orthogonal decomposition (POD),^{22–25} also known as Karhunen–Loève expansion, empirical orthogonal function, and principal component analysis, have been commonly used to provide a compact description of spatial and temporal variabilities in a dynamical system.^{26,27} Gappy POD, an extension of POD, was proposed by Everson and Sirovich²⁸ to handle incomplete or gappy data and has been successfully applied in the reconstruction of unsteady flow,²⁹ human faces,²⁸ and ocean surface temperatures.³⁰ Within the gappy POD framework, a number of algorithms have been proposed to minimize reconstruction error, including random sampling,^{31,32} extrema of the POD modes,³³ minimization of the matrix condition number (MCN),²⁹ QR with column pivoting,²¹ and discrete empirical interpolation method (DEIM).³⁴

Where to place air pollution monitoring and how the placement would benefit the depiction of PM_{2.5} remain unclear. In this study, we explore the theoretical framework and associated algorithms for optimally placing PM_{2.5}-monitoring stations. The results offer valuable implications for the future planning of monitoring stations with respect to potential studies of data assimilation or data-driven reconstruction of air pollution. In Section 2, we provide a brief introduction to POD-based reconstruction and several POD-based sensor placement algorithms. In Section 3, we comprehensively evaluate these sensor placement algorithms and discuss the benefits of optimal placements over existing locations of monitoring stations. Major findings are summarized in Section 4.

2. METHODOLOGY

2.1. POD-Based Reconstruction

POD excels in capturing dominant structures of studied fields in the forms of orthonormal eigenmodes that define a low-dimensional embedding space. The whole field's information can be encoded into such a feature space and expressed as linear combinations of a set of POD modes. Given a data matrix $\mathbf{U} \in \mathbb{R}^{N \times Q}$ with Q snapshots in \mathbb{R}^N , a snapshot $\mathbf{u}(\mathbf{x}, t) \in \mathbb{R}^N$ is a spatial map of studied fields at a given time t , which can be represented by

$$\mathbf{u}(\mathbf{x}, t) = \sum_{k=1}^{\infty} \mathbf{a}^k(t) \Phi^k(\mathbf{x}) \quad (1)$$

or approximated as a truncated expansion:

$$\mathbf{u}(\mathbf{x}, t) \approx \sum_{k=1}^K \mathbf{a}^k(t) \Phi^k(\mathbf{x}) \quad (2)$$

where K represents the number of orthonormal POD basis modes $\Phi^k(\mathbf{x})$ and $\mathbf{a}^k(t)$ denotes the time-dependent POD coefficients. Suppose the gappy vector is denoted as $\tilde{\mathbf{u}}(\mathbf{x}, t)$ and it is a point-wise product of a mask vector $\mathbf{m}(\mathbf{x}, t)$ and complete vector $\mathbf{u}(\mathbf{x}, t)$, that is

$$\tilde{\mathbf{u}}(\mathbf{x}, t) = \mathbf{m}(\mathbf{x}, t) \cdot \mathbf{u}(\mathbf{x}, t) \quad (3)$$

The mask vector $\mathbf{m}(\mathbf{x}, t)$ consists of values of 0 or 1, indicating whether the data at the corresponding locations are missing or available. Our aim is to construct an estimator $\tilde{\mathbf{u}}_{\zeta}$ to approximate $\tilde{\mathbf{u}}$ using the existing spatial POD modes $\Phi^k(\mathbf{x})$. This can be represented as

$$\tilde{\mathbf{u}}_{\zeta} = \sum_{k=1}^K \mathbf{b}^k(t) \Phi^k(\mathbf{x}) \quad (4)$$

where $\mathbf{b}^k(t)$ stands for unknown POD coefficients. The error between the intermediate repaired vector $\tilde{\mathbf{u}}_{\zeta}$ and the gappy vector $\tilde{\mathbf{u}}$ can be minimized by

$$\xi = \|\tilde{\mathbf{u}} - \tilde{\mathbf{u}}_{\zeta}\|_m^2 \quad (5)$$

where the subscript m denotes a gappy norm, not a standard L2-norm, which means only original existing elements in $\tilde{\mathbf{u}}$ are compared. As described by Willcox,²⁹ the gappy inner product is defined

$$(\mathbf{w}, \mathbf{v})_m = ((\mathbf{m} \cdot \mathbf{w}), (\mathbf{m} \cdot \mathbf{v})) \quad (6)$$

and the corresponding gappy norm is

$$\|\mathbf{v}\|_m^2 = (\mathbf{v}, \mathbf{v})_m \quad (7)$$

Then, the error in eq 5 can be unfolded as

$$\begin{aligned} \xi &= \|\tilde{\mathbf{u}}\|_m^2 - 2 \sum_{k=1}^K \mathbf{b}^k(\tilde{\mathbf{u}}, \Phi^k(\mathbf{x}))_m \\ &\quad + \sum_{k=1}^K \sum_{j=1}^K \mathbf{b}^k \mathbf{b}^j (\Phi^k(\mathbf{x}), \Phi^j(\mathbf{x}))_m \end{aligned} \quad (8)$$

We differentiate it with respect to $\mathbf{b}^k(t)$ and yield a linear equation system for the coefficient \mathbf{b}^k

$$\mathbf{M} \mathbf{b} = \mathbf{f} \quad (9)$$

where $\mathbf{M}_{ij} = (\Phi^i, \Phi^j)_m$ and $\mathbf{f}_i = (\tilde{\mathbf{u}}, \Phi^i)_m$. Specifically, the matrix \mathbf{M} depends on the mask vector and POD basis vectors. Once the coefficients \mathbf{b}^k are determined, we can obtain the intermediate-repaired vector $\tilde{\mathbf{u}}^\zeta$ through eq 4. Finally, the missing elements in gappy data $\tilde{\mathbf{u}}$ can be repaired by the corresponding elements in $\tilde{\mathbf{u}}^\zeta$.

2.2. Algorithms for Efficient Sensor Placement

Considerable previous studies proposed multiple algorithms to determine sensor locations within the gappy framework. The simplest approach is to sample randomly (hereafter referred to as Random algorithm).^{21,28,32,35,36} Some studies^{29,37} leveraged the properties of the linear system and designed a method that minimizes the matrix condition number \mathbf{M} or $\kappa(\mathbf{M})$ (hereafter referred to as MCN algorithm). Mathematically, the condition number of a matrix \mathbf{M} refers to the ratio of the maximum to minimum singular values of \mathbf{M} , which reveals the orthogonality of the matrix. The matrix \mathbf{M} is a $K \times K$ identity matrix in cases of complete data, and all singular values equal to 1. Its condition number $\kappa(\mathbf{M})$ equals to 1 accordingly for complete data, yet orthogonality is lost ($\kappa(\mathbf{M}) > 1$) for gappy data. We can thus set up an optimization problem which minimizes the condition number \mathbf{M} by selecting sensor locations to preserve orthogonality in order to approach closer to conditions of complete data. The MCN algorithm is computationally costly as it needs to look over all remaining locations and calculate the condition number at each iteration.

