The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1064–1089
https://doi.org/10.1214/23-AOAS1823

EFFICIENT AND EFFECTIVE CALIBRATION OF NUMERICAL MODEL OUTPUTS USING HIERARCHICAL DYNAMIC MODELS

BY YEWEN CHEN[1,a], XIAOHUI CHANG[2,b], BOHAI ZHANG[3,c] AND HUI HUANG[4,d]

[1]*College of Public Health, University of Georgia,* [a]*Yewen.Chen@uga.edu*
[2]*College of Business, Oregon State University,* [b]*xiaohui.chang@oregonstate.edu*
[3]*Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science, BNU-HKBU United International College,* [c]*bohaizhang@uic.edu.cn*
[4]*Center for Applied Statistics and School of Statistics, Renmin University of China,* [d]*huangh89@mail.sysu.edu.cn*

Numerical air quality models, such as the Community Multiscale Air Quality (CMAQ) system, play a critical role in characterizing pollution levels at fine spatial and temporal scales. The model outputs, however, tend to systematically over- or underestimate the real pollutant concentrations. In this study we propose a Bayesian hierarchical dynamic model to calibrate large-scale grid-level CMAQ model outputs using data from other sources, especially point-level observations from sparsely located monitoring stations. In our model a stochastic integro-differential equation (IDE) is implemented to account for space-time interactions of air pollutants. To better approximate the spatial pattern of pollutants, we employ nonregular meshes to discretize IDEs. A spatial partitioning procedure is embedded to improve the scalability of the approach for very large meshes. An algorithm based on variational Bayes and ensemble Kalman smoother is developed to accelerate the parameter estimation and calibration procedure. We apply the proposed approach to calibrate CMAQ outputs for China's Beijing–Tianjin–Hebei region. In contrast to existing methods, the proposed approach captures space-time interactions, produces more accurate calibration results, and operates at a higher computational efficiency. A reanalysis dataset is also adopted to demonstrate the effectiveness and efficiency of our approach to large spatial data.

1. Introduction. High-quality and high-resolution air pollution maps are pivotal to the assessments of regional air pollution levels and emission control strategies (Qi et al. (2017), Guan et al. (2020)). In recent years numerical air quality systems, such as the Community Multiscale Air Quality (CMAQ) (Byun and Schere (2006)) or the Nested Air Quality Prediction Modeling System (NAQPMS) (Wang et al. (2006)), coupled with meteorological and emission models, have been widely used in air pollution forecasting. Pollutant concentrations in horizontally regular grids are usually estimated by solving large hydrodynamic equations, running chemical transport models, and embedding many other components (Byun and Schere (2006), Wang et al. (2006), Appel et al. (2017)). These complex numerical model systems are valuable for forecasting large spatial domains with no missing values but often introduce prediction bias due to errors in input parameters or initial conditions, especially under extreme conditions; see Vannitsem et al. (2021) and references therein. As an illustration, Figures 1(c), (d), and (e) compare CMAQ $PM_{2.5}$ (i.e., fine particulate matters with aerodynamic diameters less than 2.5 micrometers) forecasts with the actual observations in China's Beijing–Tianjin–Hebei (BTH) region. CMAQ forecasts systematically overestimate $PM_{2.5}$ concentrations in Zhangjiakou while underestimating the concentrations in Beijing and Hengshui.

Received February 2023; revised August 2023.
Key words and phrases. Calibration, numerical model outputs, hierarchical dynamic models, stochastic integro-differential equations, variational Bayes, space-partitioning-based ensemble Kalman smoother.

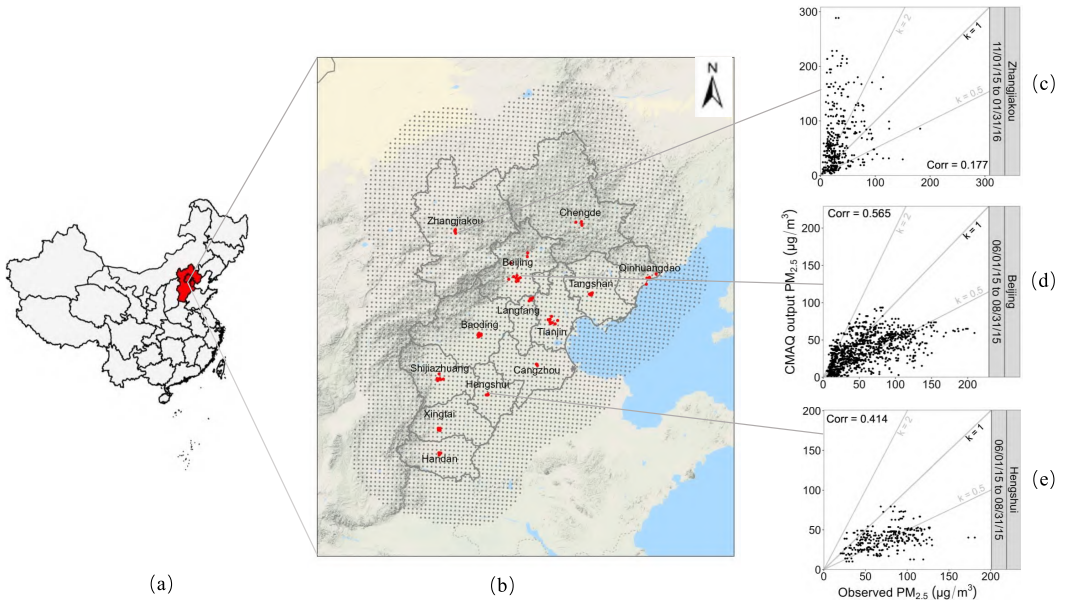


FIG. 1. (*a*) *Map of China*. (*b*) *Zoomed-in map of the BTH region along with locations of* 68 *monitoring stations* (*red dots*) *and centroids of* 5587 *nine-km CMAQ grids* (*gray dots*). (*c*)–(*e*) *Scatter plots of CMAQ* $PM_{2.5}$ *forecasts vs. actual observations at Zhangjiakou*, *Beijing*, *and Hengshui in different seasons*, *where "Corr" represents the Pearson correlation coefficient between CMAQ* $PM_{2.5}$ *outputs and observed* $PM_{2.5}$ *concentrations*. *Three reference lines with the slope* $k = 0.5, 1,$ *and* 2 *are colored in gray.*

To improve forecast accuracy, methods, such as calibration in statistics or data assimilation in geosciences, are often applied to fuse data from multiple sources. The most important data source is readings retrieved from monitoring networks; see Fuentes and Raftery (2005), Guillas et al. (2008), Berrocal, Gelfand and Holland (2010). Although monitoring stations are usually distributed sparsely in space and clustered in urban areas, observations collected from the stations reflect the actual pollution on a local scale and provide a reliable reference for regional calibrations (Berrocal, Gelfand and Holland (2010)). Early studies, such as Kennedy and O'Hagan (2001), proposed a Bayesian calibration method to use physical observations to adjust the input parameters of computer models; see Higdon et al. (2008) and Salter et al. (2019) for more recent works. Different from these studies that calibrate the input parameters of models, we are primarily interested in bias-correcting raw outputs of numerical models using observed data (Guillas et al. (2008)).