A heuristic physics-based approach based on the extrema of dominant POD was developed to maximally capture the variance in the data.³² Specifically, Cohen et al.³³ placed sensors at the energetic maxima and minima of each mode for modeling unsteady flow past a circular cylinder since the positions of the extrema are the areas with highest modal activity. Normally, the number of sensor locations that are maxima and minima of POD modes is often less than the double of the number of modes as locations of maxima and minima for different modes may coincide. To relax the quantity limit of the potential sensor locations, the local extrema (hereafter referred to as Extrema algorithm) of each mode can also be involved to introduce more sensor locations.³⁶ More recently, variants of the empirical interpolation method (EIM) were developed to provide a principled selection of interpolation points for data reconstruction.²¹ As a discrete variant of EIM, the discrete EIM (hereafter referred to as DEIM algorithm)³⁸ has been applied with success and demonstrated to be a good candidate for sensor placement problems.^{21,32,38,39} In the POD-based DEIM algorithm, the nonlinear function $\mathbf{f}(\tau)$ of a system can be onto the subspace spanned by the basis $\{\Phi_1, \dots, \Phi_K\} \subset \mathbb{R}^N$ with the form

$$\mathbf{f}(\tau) = \Phi \mathbf{c}(\tau) \quad (10)$$

where $\Phi = \{\Phi_1, \dots, \Phi_M\} \in \mathbb{R}^{N \times K}$ and $\mathbf{c}(\tau) \in \mathbb{R}^K$ denotes the coefficient vector ($K \ll N$). $\mathbf{c}(\tau)$ can be simply estimated by $\mathbf{c}(\tau) = \Phi^T \mathbf{f}(\tau)$ when Φ represents POD bases, but it requires handling higher dimensional-state vectors (N -dimensional), leading to expensive computation. To deal with this over-determined system, we estimate $\mathbf{c}(\tau)$ by carefully selecting K interpolation points. Given the interpolation or measurement matrix $\mathbf{D} = [\mathbf{e}_{q_1}, \dots, \mathbf{e}_{q_K}] \in \mathbb{R}^{N \times K}$, where

$\mathbf{e}_{q_i} = [0, \dots, 0, \underset{q_i}{1}, 0, \dots, 0]^T \in \mathbb{R}^N$, $\mathbf{c}(\tau)$ can be uniquely determined with

$$\mathbf{D}^T \mathbf{f}(\tau) = \mathbf{D}^T \Phi \mathbf{c}(\tau) \quad (11)$$

Then, the DEIM approximation of $\mathbf{f}(\tau)$ becomes

$$\mathbf{f}_{\text{DEIM}}(\tau) = \Phi (\mathbf{D}^T \Phi)^{-1} \mathbf{D}^T \mathbf{f}(\tau) \quad (12)$$

DEIM provides a principled method to construct a set of indices inductively on basis vectors.³⁴

QR factorization with column pivoting is a specific version of the common QR-factorization method that was developed to handle rank-deficient least-squares problems. The reduced matrix QR factorization decomposes a matrix \mathbf{A} into a product of a unitary matrix \mathbf{Q} and an upper triangular matrix \mathbf{R} , that is, $\mathbf{A} = \mathbf{Q}\mathbf{R}$. Therefore, it follows $|\det(\mathbf{A})| = |\det(\mathbf{Q}) \cdot \det(\mathbf{R})| = |\det(\mathbf{R})| = \prod |r_{ii}| = \prod \lambda_{ip}$ where r_{ii} denotes diagonal entries and λ_i is the eigenvalues of \mathbf{R} . Thus, the condition number of \mathbf{A} can be controlled by optimizing the determinant or spectral radius of this matrix, that is, maximize $|\prod |r_{ii}||$. In general, the diagonal values r_{ii} of \mathbf{R} have no particular sequences. However, when combined with the column pivoting procedure, we introduce a column permutation matrix \mathbf{D} containing ones and zeros, thus $\mathbf{A}\mathbf{D} = \mathbf{Q}\mathbf{R}$. The pivoting procedure enforces the diagonal values of \mathbf{R} , r_{ip} to form decreasing sequence. Therefore, QR column pivoting increments the volume of the submatrix by reordering the columns of \mathbf{A} , which also maximizes the absolute value of the determinant. This approach can be applied for sensor placement problems based on the connections between the permutation matrix \mathbf{D} and the measurement matrix. Suppose a measurement matrix is defined as \mathbf{C} , due to the fact that sensor locations are fixed and do not vary with time, \mathbf{C} can be viewed as the matrix representation form of the mask vector \mathbf{m} (eq 3). Therefore, the observations \mathbf{y} can be represented as P elements selected from \mathbf{f} (where $\mathbf{f} = \Phi \mathbf{c}$ seen in eq 10) by $\mathbf{C} \in \mathbb{R}^{P \times N}$

$$\mathbf{y} = \mathbf{C}\mathbf{f} = \mathbf{C}\Phi \mathbf{c} = \Theta \mathbf{c} \quad (13)$$

where P is the number of sensors and N is the number of state dimension.

For the case of $P = K$, Θ becomes a square matrix and $\Theta = \mathbf{C}\Phi_K$ where Φ_K is the given tailored basis and K is the number of basis. For improved reconstruction, sensor locations need to be carefully selected to maximize the determinant of Θ , which in turn is expected to control the condition number. The following relationship holds for such a square matrix

$$\det(\Theta) = \det(\Theta^T) = \det(\Phi_K^T \mathbf{C}^T) \quad (14)$$

Thus, the QR factorization of $\Phi_K \in \mathbb{R}^{N \times K}$ with column pivoting yields

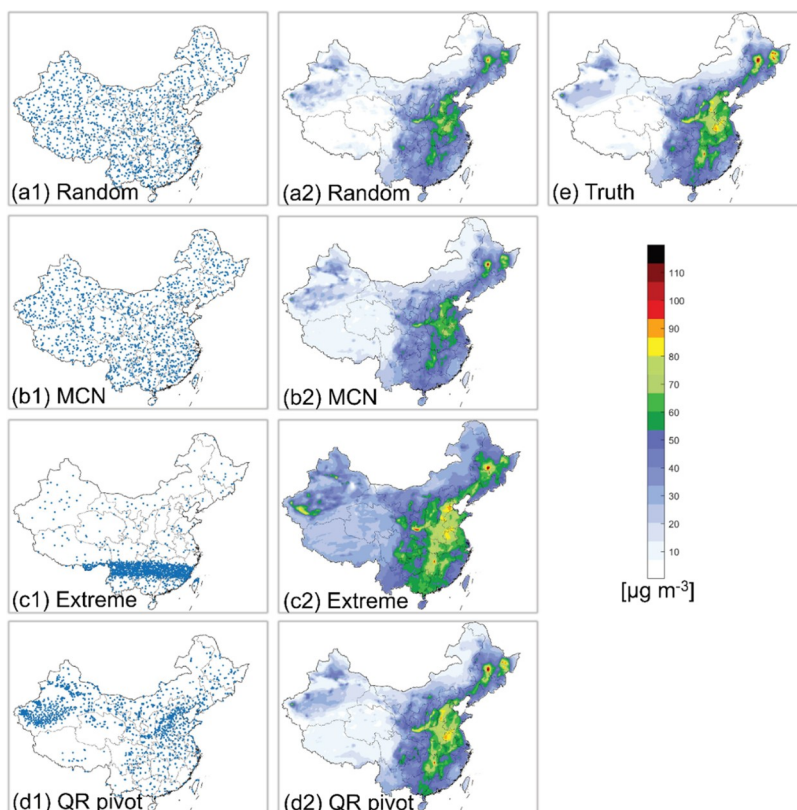
$$\Phi_K^T \mathbf{D} = \mathbf{Q}\mathbf{R} \quad (15)$$

where $\mathbf{D} \in \mathbb{R}^{N \times N}$ is a square permutation matrix. When \mathbf{C} builds upon the first P rows of \mathbf{D}^T , the right-hand side of eq 14 will be maximized. The index locations of ones in each row of \mathbf{C} are denoted by $[q_1, q_2, \dots, q_p]$, corresponding to point sensor locations in state space.

For oversampled cases with $P > K$, the determinant of $\mathbf{M} = \Theta^T \Theta \in \mathbb{R}^{K \times K}$ ($\mathbf{M} = \Theta$, if $P = K$) needs to be maximized, so that the condition number of \mathbf{M} is bounded. Based on this principle of pivoted QR, we have the following relationships

Table 1. Performance of Different Sensor Placement Algorithms Using Different Numbers of Sensors under POD Modes without Modal Truncation, i.e., $K = 657$ (Units of RMSE: $\mu\text{g}/\text{m}^3$)

algorithms	$P = K$			$P = 1.5K$			$P = 2K$		
	ϵ	R^2	RMSE	ϵ	R^2	RMSE	ϵ	R^2	RMSE
Random	109.70	0.02	4493.64	0.45	0.69	16.94	0.33	0.80	12.60
MCN	87.27	0.02	3405.29	0.44	0.70	16.73	0.32	0.81	12.55
Extrema	47.80	0.03	1721.57	1.49	0.10	55.79	0.69	0.45	25.98
DEIM	1.35	0.24	52.16						
QR pivot	1.03	0.31	40.11	0.33	0.81	12.49	0.28	0.84	10.98

**Figure 1.** Locations of sensors and reconstructed yearly mean $\text{PM}_{2.5}$ concentrations across China with Random (a1,a2), MCN (b1,b2), Extreme (c1,c2), and QR pivot (d1,d2) algorithms under the condition of $P = 2K$ and $K = 657$ and CAQRA yearly mean $\text{PM}_{2.5}$ as ground truth (e).

$$\begin{aligned} \det(\mathbf{M}) &= \det(\mathbf{\Theta}^T \mathbf{\Theta}) = \prod_i \sigma_i(\mathbf{\Theta}^T \mathbf{\Theta}) = \prod_i \sigma_i(\mathbf{\Theta} \mathbf{\Theta}^T) \\ &= \prod_i \sigma_i(\mathbf{C} \mathbf{\Phi}_K \mathbf{\Phi}_K^T \mathbf{C}^T) \end{aligned} \quad (16)$$

Thus, the maximization of $\det(\mathbf{M})$ can be realized by maximizing $\det(\mathbf{\Phi}_K \mathbf{\Phi}_K^T \mathbf{C}^T)$ by the pivoted QR factorization

$$(\mathbf{\Phi}_K \mathbf{\Phi}_K^T) \mathbf{D} = \mathbf{QR} \quad (17)$$

and \mathbf{C} is chosen as the first P rows of $\mathbf{D}^T \in \mathbb{R}^{N \times N}$. The index locations of ones in each row of \mathbf{C} are denoted by $[\mathcal{Q}_1, \mathcal{Q}_2, \dots, \mathcal{Q}_p]$.²¹

2.3. Descriptions of Data Sets

Experiments were performed with two $\text{PM}_{2.5}$ data sets, namely, the high-resolution Chinese air quality reanalysis data set (CAQRA)⁴⁰ and a long-term gap-free high-resolution air pollutant concentration data set (LGHAP) offering 1 km resolution satellite-derived daily $\text{PM}_{2.5}$ concentrations over

China.⁴¹ We chose the daily $\text{PM}_{2.5}$ concentrations across China from CAQRA at two time slots (8:00 and 20:00) in 2017. Therefore, the first data set is composed of 730 snapshots. Each snapshot has 432×339 pixels. The LGHAP data set was generated by a machine learning model with inputs of satellite observations of AOD, numerical simulations, and in situ measurements. The daily $\text{PM}_{2.5}$ concentrations in 2017 were utilized as the second data set consisting of 365 snapshots. To avoid the excessive consumption of computer memory, each snapshot in this data set was regridded to 410×680 . 10-fold cross validation (CV) was employed on these two data sets to ensure the reliability of our experiments. We also used locations of current China National Environmental Monitoring Center (CNEMC) network stations to conduct reconstructions. Then, the results were utilized to compare with the reconstructions of sensor placement algorithms to further illustrate the advantage and necessity of finding more optimal sensor locations.

To quantify the reconstruction performance of different sensor placement algorithms, we used three commonly used