To connect the gridded outputs of the numerical models to the readings collected at the monitoring stations, Fuentes and Raftery (2005) assumed a "true" latent point-level process that drives both the grid-level numerical outputs and the point-level station observations and developed a "Bayesian melding" approach by which the bias in numerical model outputs can be estimated and corrected. A spatiotemporal version of this model was formulated in McMillan et al. (2010) where CMAQ $PM_{2.5}$ outputs were calibrated using Markov chain Monte Carlo (MCMC) techniques that become computationally difficult for large-scale datasets. Another type of calibration technique is the "downscaling" method (Berrocal, Gelfand and Holland (2010), Berrocal, Gelfand and Holland (2012)), which utilizes the framework of spatially varying coefficient (SVC) models developed in Gelfand et al. (2003). In downscaling models the numerical model outputs are the explanatory variables and are "downscaled" to explain point-level observations. Under a Bayesian setting, the varying coefficients are first fitted via the "coregionalization" of Gaussian processes. The calibrations in each grid cell are then simulated and integrated from posterior distributions; see Berrocal, Gelfand and Holland (2010) for more details.

Although the techniques described earlier have been successfully applied in many environmental and atmospheric studies (e.g., Shaddick et al. (2018), Jiang and Yoo (2019)), improving the calibration accuracy and computational efficiency remains challenging, especially for large spatiotemporal datasets. These challenges are similar to those that arise in statistical postprocessing (Vannitsem et al. (2021)). The various factors, such as the physical processes of pollutants, geographical factors, and others, complicate the spatiotemporal patterns of pollutant concentrations and call for more flexible methodologies than the standard ones. The computational cost of the Bayes methods also hinders more broad applications in the era of big data. Several strategies have been adopted to reduce computational complexities, including fixed-rank approximations (Cressie and Johannesson (2008)), sparse-matrix approximations (Furrer, Genton and Nychka (2006), Lindgren, Rue and Lindström (2011), Datta et al. (2016)), Gaussian predictive processes (Banerjee et al. (2008), Zhang, Sang and Huang (2015)), and others. Compared to MCMC methods that are based on sampling, variational Bayes (VB) offers a computationally efficient alternative and approximates the posterior distribution using distributed or stochastic optimization, making it both faster and simpler to monitor for convergence (Ishiguro, Sato and Ueda (2017)). Therefore, VB can be easily scaled up to massive data (see Blei, Kucukelbir and McAuliffe (2017) for a recent review) and has been successfully applied to several applications such as inference of network data (Tabouy, Barbillon and Chiquet (2020)) and spatial data analysis (Ren et al. (2011)). However, in the context of calibration, the performance of all the aforementioned methods has yet to be fully investigated.

Within the geoscience community, ensemble-based methods are developed to minimize the computational burden of Kalman filter (KF) and Kalman smoother (KS) when estimating latent spatiotemporal processes. Notably, ensemble KF (EnKF) (Evensen (1994)) and ensemble KS (EnKS) (Evensen and Van Leeuwen (2000)) have been widely employed in numerous applications (Houtekamer et al. (2005), Stroud et al. (2010)). Using ensemble-based methods, the propagation of an ensemble is repeatedly updated with new observations via predetermined iterative algorithms. In practice, the processes are typically confined to some spatial grids. However, the ensemble size is generally smaller than the dimension of spatial grids, resulting in rank deficiencies and inducing spurious correlations (Katzfuss, Stroud and Wikle (2016)). In addition, EnKF and EnKS also face computational difficulties when both the observed data and spatial grids become very large.

In this work we propose a novel hierarchical dynamic calibration model (HDCM) to improve the raw outputs of the numerical models. Two approximation algorithms, VB and space-partitioning-based EnKS (spEnKS), are also developed to accelerate parameter estimation and calibration. Using stochastic integro-difference equation (IDE) to construct calibration models, HDCM accounts for spatiotemporal interactions of air pollution. In particular, through the use of a triangulation scheme, the discretized stochastic IDEs are able to model the spatiotemporal variations of the data more flexibly than the continuous IDEs that allow processes to be random in time alone (e.g., Xu, Wikle and Fox (2005), Richardson, Kottas and Sansó (2017)). To further improve the scalability of HDCM, a Laplace approximation and a space-partitioning-based procedure are embedded in VB and EnKS, respectively. Our approach differs from most current calibration methods in that not only does our model capture the dynamic evolution of space-time interactions in air pollution, but it is computationally efficient for large datasets. We demonstrate the proposed method using two datasets. The first dataset consists of the observed $PM_{2.5}$ concentrations from the BTH stations and the outputs in the BTH region from CMAQ between 2015 and 2016. The second dataset arises from a reanalysis of $PM_{2.5}$ concentrations from NAQPMS and the outputs from CMAQ in 2015.

In Section 2 we present $PM_{2.5}$ datasets in the BTH region and demonstrate some specific data features. The proposed dynamic calibration model and its implementation issues are

described in Section 3. In Section 4 we compare our model with other commonly used calibration methods using different assessment criteria. In Section 5 we use the proposed model to calibrate CMAQ model outputs of the BTH region and produce pollution maps. In Section 6 we use the reanalysis data to illustrate the effectiveness and efficiency of the proposed method on large datasets. The paper concludes with a discussion in Section 7. Some technical details of the proposed approach can be found in Sections S9–S10 of the Supplementary Material (Chen et al. (2024)). The code for our approach and for the computations in this work are contained in the Supplementary Material, and both the data and the latest version of the code are publicly available on GitHub (https://github.com/ChenYW68/HDCM).

2. Data description of the BTH data. The air quality of North China has drawn worldwide attention since the record-high haze events in January 2013 (Zhang et al. (2016)). As one of the largest industrial bases in China, the BTH region has suffered from severe air pollution for decades. In recent years one of the most common and harmful pollutants in the area has been $PM_{2.5}$. With the launch of a series of movements, including the short-term Air Pollution Prevention and Control Action Plan between 2013 and 2017 (China's State Council (2013)), "Blue Sky Defense Battle" between 2018 and 2020 (China's State Council (2018)), and the long-term "Beautiful China" targets through 2035 (China's State Council (2021)), significant air quality improvements have been made. Taking Beijing as an example, since 2013, SO_2 emissions have fallen by 83%, NOx by 43%, VOCs by 42%, and $PM_{2.5}$ by 59% (Lu et al. (2020)). Despite these achievements pollution prevention and control in the BTH region remains extremely important, as this region consists of 13 cities with more than 100 million residents and accounts for nearly 9% of China's GDP, playing a critical role in the balance between public health and economic growth of the country.

In this work we consider two sources of $PM_{2.5}$ data from the BTH region. The first data are CMAQ model outputs at a *nine*-km scale with a total of 5587 grid cells that cover the entire region, where 2499 grid cells are located in the BTH region and the rest are in the area adjoining the BTH region. The second data are readings from 68 national monitoring stations, which are clustered mostly in urban areas, especially the city centers. For the two datasets, we used hourly $PM_{2.5}$ concentrations from one summer (between June 1, 2015 and August 31, 2015, 92 days in total) and one winter (between November 1, 2015 and January 31, 2016, 92 days in total). The locations of CMAQ grid cells and monitoring stations are indicated in Figure 1(b). Hourly observations from monitoring stations may contain missing data and extreme values from occasional maintenance errors. For convenience, we first averaged the hourly observations to daily data and then imputed missing values (about 0.28% of the data) using ordinary kriging (Cressie and Wikle (2011)). It is worth noting that the proposed approach is able to handle datasets with sites that are completely missing, as this approach allows imputing the missing data within VB-spEnKS. In contrast, several other non-Bayesian methods presented in Section 4 have difficulties in automatically imputing missing data within their implementation procedures. To ensure a fair comparison among different methods, instead of imputing the missing data through the VB-spEnKS procedure, we filled in the missing values at this stage.

As an illustration, we present the contours of empirical space-time covariance in Figure 2. For both seasons, the empirical covariance demonstrates a complex pattern that varies across different lags in space and/or time, especially when the time lags are more than one day. This suggests that spatiotemporal interaction should be considered in the calibration models. We have also shown the scatter plots of CMAQ model outputs vs. the actual observations from five cities in the BTH region in Figure 1(c)–(e) and Figure 3. The bias of CMAQ outputs

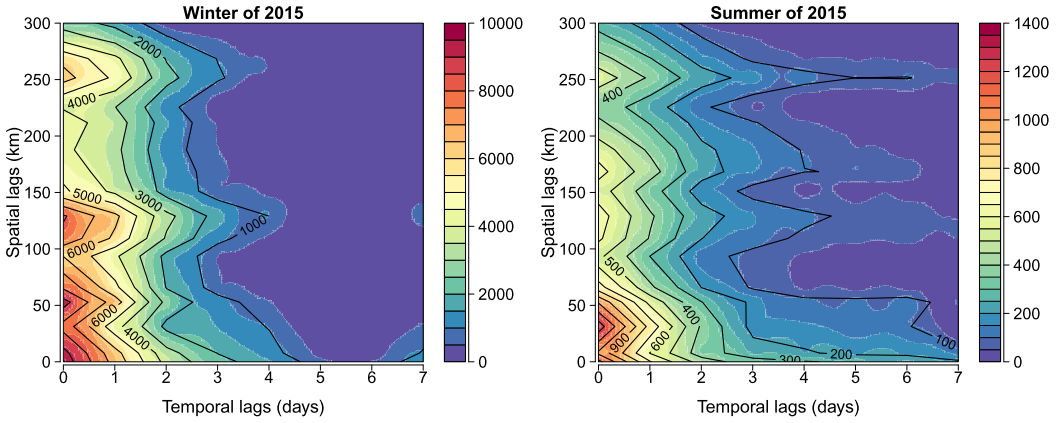


FIG. 2. *The interpolated surfaces and contour lines of empirical spatiotemporal covariances of the observed* $PM_{2.5}$ *concentrations in the BTH region from the winter* (*left*) *and the summer* (*right*).

varies across both locations and seasons. In addition, the winter $PM_{2.5}$ concentrations are generally more variable and higher than the summer concentrations, resulting in much worse air quality in winter; see Figure 3 and Table S1 of the Supplementary Material. In summary, it is necessary to include both geographical and meteorological information in the calibration models for a more accurate pollution assessment.

The actual $PM_{2.5}$ readings obtained from monitoring stations also display clear right skewness, especially in winter, mostly due to the large variation in concentrations (Liang et al. (2015)); see Figures S1(a) and (d) of the Supplementary Material. Following earlier studies (Sahu, Gelfand and Holland (2006)), we transform the $PM_{2.5}$ data, using both the square root and logarithmic transformations, and explore their distributions by histograms (Berrocal, Gelfand and Holland (2010)) and quantile-quantile (Q-Q) boxplots (Rodu and Kafadar (2022)). The histograms displayed in Figure S1 of the Supplementary Material show that the distributions of the square root-transformed data are closer to normal distributions than that of the log-transformed data. Additionally, the Q-Q boxplots in Figure S2 of the Supplementary Material demonstrate significant deviations from normal distributions in both the original and log-transformed data, with these deviations being particularly pronounced in the tails. Thus, throughout this study we model $PM_{2.5}$ concentrations on the square-root scale but present spatial predictions on the original scale for ease of interpretation and comparisons.