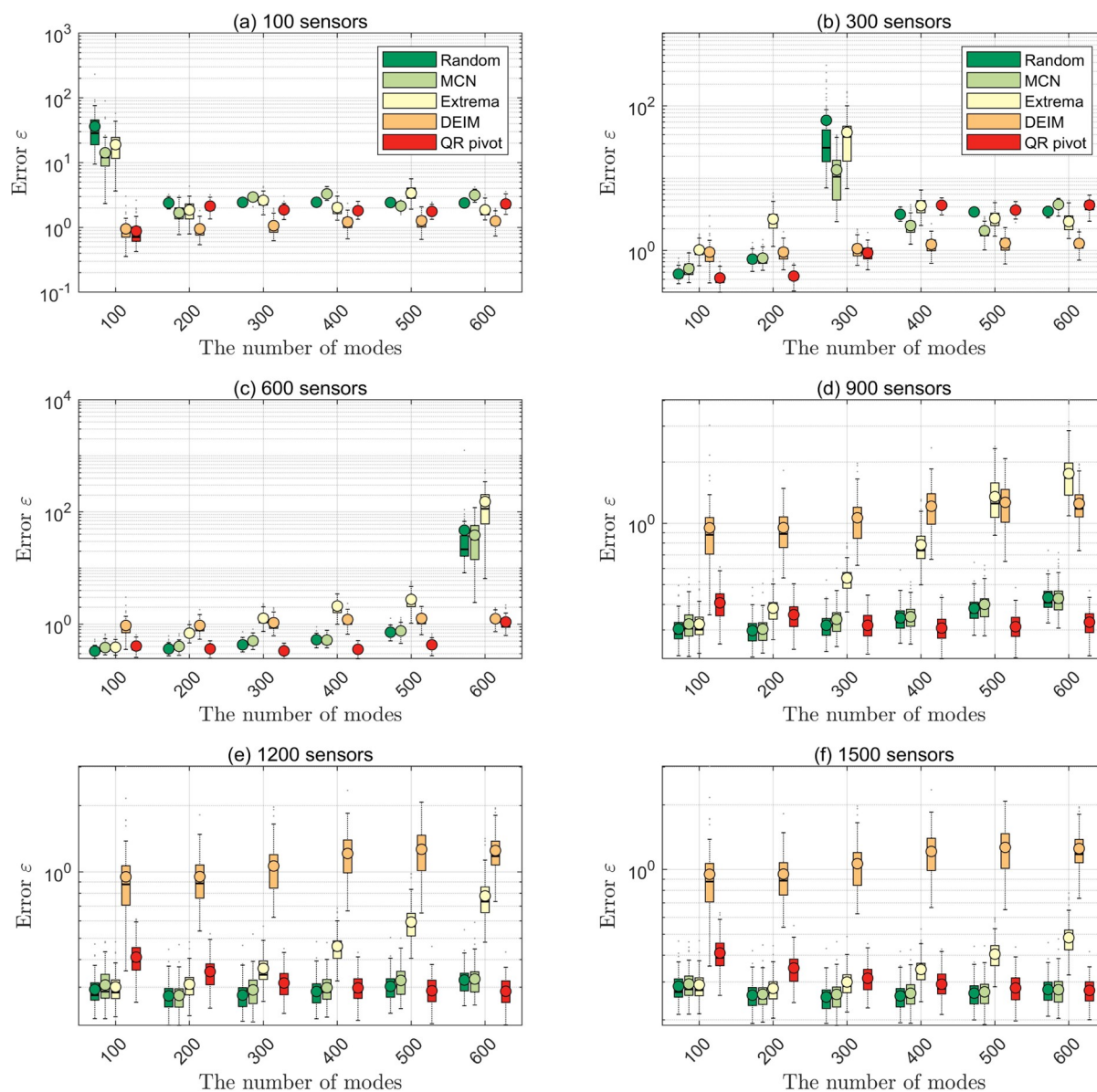


Figure 2. Reconstruction errors as a function of number of POD modes using (a) 100 sensors; (b) 300 sensors; (c) 600 sensors; (d) 900 sensors; (e) 1200 sensors; and (f) 1500 sensors.

evaluation metrics. First, we define the reconstruction error ε , as shown in eq 18

$$\varepsilon = \sqrt{\frac{\sum_{i=1}^N \|\text{PM}_{t,i} - \text{PM}_{e,i}\|^2}{\sum_{i=1}^N \|\text{PM}_{t,i}\|^2}} \quad (18)$$

where N represents the state dimension in a snapshot and PM_t and PM_e are the true $\text{PM}_{2.5}$ concentration and estimated $\text{PM}_{2.5}$ concentration, respectively. $\overline{\text{PM}}_t$ and $\overline{\text{PM}}_e$ are the mean $\text{PM}_{2.5}$ concentrations of PM_t and PM_e , respectively. In addition, the correlation of determination (R^2) and the root-mean-square error (RMSE, $\mu\text{g}/\text{m}^3$) are common metrics used in the estimation of $\text{PM}_{2.5}$ concentrations, which are defined, respectively, as follows

$$R^2 = \frac{\sum_{i=1}^N (\text{PM}_{t,i} - \overline{\text{PM}}_t)^2 (\text{PM}_{e,i} - \overline{\text{PM}}_e)^2}{\sum_{i=1}^N (\text{PM}_{t,i} - \overline{\text{PM}}_t)^2 \sum_{i=1}^N (\text{PM}_{e,i} - \overline{\text{PM}}_e)^2} \quad (19)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^N (\text{PM}_{t,i} - \text{PM}_{e,i})^2}{N}} \quad (20)$$

According to these definitions, the smaller the values of these metrics are except R^2 , the more accurate the estimation of $\text{PM}_{2.5}$ concentrations is. In particular, R^2 is no greater than 1. The value of R^2 is closer to 1, which means that the estimated $\text{PM}_{2.5}$ concentrations fit the true $\text{PM}_{2.5}$ concentrations better. Eventually, the reconstruction performance of sensor placement algorithms was assessed by the mean of ε , R^2 , and RMSE of validation snapshots.

3. RESULTS AND DISCUSSION

3.1. Performance of Five Sensor Placement Algorithms

Here, we assess five POD-based sensor placement algorithms (i.e., Random, MCN, Extrema, QR pivot, and DEIM) in estimating the ground-level distribution of $\text{PM}_{2.5}$ across China. To minimize the bias raised by randomization in Random and

Table 2. Performance of Sensor Placement Using Sensors Located by QR Pivot and Existing CNEMC Sites; Results of Both CAQRA and LGHAP Data Sets Are Presented

sensors	CAQRA			LGHAP		
	ϵ	R^2	RMSE	ϵ	R^2	RMSE
QR pivot	0.27	0.87	10.60	0.22	0.81	11.05
existing sites	0.98	0.37	37.72	0.36	0.60	17.90

Extrema algorithms, we take the average of 50 independent runs in each round of cross-validation. Due to the extremely expensive computation of MCN, we conduct a simplified version which traverses all possible sensor locations at once and picks the first P positions with minimum values of $\kappa(\mathbf{M})$. Table 1 summarizes 10-fold CV (10-CV) results of the CAQRA data set generated by different sensor placement algorithms without modal truncation, that is, using full-state modes, $r = 657$ (the number of training snapshots). Three conditions are considered, namely, $P = K$ (sensor quantity equals to the number of modes used in reconstruction), $P = 1.5K$, and $P = 2K$ (number of sensors exceeds the number of modes used). Under the $P = K$ condition, reconstruction with sensors generated by Random, MCN, and Extrema yields relatively poor evaluation metrics, with reconstruction error greater than 40, $R^2 < 0.03$, and $\text{RMSE} > 1000$ (Table 1). DEIM and QR achieve better reconstructions under the $P = K$ condition as indicated by large declines in reconstruction error and notable enhancement in R^2 .

Although QR pivot offers the best reconstruction under the $P = K$ condition, its performance is still unacceptable. The poor performance is likely to be associated with the relatively small number of used sensors ($P = 657$). With growing sensor quantity, we observe significant improvements in results using all the algorithms under conditions of both $P = 1.5K$ and $P = 2K$ (Table 1). Note that the DEIM algorithm generates sensor locations whose quantity equals to the dimension of POD basis, and this feature limits the occurrences of oversampling. Among other four algorithms, QR pivot exhibits distinct advantages and achieves the best performance, with the lowest error ϵ of 0.28, the highest R^2 value of 0.84, and the lowest RMSE value of $10.98 \mu\text{g}/\text{m}^3$ under the $P = 2K$ (~ 1300 sensors) condition.

Figure 1 displays spatial distribution of $\text{PM}_{2.5}$ concentrations over China from CAQRA (e) and reconstructed ones using Random (a2), MCN (b2), Extreme (c2), and QR pivot (d2)

algorithms under the $P = 2K$ condition. These four algorithms generally produce satisfactory reconstruction under such an oversampled condition. Nevertheless, the Extrema algorithm underestimates $\text{PM}_{2.5}$ pollution in eastern China, and it produces more hotspots of $\text{PM}_{2.5}$ pollution in northwestern China, compared with ground truth (Figure 1). The Random, MCN, and QR pivot algorithms achieve better reconstructions with respect to spatial distribution, consistent with the evaluation metrics displayed in Table 1. The corresponding sensor locations generated by these four algorithms are shown in the left column of Figure 1. The sensor locations generated by the Random and MCN algorithms are evenly dispersed across China, while the Extrema algorithm distributes sensors densely in an approximately rectangular area. Such a clustering pattern was also mentioned previously,⁴² which is likely to be caused by proximity or coincidence of the locations of the extrema of POD modes.⁴² Different from the unreasonably dense distribution of Extrema-generated sensor locations, QR pivot enables the sensors relatively concentrated in northern and northwest China, where $\text{PM}_{2.5}$ pollution is usually severe (Figure 1). Potential impacts brought by seasonal variability are shown in the Supporting Information.

3.2. Reconstruction Accuracy Varies with Sensor Quantity and POD Modes

Normally, dimension reduction is performed in POD-based reconstruction through the truncation of POD modes to obtain a low-rank reduced-order model. The first K most dominant modes are usually retained to build a low-rank approximation. Accordingly, the number of POD modes K is essential with respect to reconstruction quality. Here, we explore the interplay between POD basis, sensor quantity, and sensor locations on reconstruction accuracy. Figure 2 illustrates how reconstruction errors ϵ change along with the dimension of POD modes when sensor quantity is fixed at (a) 100, (b) 300, (c) 600, (d) 900, (e) 1200, and (f) 1500. Under $P=K$ conditions, the reconstruction errors of the Random, MCN, and Extrema algorithms reach peaks, distinctly greater than those of DEIM and QR pivot (Figure 2a–c). Additionally, QR pivot tends to be superior to DEIM, as indicated by lower errors under $P=K$ conditions (Figure 2a–c). Under $P < K$ conditions, reconstruction errors of these algorithms are generally greater than 1.0 (Figure 2a,b). We observe that the Random, MCN, and Extrema algorithms achieve the worst performance under the $P = K$ condition, even worse than their

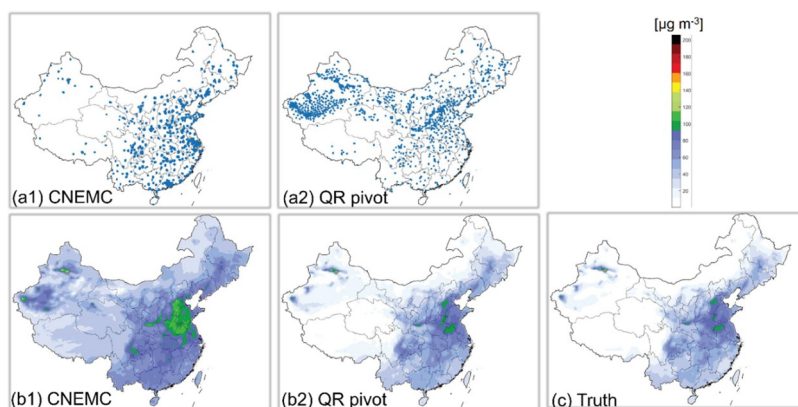


Figure 3. Locations of sensors and reconstructed $\text{PM}_{2.5}$ concentrations across China using existing CNEMC sites (a1,b1) and QR pivot algorithm (a2,b2) and CAQRA mean $\text{PM}_{2.5}$ of evaluation snapshots in one cross-validation round as ground truth (c).

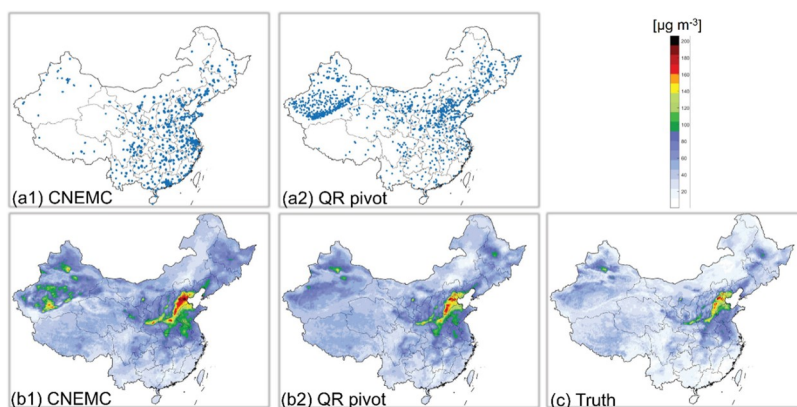


Figure 4. Locations of sensors and reconstructed $\text{PM}_{2.5}$ concentrations across China using existing CNEMC sites (a1,b1) and QR pivot algorithm (a2,b2) and LGHAP mean $\text{PM}_{2.5}$ of evaluation snapshots in one cross-validation round as ground truth (c).

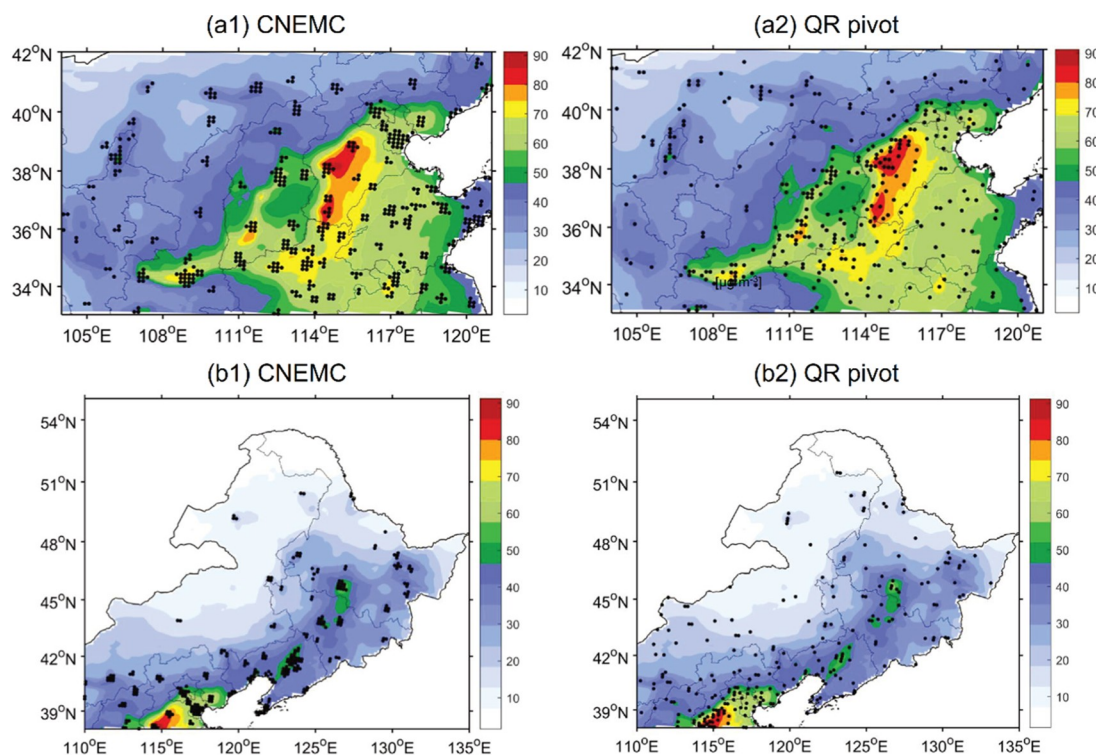


Figure 5. Locations of CNEMC and QR pivot-derived sensors and CAQRA $\text{PM}_{2.5}$ in northern China (a1,a2) and northeastern China (b1,b2).

performances under under-sampling ($P < K$) conditions. These results suggest that these three algorithms are relatively unstable, compared to the DEIM and QR pivot algorithms.