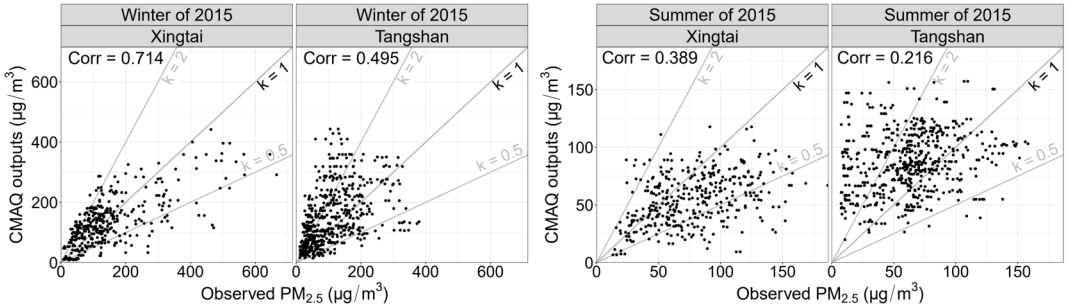


FIG. 3. *Scatter plots of CMAQ* $PM_{2.5}$ *predictions vs. actual observations at Xintai and Tangshan in different seasons, where "Corr" and three gray reference lines are as defined in Figure 1.*

3. Methodology.

3.1. *Hierarchical dynamic calibration model* (*HDCM*). To connect the point-level observations to the grid-level CMAQ outputs and eventually calibrate CMAQ outputs, we develop a hierarchical dynamic space-time model. Suppose $\boldsymbol{D}$ is the spatial domain of interest that consists of n point-level observations and is covered by m grid cells. For the BTH data, $n = 68$. The choice of m could be the dimension of the prespecified CMAQ grids and its extended area, that is, $m = 5587$, as seen in Figure 1(b), or the number of vertices in a reconstructed mesh, such as the triangulated mesh in Figure 4, that is, $m = 882$. Let $t = 1, \ldots, N_t$, where $N_t = 92$. We then denote an n-dimensional vector $\boldsymbol{y}_t$ as the observations at time t and assume that there is an n-dimensional true underlying Gaussian process $\boldsymbol{w}_t$ that drives the behavior of $\boldsymbol{y}_t$. The first level of our model is then

$$\boldsymbol{y}_t = \boldsymbol{w}_t + \boldsymbol{\varepsilon}_t, \tag{1}$$

where $\boldsymbol{\varepsilon}_t$ is a zero-mean white noise. We also assume $\boldsymbol{w}_t$ and $\boldsymbol{\varepsilon}_t$ are mutually and serially independent, that is, $\operatorname{cov}(w_t(\boldsymbol{s}), \varepsilon_{t'}(\boldsymbol{s}')) = 0$ for all spatiotemporal points $(\boldsymbol{s}, t)$ and $(\boldsymbol{s}', t')$ (similarly hereinafter). We further assume a Gaussian distribution for the error such that $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$, where $\sigma > 0$ and $\boldsymbol{I}_n$ is an $n \times n$ identity matrix. It is worth noting that this assumption can be relaxed to other distributions of $\boldsymbol{\varepsilon}_t$ with more complex covariance structures. From the implementation perspective, when the value of $n \times N_t$ is large, the computation efficiency of (1) is closely related to the model specification of the process $\boldsymbol{w}_t$. For example, the descriptive geostatistical approaches to model the process $\boldsymbol{w}_t$ usually face computational difficulties, as they require an inverse of an $(n \times N_t) \times (n \times N_t)$ matrix at each iteration (Wikle, Zammit-Mangion and Cressie (2019)).

To ensure the applicability of (1) for a very large matrix of size $(n \times N_t) \times (n \times N_t)$, the latent process $\boldsymbol{w}_t$ is further decomposed into two parts: an n-dimensional point-level mean term $\boldsymbol{\mu}_t$ and an m-dimensional grid-level spatiotemporal process $\boldsymbol{v}_t$ transformed by a predetermined $n \times m$ mapping matrix $\boldsymbol{H}$. In particular, the second level of our model is

$$\boldsymbol{w}_t = \boldsymbol{\mu}_t + \boldsymbol{H}\boldsymbol{v}_t. \tag{2}$$

This model plays a key role in our methodology since it links the point-level process $\boldsymbol{w}_t$ with the grid-level process $\boldsymbol{v}_t$. The term $\boldsymbol{\mu}_t$ reflects the average level of the true process $\boldsymbol{w}_t$

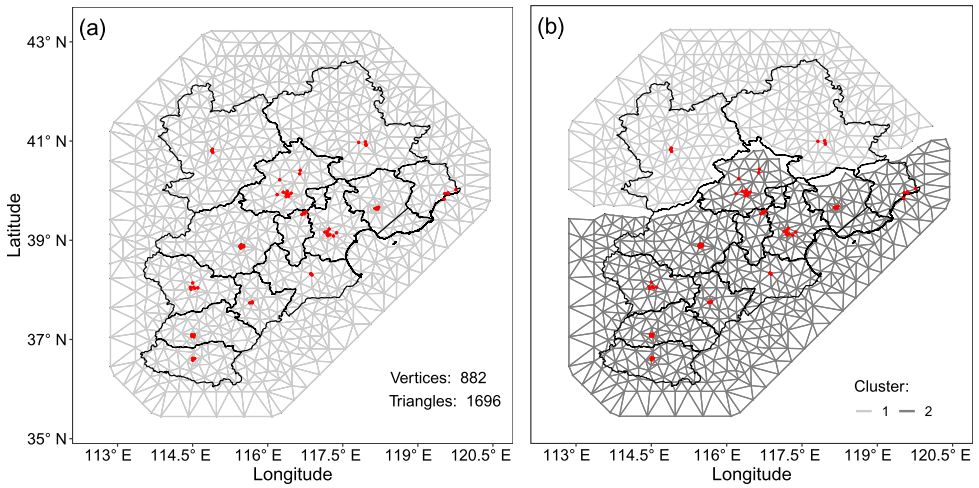


FIG. 4. *Triangulated mesh of the BTH region and extended area*: (*a*) *Triangulated mesh with a region*. (*b*) *Triangulated mesh with two subregions*.

and is used to incorporate any covariate variables in the model. The specific model of $\boldsymbol{\mu}_t$ depends on scientific knowledge from other studies and results from data explorations; one can use either linear regression models or more complex nonlinear models. The term $\boldsymbol{v}_t$ can be seen as random effects from CMAQ grids caused by uncertainties in physical and chemical models, source inventories, or initial conditions of CMAQ system, among others. In the cases where the observed data are from sparsely located monitoring stations and $\boldsymbol{v}_t$ is defined on the centroids of numerical model grid cells, the dimension of $\boldsymbol{v}_t$ is usually much higher than that of $\boldsymbol{w}_t$. Therefore, a mapping matrix $\boldsymbol{H}$ is critical in dimension reduction; more details on this are provided in Section 3.2.1.

Following Lindgren, Rue and Lindström (2011), we use a triangulation scheme to restrict the spatial support of $\boldsymbol{v}_t$ on the triangle vertices rather than on the centroids of the grid cells. The spatial support here is not related to the grid cells, which is different from "Bayesian melding" approaches where the processes usually depend on the grid cells (Fuentes and Raftery (2005), McMillan et al. (2010)). Compared to other methods based on the regular model grids, the proposed approach makes use of the triangulated mesh to better approximate the true latent processes around the spatial observed points and to achieve a more flexible trade-off between computation cost and performance.