Under oversampling conditions ($P > K$), we observe a marginally upward trend of errors along with the increasing number of POD modes (Figure 2c). Lower errors are obtained by these algorithms under highly oversampling conditions (e.g., $K = 100$, $P = 600$), compared to marginal oversampling (e.g., $K = 500$, $P = 600$). More modes are involved in the latter, and sensor location algorithms are thus more likely to decide locations based on less-informative modes, especially when the number of sensors is relatively small. In general, the fewer the number of the POD modes, the better performance can be achieved for most sensor placement methods, when available sensors are not sufficient.

It is worth noting that the actual available sensor quantity in DEIM equals to the number of modes and does not exactly

correspond to the sensor quantity displayed in each subfigure (Figure 2). Therefore, the errors of DEIM are consistent in all subfigures. When we increase the sensor quantity to 900, 1200, and 1500, a substantial gap between DEIM and other algorithms appears (Figure 2). This signifies that a large increase in the number of sensors is able to significantly improve reconstruction abilities of these algorithms, except DEIM whose sensor quantity cannot be arbitrarily set. We also find in Figure 2c–f that the error of Extrema shows an upward trend when the dimension of POD increases. However, the dimension of modes does not exert notable influences on the performances of Random, MCN, and QR pivot, when the number of sensors is sufficient. More precisely, we find that QR pivot achieves marginally better approximation with the increased number of POD modes (Figure 2e,f), which is able to take advantage of the low-energy information contributed by additional features. On the whole, QR pivot algorithm can

yield competitive and stable reconstruction results under most conditions.

3.3. Comparison with Reconstruction Using Existing Sites

We further compare the reconstruction accuracy using QR pivot-guided sensors (sites) ($K = 600$ and $P = 1500$) and existing CNEMC sites (1583 sites), and the results with both CAQRA and LGHAP data sets as ground truth are listed in Table 2. Compared with the reconstruction results using existing CNEMC sites, QR pivot sensors (sites) yield better reconstruction by a large margin in the case of fewer sensors. As shown in Figure 3, QR pivot sensors offer higher degree agreements with ground truth, while existing sites result in an extensive overestimation in northwestern China. Similar agreements and differences are found when we use the LGHAP data set as ground truth (Table 2 and Figure 4). As indicated by Figure 5a1,b1, most existing CNEMC stations are concentrated in densely populated urban regions in eastern China, leading to missing information in other polluted suburban or rural regions. Instead, QR pivot sensors are more evenly distributed in populated eastern China and densely distributed in northeastern China (Figure 5).

We notice that notable discrepancies exist between the two data sets as ground truth with respect to magnitudes of $PM_{2.5}$ concentrations. Yet the spatial distributions of pollution hotspots are generally consistent. The CAQRA data set was derived with chemical transport modeling with assimilated surface observations of air pollutants across China, while the LGHAP data were inferred by combining satellite AOD and associated variables. Satellite AOD offers only daytime products, and results are relatively less reliable in northern regions that affected by ice cover. Accordingly, differences occur in the magnitudes of $PM_{2.5}$. Instead of locating monitoring sites empirically before, we could determine optimal site locations using theoretical guidance by QR pivot to better estimate the spatial distribution of $PM_{2.5}$ pollution across China. Although ground truth of gap-free variations of $PM_{2.5}$ does not exist, the inferred locations of sites using two data sets are generally consistent. This further suggests that we are able to demonstrate the rationality of the proposed framework when we use reasonable data sets as inputs.

4. CONCLUSIONS

Numerous studies attempted to use ground-level observations to obtain a better depiction of spatiotemporal variations of $PM_{2.5}$, in support of air quality management and impact studies. Statistical methods (machine learning, etc.) or numerical methods by combing chemical transport modeling and observations with data assimilation techniques are typically applied, yet the significance of site placement has not been well considered in these studies. In this work, we seek to derive a spatially complete distribution of $PM_{2.5}$ across China using ground-level site measurements. We applied five POD-based sensor placement algorithms to identify optimal site locations and systematically compared their reconstruction ability. Additionally, we discussed how sensor quantity, locations, and dimension of modes affect the accuracy of reconstruction. We demonstrated that QR pivot is relatively more reliable in deciding optimal monitoring site locations. When the number of planned sites (sensors) is limited, using lower number of modes would yield lower estimation errors. However, the dimension of POD modes has little impact on reconstruction quality when sufficient sensors are available. The locations of

sites guided by QR pivot algorithm are mainly located in regions where $PM_{2.5}$ pollution is severe. We compared the reconstruction of $PM_{2.5}$ pollution based on QR pivot-guided sites and existing CNEMC sites, and we found that QR pivot-guided sites are superior to existing sites with respect to reconstruction accuracy.

The current planning of monitoring stations is likely to miss sources of pollution in less-populated regions, while our QR pivot-guided sites are planned based on the severity of $PM_{2.5}$ pollution. This planning methodology has additional potentials in chemical data assimilation studies as duplicate information from concentrated stations is not likely to boost the performance. Previously, we used CNEMC-observed $PM_{2.5}$ to constrain modeled spatiotemporal variations of $PM_{2.5}$ in northern China with a three-dimensional variational data assimilation system.¹⁸ We used only one station in each city to avoid duplicate information, while observation constraints over suburban and rural regions were not sufficient due to lack of monitoring stations. With a better plan of placing monitoring sites, these issues would be better addressed. Furthermore, as OSP is a pervasive problem in various fields, such as ocean modeling, fluid flow control, and aerodynamics, our work also offers inspiration and guidance for selecting sensor locations in other disciplines. In practical implementation, cost is an important concern as the costs of site construction and maintenance vary across locations due to terrain, environment, and other factors. Therefore, how to make a cost-effective planning of monitoring sites needs further investigations.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acsenvironau.1c00051>.