For models with a very large number L of grid cells, the dimension reduction from the triangulation scheme is remarkable. As an illustration, Section 6 shows that when $L = 16{,}093$ and the number of triangles on the triangulated mesh is set to 3591 (corresponding to 2042 triangle vertices), the proposed approach still presents a competitive performance. Particularly, the BTH region is divided into a set of nonintersecting triangles where any two triangles meet in at most one common edge or vertex. In urban areas where the local spatial variations are usually large, we try to use triangles that are as small as possible with each triangle covering the fewest monitoring stations. In rural areas we use large triangles so that one triangle can contain multiple grids that are sufficiently close to each other. This procedure substantially reduces the computational time in largely rural areas with no monitoring station. To avoid boundary effects, we extend the BTH region to about 100 km beyond the region boundary and employ coarse triangles outside the region. Eventually, we have a total of 1646 triangles with 882 vertices that are denoted as $\{B_k : k = 1, \ldots, m\}$. The triangulated mesh is illustrated in Figure 4(a). In summary, given the computational cost for a very large L, L can be considered as the upper bound on m, and triangulated meshes with $m < L$ typically prove sufficient for a wide range of applications, including ours.

The applicability of the proposed method for a very large n can be guaranteed because the model (2) allows the spatial support of $\boldsymbol{v}_t$ to differ from that of observations. In the triangulation scheme, when n is very large, we can improve the computation at the expense of approximation accuracy by constructing a triangulated mesh such that $m < n$. This strategy is often adopted in low-rank methods such as predictive processes (Banerjee et al. (2008)), fixed-rank approximations (Cressie and Johannesson (2008)), etc. In HDCM, when $n > m$, the Sherman–Morrison–Woodbury formula can be applied to ensure that the computation depends mainly on m than n; see Section 3.2.4 for more details.

The choice of triangulated mesh usually depends on the distribution of spatial observed points; thus, m can also be quite large when n is very large. More specifically, as the spatial domain of interest and/or the number of the spatial observed points (i.e., n) become very large, a high-dimensional process $\boldsymbol{v}_t$ is required to capture small-scale spatial variations of data, leading to a very large m and expensive computations for parameter estimation. A well-designed procedure is needed.

To reduce the computational burden when m is very large, a spatial partitioning procedure is embedded in our triangulation scheme. We assume the spatial domain $\boldsymbol{D}$ can be partitioned into R subregions such that $\boldsymbol{D} = \bigcup_{r=1}^{R} \boldsymbol{D}_r$ and $\boldsymbol{D}_r \cap \boldsymbol{D}_{r'} = \varnothing$ for $r \neq r'$, where

$r' = 1, 2, \ldots, R$. This assumption is often used to accelerate data analysis in spatial statistics such as Heaton et al. (2019). For the choice of partitioning, various approaches have been proposed in the literature, including partitioning the region into equal areas (Sang, Jun and Huang (2011)), clustering-based partitioning (Kim, Mallick and Holmes (2005), Heaton, Christensen and Terres (2017)), and model-based methods (Konomi, Sang and Mallick (2014), Liang et al. (2021)), among others. For illustration purposes we partition the triangle vertices of Figure 4(a) into two clusters though the partitioning is not required for small n. As seen in Figure 1(b), the northern part of the BTH region is mountainous. Two cities in the north, namely, Chengde and Zhangjiakou, are not only at higher elevations than many other cities but also have lower $PM_{2.5}$ concentrations than the rest (see Table S1 in the Supplementary Material). Therefore, we select the triangle vertices located in the two cities as one cluster and the vertices in the remaining cities as another cluster, leading to $R = 2$ subregions; see Figure 4(b). For simplification, let $B_l^{(r)} \in \{B_1, B_2, \ldots, B_m\}$ and $B_l^{(r)}$ represent the vertice that belongs to the rth cluster (or subregion), where $l = 1, 2, \ldots, m_r$, and $m = \sum_{r=1}^{R} m_r$.

Earlier studies (Wan et al. (2021)) and Figure 2 suggest that $PM_{2.5}$ concentration of a particular location is usually affected by that of its neighbors at an earlier time. More specifically, the state of $\boldsymbol{v}_t$ at the kth vertex B_k is affected not only by its own previous state $\boldsymbol{v}_{t-1}$ but also by the neighbors' previous states. In this work, these complex spatiotemporal patterns of $PM_{2.5}$ concentrations are characterized by IDE models. In the spatiotemporal context, although IDE has been employed to model complex ecological dispersal patterns such as leptokurtic dispersal curves (Kot, Lewis and van den Driessche (1996)) and the meteorological dynamic systems (Wikle (2002), Xu, Wikle and Fox (2005), Wikle and Holan (2011), Zammit-Mangion and Wikle (2020)), the performance of IDE for bias correction remains unknown and deserves further investigation. Using IDE, spatial interactions of the processes can be described via a redistribution kernel $\mathcal{M}(\cdot, \cdot; \boldsymbol{\theta})$ with unknown parameters $\boldsymbol{\theta}$. More specifically, for the point-level process $w_t(\boldsymbol{s})$ of (1) at location $\boldsymbol{s}$ and at time point t, the stochastic IDE (Cressie and Wikle (2011)) can be formulated as

$$w_t(\boldsymbol{s}) = \int \mathcal{M}(\boldsymbol{s}, \boldsymbol{s}'; \boldsymbol{\theta}) w_{t-1}(\boldsymbol{s}')\, d\boldsymbol{s}' + \eta_t(\boldsymbol{s}), \tag{3}$$

where $\eta_t(\boldsymbol{s})$ represents random perturbations, called innovation. We further modify (3) by constructing IDE with discretized spatial knots defined on the triangle vertices and introduce a discretized stochastic IDE of the form

$$v_t^{(r)}(B_l^{(r)}) = \sum_{l'=1}^{m_r} \mathcal{M}(B_l^{(r)}, B_{l'}^{(r)}; \boldsymbol{\theta}^{(r)}) v_{t-1}^{(r)}(B_{l'}^{(r)}) + \eta_t(B_l^{(r)}), \tag{4}$$