Model performance metric R^2 values varying with seasons (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Meng Gao – Department of Geography, Faculty of Social Sciences, Hong Kong Baptist University, Hong Kong SAR 999077, China; orcid.org/0000-0002-8657-3541;
Email: mmgao2@hkbu.edu.hk

Yike Guo – Department of Computer Science, Faculty of Science, Hong Kong Baptist University, Hong Kong SAR 999077, China; Email: yikeguo@hkbu.edu.hk

Authors

Chenhong Zhou – Department of Computer Science, Faculty of Science, Hong Kong Baptist University, Hong Kong SAR 999077, China

Jianjun Li – China National Environmental Monitoring Center, Beijing 100012, China

Kaixu Bai – Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University, Shanghai 200241, China; orcid.org/0000-0002-6529-712X

Xiao Tang – State Key Laboratory of Atmospheric Boundary Layer Physics and Atmospheric Chemistry, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029, China; orcid.org/0000-0003-1181-2103

Xiao Lu – School of Atmospheric Sciences, Sun Yat-sen University, Zhuhai 519082 Guangdong Province, China

Cheng Liu – Department of Precision Machinery and Precision Instrumentation, University of Science and Technology of China, Hefei 230026 Anhui, China

Zifa Wang – State Key Laboratory of Atmospheric Boundary Layer Physics and Atmospheric Chemistry, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029, China

Complete contact information is available at:

<https://pubs.acs.org/10.1021/acsenvironau.1c00051>

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This study was supported by grants from Research Grants Council of the Hong Kong Special Administrative Region, China (project no. HKBU22201820 and HKBU12202021) and National Natural Science Foundation of China (No. 42005084).

ABBREVIATION

Dimensions

N dimension of the state

P number of sensors (measurements/interpolation points)

K number of POD modes

Q number of snapshots

Matrices

$\mathbf{U} \in \mathbb{R}^{N \times Q}$ data matrix with Q snapshots

$\Phi \in \mathbb{R}^{N \times K}$ matrix of POD modes

$\mathbf{C} \in \mathbb{R}^{P \times N}$ measurement matrix (matrix representation form of \mathbf{m})

$\Theta = \mathbf{C}\Phi$ product of measurement matrix and basis matrix

$\mathbf{M} \in \mathbb{R}^{K \times K}$ $\Theta^T\Theta$ or the gappy inner product of Φ^T and Φ

$\mathbf{D} \in \mathbb{R}^{N \times K}$ interpolation matrix in DEIM

\mathbf{Q} unitary matrix from QR factorization

\mathbf{R} upper triangular matrix from QR factorization

$\Phi_K \in \mathbb{R}^{N \times K}$ tailored basis of rank K in QR column pivoting

$\mathbf{D} \in \mathbb{R}^{N \times N}$ permutation matrix in QR column pivoting

Vectors

$\mathbf{u} \in \mathbb{R}^N$ data vector

$\mathbf{a} \in \mathbb{R}^K$ POD coefficient vector

$\Phi^k \in \mathbb{R}^N$ POD basis (k th column of Φ)

$\tilde{\mathbf{u}} \in \mathbb{R}^N$ Gappy vector

$\mathbf{m} \in \mathbb{R}^N$ mask vector

$\mathbf{b} \in \mathbb{R}^K$ unknown POD coefficient vector

$\tilde{\mathbf{u}}_\zeta \in \mathbb{R}^N$ reconstruction of $\tilde{\mathbf{u}}$

$\mathbf{f} \in \mathbb{R}^K$ Gappy inner product of $\tilde{\mathbf{u}}$ and Φ (Section 2.1)

$\mathbf{c} \in \mathbb{R}^K$ coefficient vector

$\mathbf{y} \in \mathbb{R}^P$ observations/measurements

REFERENCES

- (1) Song, W.; Jia, H.; Huang, J.; Zhang, Y. A satellite-based geographically weighted regression model for regional PM_{2.5} estimation over the Pearl River Delta region in China. *Remote Sens. Environ.* **2014**, *154*, 1–7.
- (2) Geng, G.; Zhang, Q.; Martin, R. V.; van Donkelaar, A.; Huo, H.; Che, H.; Lin, J.; He, K. Estimating long-term PM_{2.5} concentrations

in China using satellite-based aerosol optical depth and a chemical transport model. *Remote Sens. Environ.* **2015**, *166*, 262–270.

(3) Zhang, Y.; Li, Z. Remote sensing of atmospheric fine particulate matter (PM_{2.5}) mass concentration near the ground from satellite observation. *Remote Sens. Environ.* **2015**, *160*, 252–262.

(4) Peng, J.; Chen, S.; Lü, H.; Liu, Y.; Wu, J. Spatiotemporal patterns of remotely sensed PM_{2.5} concentration in China from 1999 to 2011. *Remote Sens. Environ.* **2016**, *174*, 109–121.

(5) Pope III, C. A.; Iii, Burnett, R. T.; Thun, M. J.; Calle, E. E.; Krewski, D.; Ito, K.; Thurston, G. D. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *Jama* **2002**, *287*, 1132–1141.

(6) Gauderman, W. J.; Avol, E.; Gilliland, F.; Vora, H.; Thomas, D.; Berhane, K.; McConnell, R.; Kuenzli, N.; Lurmann, F.; Rappaport, E.; Margolis, H.; Bates, D.; Peters, J. The effect of air pollution on lung development from 10 to 18 years of age. *N. Engl. J. Med.* **2004**, *351*, 1057–1067.

(7) Hoek, G.; Krishnan, R. M.; Beelen, R.; Peters, A.; Ostro, B.; Brunekreef, B.; Kaufman, J. D. Long-term air pollution exposure and cardio-respiratory mortality: a review. *Environ. Health* **2013**, *12*, 43.

(8) Pascal, M.; Falq, G.; Wagner, V.; Chatignoux, E.; Corso, M.; Blanchard, M.; Host, S.; Pascal, L.; Larrieu, S. Short-term impacts of particulate matter (PM₁₀, PM_{10-2.5}, PM_{2.5}) on mortality in nine French cities. *Atmos. Environ.* **2014**, *95*, 175–184.

(9) Liu, C.; Gao, M.; Hu, Q.; Brasseur, G. P.; Carmichael, G. R. Stereoscopic Monitoring: A Promising Strategy to Advance Diagnostic and Prediction of Air Pollution. *Bull. Am. Meteorol. Soc.* **2021**, *102*, E730–E737.

(10) Liang, F.; Gao, M.; Xiao, Q.; Carmichael, G. R.; Pan, X.; Liu, Y. Evaluation of a data fusion approach to estimate daily PM_{2.5} levels in North China. *Environ. Res.* **2017**, *158*, 54–60.

(11) Zhang, N.-N.; Ma, F.; Qin, C.-B.; Li, Y.-F. Spatiotemporal trends in PM_{2.5} levels from 2013 to 2017 and regional demarcations for joint prevention and control of atmospheric pollution in China. *Chemosphere* **2018**, *210*, 1176–1184.

(12) Wang, W.; Zhao, S.; Jiao, L.; Taylor, M.; Zhang, B.; Xu, G.; Hou, H. Estimation of PM_{2.5} concentrations in China using a spatial back propagation neural network. *Sci. Rep.* **2019**, *9*, 13788.

(13) Xue, W.; Wei, J.; Zhang, J.; Sun, L.; Che, Y.; Yuan, M.; Hu, X. Inferring Near-Surface PM_{2.5} Concentrations from the VIIRS Deep Blue Aerosol Product in China: A Spatiotemporally Weighted Random Forest Model. *Remote Sens.* **2021**, *13*, 505.