where the innovation $\boldsymbol{\eta}_t^{(r)} = (\eta_t(B_1^{(r)}), \ldots, \eta_t(B_{m_r}^{(r)}))$ is a random field used to capture any remaining spatial correlation information. Let $\boldsymbol{v}_t^{(r)} = (v_t(B_1^{(r)}), \ldots, v_t(B_{m_r}^{(r)}))$, $\boldsymbol{v}_t = (\boldsymbol{v}_t^{(1)}, \ldots, \boldsymbol{v}_t^{(R)})^T$ and $\boldsymbol{\eta}_t = (\boldsymbol{\eta}_t^{(1)}, \ldots, \boldsymbol{\eta}_t^{(R)})^T$. We assume $\boldsymbol{v}_t^{(1)}, \ldots, \boldsymbol{v}_t^{(R)}$ are mutually and serially independent and the same for $\boldsymbol{\eta}_t^{(1)}, \ldots, \boldsymbol{\eta}_t^{(R)}$. We also assume $\boldsymbol{\eta}_t$ and $\boldsymbol{\varepsilon}_t$ in (1) are mutually and serially independent and the same for $\boldsymbol{\eta}_t$ and $\boldsymbol{v}_{t-1}$. Although the evolutions of IDEs of different subregions are independent of each other, the data located at different subregions can be dependent upon each other.

The proposed discretized IDE (4) is one of the major modeling differences between the proposed HDCM and many current calibration models. Many spatiotemporal calibration methods are based on first-order autoregression models, such as the dynamic downscaler model (Berrocal, Gelfand and Holland (2012)), and are unable to capture spatial interactions of high-dimensional dynamic processes. In contrast, IDE can model spatial interactions of spatiotemporal processes $\boldsymbol{v}_t$ through the use of a redistribution kernel and also account for

both spatial and temporal variations of the data. Some earlier IDE models employed basis expansion to model $w_t(\boldsymbol{s})$ of (3) as a product of fixed spatial basis functions and random temporal coefficients while assuming $\boldsymbol{v}_t$ of IDE (4) random in both space and time (Xu, Wikle and Fox (2005), Richardson, Kottas and Sansó (2017)).

3.2. *Model implementation*.

3.2.1. *The mean trend and mapping matrix*. Based on the model (2), we try to incorporate CMAQ outputs as a model covariate in a similar fashion as Berrocal, Gelfand and Holland (2012). More specifically, for any specific location s, we first calculate an inverse distance weighted average of CMAQ outputs within a circle of radius of 50 km centered at s and then use the weighted average as the covariate. In addition, earlier research revealed that $PM_{2.5}$ concentrations in the BTH region are significantly influenced by meteorological variables such as temperature, and eastern and northern cumulative wind powers (Liang et al. (2015)). Therefore, in addition to the weighted CMAQ output, we also include these three meteorological variables as covariates in the model. For simplicity, we use a linear form for the term $\boldsymbol{\mu}_t = (\mu_t(\boldsymbol{s}_1), \ldots, \mu_t(\boldsymbol{s}_n))^T$ in (2), that is, $\mu_t(\boldsymbol{s}) = \beta_0 + \boldsymbol{x}_t(C_s)\boldsymbol{\beta}$, where β_0 is the intercept and $\boldsymbol{x}_t(C_s)$ represent the weighted outputs of the numerical model at location $\boldsymbol{s}$ and time point t. It is worth noting that this model for $\mu_t(\boldsymbol{s})$ can be extended to nonlinear additive models (Chen et al. (2023)).

The mapping matrix $\boldsymbol{H}$ is an $n \times m$ dimensional matrix with the (i,k)th element denoted by $h(i,k)$, where $i = 1, 2, \ldots, n$ and $k = 1, 2, \ldots, m$. In the literature $h(i,k)$ has been modeled using the projection of a piecewise linear basis on a mesh (Lindgren, Rue and Lindström (2011)), but because the BTH data are sparsely observed in space, it would be difficult to identify $\boldsymbol{v}_t$ located in the rural areas with no stations using piecewise linear bases. Instead, we propose to map the grid-level process $\boldsymbol{v}_t$ to the point-level process $\boldsymbol{w}_t$ using a Wendland function $W(d;\theta_w)$ with $d \geq 0$ and the parameter $\theta_w > 0$. The function $W(d;\theta_w)$ is given by

$$W(d;\theta_w) = \begin{cases} \frac{1}{12}(1 - d/\theta_w)^3(1 + 3d/\theta_w) & \text{if } 0 \leq d \leq \theta_w, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

More specifically, $h(i,k) = W(d_{i,k}; c_h d_{\max}^H)$, where $d_{i,k} = \|\boldsymbol{s}_i - B_k\|$, $c_h \in (0,1]$ is a constant, and $d_{\max}^H = \max_{i,k}\{\|\boldsymbol{s}_i - B_k\|\}$; here $\|\boldsymbol{s} - B\|$ denotes the Euclidean distance between $\boldsymbol{s}$ and B. The Wendland functions have many attractive properties, including compact support, and are widely used in spatial statistics methods such as covariance tapering (Kaufman, Schervish and Nychka (2008)) and multiresolution Gaussian process models (Nychka et al. (2015)). More details on the Wendland functions can be found in Wendland (1995).

3.2.2. *Redistribution kernel and Gaussian Markov random fields (GMRFs) for IDEs*. For the redistribution kernel $\mathcal{M}$ in (4), the proposed approach allows $\mathcal{M}$ to be a spatially varying kernel that can be approximated by the basis expansion described in Chapter 5 of Wikle, Zammit-Mangion and Cressie (2019). Because the BTH monitoring stations are sparsely distributed in space and clustered in urban areas, to reduce the model complexity, we define the redistribution kernel with $\boldsymbol{\theta}^{(r)} = (\theta_1^{(r)}, \theta_2^{(r)})$ for the rth subregion as $\mathcal{M}(B_l^{(r)}, B_{l'}^{(r)}; \theta_1^{(r)}, \theta_2^{(r)}) = \theta_1^{(r)} W(d_{l,l'}^{(r)}; \theta_2^{(r)})$, where $\theta_1^{(r)} \neq 0$, $d_{l,l'}^{(r)} = \|B_l^{(r)} - B_{l'}^{(r)}\|$, and $l' = 1, 2, \ldots, m_r$.

Based on the relationships among the vertices of the triangulated mesh in Figure 4, we obtain spatial Markov structures. In light of this, we assume the innovation $\boldsymbol{\eta}_t^{(r)}$ from the rth subregion as a GMRF (Rue and Held (2005)) with a sparse precision matrix $\boldsymbol{Q}^{(r)}$, that is,

$\boldsymbol{\eta}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{Q}^{-1,(r)})$, where $\boldsymbol{Q}^{(r)} = \tau^{2,(r)}(\boldsymbol{G}^{(r)} + \zeta^{2,(r)}\boldsymbol{I}_m)$, $\tau^{(r)} > 0$, and $\zeta^{(r)} > 0$ are hyperparameters, and $\boldsymbol{G}^{(r)}$ is a Laplacian matrix defined on the triangulated mesh with elements

$$G_{ll'}^{(r)} = \begin{cases} \text{the degree of } B_l & \text{if } l = l', \\ -1 & \text{if } B_l \text{ is adjacent to } B_{l'} \text{ in the lattice,} \\ 0 & \text{otherwise.} \end{cases}$$

We also assume that IDE (4) initiates from another GMRF denoted as $\boldsymbol{v}_0^{(r)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{Q}_0^{-1,(r)})$, where $\boldsymbol{Q}_0^{(r)} = \tau_0^{2,(r)}(\boldsymbol{G} + \zeta_0^{2,(r)}\boldsymbol{I}_m)$, $\tau_0^{(r)} > 0$, and $\zeta_0{}^{(r)} > 0$; also, refer to Bolin, Wallin and Lindgren (2019) for a similar definition. The reasons behind the definition of the precision matrices are discussed in Section S3.1 of the Supplementary Material. The advantage of this definition is also discussed in Chen et al. (2023).

Let $\boldsymbol{y}_{1:t} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_t\}$ and $\boldsymbol{v}_{0:t} = \{\boldsymbol{v}_0, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_t\}$. All observed data and random effects can be represented by $\boldsymbol{y} = \boldsymbol{y}_{1:N_t}$ and $\boldsymbol{v} = \boldsymbol{v}_{0:N_t}$, respectively. There are many parameters involved in the proposed HDCM. All parameters other than $\boldsymbol{v}$ are summarized by $\boldsymbol{\Theta} = \{\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\tau}_0^2, \boldsymbol{\tau}^2, \boldsymbol{\zeta}_0^2, \boldsymbol{\zeta}^2\}$, where $\boldsymbol{\theta}_1 = \{\theta_1^{(1)}, \ldots, \theta_1^{(R)}\}$, $\boldsymbol{\theta}_2 = \{\theta_2^{(1)}, \ldots, \theta_2^{(R)}\}$, $\boldsymbol{\tau}_0^2 = \{\tau_0^{2,(1)}, \ldots, \tau_0^{2,(R)}\}$, $\boldsymbol{\tau}^2 = \{\tau^{2,(1)}, \ldots, \tau^{2,(R)}\}$, $\boldsymbol{\zeta}_0^2 = \{\zeta_0^{2,(1)}, \ldots, \zeta_0^{2,(R)}\}$, and $\boldsymbol{\zeta}^2 = \{\zeta^{2,(1)}, \ldots, \zeta^{2,(R)}\}$. Then the posterior distribution of parameters, given the point-level observations $\boldsymbol{y}$, can be expressed as

$$p(\boldsymbol{\Theta}, \boldsymbol{v}|\boldsymbol{y}) = \frac{L(\boldsymbol{y}; \boldsymbol{v}, \boldsymbol{\Theta}) p(\boldsymbol{v}_{1:N_t}|\boldsymbol{v}_0, \boldsymbol{\Theta}) p(\boldsymbol{v}_0|\boldsymbol{\Theta}) \pi(\boldsymbol{\Theta})}{p(\boldsymbol{y})}, \tag{6}$$

where $p(\boldsymbol{y})$ is the joint distribution of data, $L(\boldsymbol{y}; \boldsymbol{v}, \boldsymbol{\Theta})$ is the data likelihood function from (1) and (2), $p(\boldsymbol{v}_{1:N_t}|\boldsymbol{v}_0, \boldsymbol{\Theta}) = \prod_{t=1}^{N_t} p(\boldsymbol{v}_t|\boldsymbol{v}_{t-1}, \boldsymbol{\Theta})$ is the distribution from the model (4), $p(\boldsymbol{v}_0|\boldsymbol{\Theta})$ is the distribution for the initial state $\boldsymbol{v}_0$, $\pi(\boldsymbol{\Theta})$ is the prior distribution for $\boldsymbol{\Theta}$. Specific priors for each parameter can be found in Section S3.2 of the Supplementary Material. For model implementation we adopt a VB method along with a space-partitioning-based EnKS approximation technique.

3.2.3. *Variational Bayes for HDCM.* The posterior distribution in (6) cannot be expressed analytically, and hence we carry out an approximate inference using the VB method. The VB method is a fast alternative to MCMC and can be easily scaled up to massive data. Especially, a Laplace approximation is embedded to accelerate the approximation of parameter posteriors without conjugate priors. Technical details on the VB and Laplace approximation are referred to Section S9 and Section S10 of the Supplementary Material, respectively. The excellent performance of this approach has been verified by our analysis and Ren et al. (2011); see Blei, Kucukelbir and McAuliffe (2017) and references therein for a further detailed discussion of VB.

3.2.4. *Space-partitioning-based EnKS for HDCM.* We approximate the posterior distribution of $\boldsymbol{v}$ using an EnKS. Conventional methods, such as KF, can generate samples from the posterior distribution, but storing and inverting large matrices become expensive and infeasible when m and/or n is large. Therefore, approximations are needed. EnKF, developed by Evensen (1994) for geophysical data assimilation, is an approximate version of KF in which the distribution of the state vectors (i.e., $\boldsymbol{v}$ in our model) is approximated by a sample or an "ensemble" from distributions. The ensemble is then propagated forward through time and updated when new data become available. This ensemble-based representation is a form of dimension reduction that leads to computational feasibility, even for very high-dimensional

systems. EnKS proposed by Evensen and Van Leeuwen (2000) is an extension of EnKF to a smoother formulation. Compared with EnKF that uses only past and present observations, EnKS also includes future observations, that is, $\boldsymbol{y}_t$ for $t = 1, \ldots, N_t$, to calibrate CMAQ outputs at each time $t' \in \{1, \ldots, N_t\}$, so it provides more accurate spatial interpolations than EnKF. It works particularly well for data with multimodality and skewness properties that are shared by many real-life data applications including ours (Katzfuss, Stroud and Wikle (2016), Katzfuss, Stroud and Wikle (2020)).