(14) Li, T.; Shen, H.; Zeng, C.; Yuan, Q.; Zhang, L. Point-surface fusion of station measurements and satellite observations for mapping PM_{2.5} distribution in China: Methods and assessment. *Atmos. Environ.* **2017**, *152*, 477–489.

(15) Zang, Z.; Li, D.; Guo, Y.; Shi, W.; Yan, X. Superior PM_{2.5} Estimation by Integrating Aerosol Fine Mode Data from the Himawari-8 Satellite in Deep and Classical Machine Learning Models. *Remote Sens.* **2021**, *13*, 2779.

(16) Fang, X.; Zou, B.; Liu, X.; Sternberg, T.; Zhai, L. Satellite-based ground PM_{2.5} estimation using timely structure adaptive modeling. *Remote Sens. Environ.* **2016**, *186*, 152–163.

(17) Zhang, X.; Wang, H.; Che, H.-Z.; Tan, S.-C.; Shi, G.-Y.; Yao, X.-P.; Zhao, H.-J. Improvement of snow/haze confusion data gaps in MODIS Dark Target aerosol retrievals in East China. *Atmos. Res.* **2020**, *245*, 105063.

(18) Gao, M.; Saide, P. E.; Xin, J.; Wang, Y.; Liu, Z.; Wang, Y.; Wang, Z.; Pagowski, M.; Guttikunda, S. K.; Carmichael, G. R. Estimates of Health Impacts and Radiative Forcing in Winter Haze in Eastern China through Constraints of Surface PM_{2.5} Predictions. *Environ. Sci. Technol.* **2017**, *51*, 2178–2185.

(19) Carmichael, G. R.; Sandu, A.; Chai, T.; Daescu, D. N.; Constantinescu, E. M.; Tang, Y. Predicting air quality: Improvements through advanced methods to integrate models and measurements. *J. Comput. Phys.* **2008**, *227*, 3540.

(20) Ancell, B.; Hakim, G. J. Comparing Adjoint- and Ensemble-Sensitivity Analysis with Applications to Observation Targeting. *Mon. Weather Rev.* **2007**, *135*, 4117–4134.

- (21) Manohar, K.; Brunton, B. W.; Kutz, J. N.; Brunton, S. L. Data-driven sparse sensor placement for reconstruction: Demonstrating the benefits of exploiting known patterns. *IEEE Control Syst. Mag.* **2018**, *38*, 63–86.
- (22) Rempfer, D. Low-dimensional modeling and numerical simulation of transition in simple shear flows. *Annu. Rev. Fluid. Mech.* **2003**, *35*, 229–265.
- (23) Berkooz, G.; Holmes, P.; Lumley, J. L. The proper orthogonal decomposition in the analysis of turbulent flows. *Annu. Rev. Fluid. Mech.* **1993**, *25*, 539–575.
- (24) Aubry, N.; Guyonnet, R.; Lima, R. Spatiotemporal analysis of complex signals: theory and applications. *J. Stat. Phys.* **1991**, *64*, 683–739.
- (25) Sirovich, L. Turbulence and the dynamics of coherent structures, Parts I, II and III. *Quart. Appl. Math.* **1987**, *45*, 561–571.
- (26) Thomson, R. E.; Emery, W. J. *Data Analysis Methods in Physical Oceanography*; Newnes, 2014.
- (27) Zhang, Y.; Bellingham, J. G. An efficient method of selecting ocean observing locations for capturing the leading modes and reconstructing the full field. *J. Geophys. Res.: Oceans* **2008**, *113*, C04005.
- (28) Everson, R.; Sirovich, L. Karhunen–Loeve procedure for gappy data. *J. Opt. Soc. Am.* **1995**, *12*, 1657–1664.
- (29) Willcox, K. Unsteady flow sensing and estimation via the gappy proper orthogonal decomposition. *Comput. Fluids* **2006**, *35*, 208–226.
- (30) Yildirim, B.; Chrysostomidis, C.; Karniadakis, G. Efficient sensor placement for ocean measurements using low-dimensional concepts. *Ocean Model.* **2009**, *27*, 160–173.
- (31) Al Mamun, S.; Lu, C.; Jayaraman, B. Extreme learning machines as encoders for sparse reconstruction. *Fluids* **2018**, *3*, 88.
- (32) Jayaraman, B.; Al Mamun, S. M. A.; Lu, C. Interplay of sensor quantity, placement and system dimension in pod-based sparse reconstruction of fluid flows. *Fluids* **2019**, *4*, 109.
- (33) Cohen, K.; Siegel, S.; McLaughlin, T. Sensor placement based on proper orthogonal decomposition modeling of a cylinder wake. *33rd AIAA Fluid Dynamics Conference and Exhibit*, 2003; p 4259.
- (34) Chaturantabut, S.; Sorensen, D. C. Nonlinear model reduction via discrete empirical interpolation. *SIAM J. Sci. Comput.* **2010**, *32*, 2737–2764.
- (35) Jayaraman, B.; Mamun, S. M. A. A. On data-driven sparse sensing and linear estimation of fluid flows. *Sensors* **2020**, *20*, 3752.
- (36) Brunton, S. L.; Kutz, J. N. *Data-driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*; Cambridge University Press, 2019.
- (37) Mokhasi, P.; Rempfer, D. Optimized sensor placement for urban flow measurement. *Phys. Fluids* **2004**, *16*, 1758–1764.
- (38) Barrault, M.; Maday, Y.; Nguyen, N. C.; Patera, A. T. An ‘empirical interpolation’ method: application to efficient reduced-basis discretization of partial differential equations. *C. R. Math.* **2004**, *339*, 667–672.
- (39) Clenet, S.; Henneron, T.; Korecki, J. Sensor placement for field reconstruction in rotating electrical machines. *IEEE Trans. Magn.* **2021**, *57*, 1–4.
- (40) Kong, L.; Tang, X.; Zhu, J.; Wang, Z.; Li, J.; Wu, H.; Wu, Q.; Chen, H.; Zhu, L.; Wang, W.; Liu, B.; Wang, Q.; Chen, D.; Pan, Y.; Song, T.; Li, F.; Zheng, H.; Jia, G.; Lu, M.; Wu, L.; Carmichael, G. R. A 6-year-long (2013–2018) high-resolution air quality reanalysis dataset in China based on the assimilation of surface observations from CNEMC. *Earth Syst. Sci. Data* **2021**, *13*, 529–570.
- (41) Bai, K. L.; Ke, Tan, Z.; Han, D.; Guo, J. *Daily 1-km Gap-free PM_{2.5} Grids in China, VI, 2000–2020*. [Data set].
- (42) Yang, X.; Venturi, D.; Chen, C.; Chrysostomidis, C.; Karniadakis, G. E. EOF-based constrained sensor placement and field reconstruction from noisy ocean measurements: Application to Nantucket Sound. *J. Geophys. Res.: Oceans* **2010**, *115*, C12072.