We develop a space-partitioning-based EnKS (spEnKS) to accelerate parameter estimation when both n and m are very large. EnKF and EnKS are computationally expensive for large n and m, as they need to invert an $n \times n$ or $m \times m$ matrix multiple times at each iteration. The proposed spEnKS embeds a space-partitioning-based procedure in EnKS and is described in Algorithm 1. To avoid rank deficiency and spurious correlations for small ensemble members, the sample cross-covariance matrices are regularized using a spatiotemporal tapering method; see (Katzfuss, Stroud and Wikle (2020)) and references therein. Let $\Gamma^{(r)}_{\rho^{(r)}}(l, l')$ be the covariance tapering function in space for the rth subregion, where $\rho^{(r)}$ is a tuning parameter and $r = 1, 2, \ldots, R$. The optimal $\rho^{(r)}$ is usually selected from the interval $(0, d^{B^{(r)}}_{\max}]$ through data-driven methods, such as cross-validation, where $d^{B^{(r)}}_{\max} = \max_{l,l'}\{\|\boldsymbol{B}^{(r)}_l - \boldsymbol{B}^{(r)}_{l'}\|\}$, but the selection of R tuning parameters can come with a high computational cost. In this work we define all $\rho^{(r)}$ parameters using only one parameter c_s, that is, $\rho^{(r)} = c_s d^{B^{(r)}}_{\max}$, where $c_s \in (0, 1]$ and is quickly optimized using an automated grid search. Therefore, the covariance tapering function in space becomes $\Gamma^{(r)}_{c_s}(l, l') = W(d_{l,l'}; c_s d^{B^{(r)}}_{\max})$. And the covariance tapering function in time is set as $\gamma_{c_t}(t, t') = W(d_t; c_t)$, where $d_t = |t - t'|$ and $c_t \in \{1, 2, \ldots, N_t\}$ is another tuning parameter.

The efficiency of ensemble-based methods is affected mostly by the inverse of an $n \times n$ matrix in (8), denoted as $\boldsymbol{\Phi} = \boldsymbol{H}\widehat{\boldsymbol{\Sigma}}_{t,t|t-1}\boldsymbol{H}^T + \langle\sigma^2\rangle\boldsymbol{I}_n$, where $\langle\cdot\rangle$ denote the expectation. When n is very large and $m > n$, some dimension reduction techniques are often used to accelerate the inverse of $\boldsymbol{\Phi}$, such as the eigenpair-based method (Heinrich et al. (2021)). For the usual case where $n > m$, the Sherman–Morrison–Woodbury formula can be applied to obtain $\boldsymbol{\Phi}^{-1} = \langle\sigma^{-2}\rangle\boldsymbol{I}_n - \langle\sigma^{-2}\rangle\boldsymbol{H}\boldsymbol{\Omega}^{-1}\boldsymbol{H}^T$, where $\boldsymbol{\Omega} = \langle\sigma^2\rangle\widehat{\boldsymbol{\Sigma}}^{-1}_{t,t|t-1} + \boldsymbol{H}^T\boldsymbol{H}$, and is an $m \times m$ matrix. Thus, the inverse of an $n \times n$ matrix $\boldsymbol{\Phi}$ is reduced to the inverse of an $m \times m$ matrix $\boldsymbol{\Omega}$. Note that $\boldsymbol{\Omega}$ is not dense because $\widehat{\boldsymbol{\Sigma}}_{t,t|t-1}$ is a block-diagonal matrix consisting of R tapered sample covariance matrices, and $\boldsymbol{H}$ is also sparse. As a result, (7) can be modified to

$$\boldsymbol{v}^{(e)}_{t'|t} = \boldsymbol{v}^{(e)}_{t'|t-1} + \langle\sigma^{-2}\rangle\widehat{\boldsymbol{\Sigma}}_{t,t'|t-1}\boldsymbol{H}^T\{\boldsymbol{I}_n(\boldsymbol{z}_t - \widetilde{\boldsymbol{z}}^{(e)}_t) - \boldsymbol{H}\widehat{\boldsymbol{\xi}}\}, \tag{10}$$

where $\widehat{\boldsymbol{\xi}}$ is a solution of the linear system $\boldsymbol{\Omega}\boldsymbol{\xi} = \boldsymbol{H}^T(\boldsymbol{z}_t - \widetilde{\boldsymbol{z}}^{(e)}_t)$.

The proposed spEnKS has several other computational advantages. First, by using several sparse precision matrices, such as $\boldsymbol{Q}^{(r)}$ and $\boldsymbol{Q_0}^{(r)}$, spEnKS carries out a fast Cholesky factorization leading to efficient sampling (Rue and Held (2005)). Second, for very large m, the spatial partitioning accelerates the computation as the calculation of high-dimensional matrices becomes the calculation of multiple low-dimensional matrices, namely, from an m-dimensional $\boldsymbol{\Sigma}_{t',t|t-1}$ to R m_r-dimensional $\boldsymbol{\Sigma}^{(r)}_{t',t|t-1}$. Lastly, different from the standard forward-backward methods, such as KS, spEnKS is a forward-only process.

Finally, we fit the proposed HDCM using an iterative procedure, VB-spEnKS; refer to Algorithm 2. The performance of VB-spEnKS is verified in the simulation study, where an MCMC implementation with direct KS of HDCM is used to assess the performance of VB-spEnKS. More details are provided in Section S5 of the Supplementary Material.

Algorithm 1: Space-partitioning-based Ensemble Kalman smoother (spEnKS)

We use $\{\boldsymbol{v}_{t'|t}^{(e)}\}_{e=1}^{N_e}$ to denote N_e ensemble members, which are approximate samples from $p(\boldsymbol{v}_{t'}|\boldsymbol{y}_{1:t}, \boldsymbol{\Theta})$ where $t', t \in \{0, 1, \ldots, N_t\}$. Let $\langle\cdot\rangle$ denote the expectation.

Input: Given $\langle\boldsymbol{\Theta}\rangle$, start with an initial ensemble $\{\boldsymbol{v}_{0|0}^{(e)}\}_{e=1}^{N_e}$ where

$\boldsymbol{v}_{0|0}^{(e)} = (\widetilde{\boldsymbol{v}}_{0|0}^{(1,e)}, \ldots, \widetilde{\boldsymbol{v}}_{0|0}^{(r,e)})^T$ with $\widetilde{\boldsymbol{v}}_{0|0}^{(r,e)} \sim \mathcal{N}(\boldsymbol{0}, \langle \boldsymbol{Q}_0^{-1,(r)}\rangle)$, here

$\langle \boldsymbol{Q}_0^{(r)}\rangle = \langle\widehat{\tau}_0^{2,(r)}\rangle(\boldsymbol{G}^{(r)} + \langle\widehat{\zeta}_0^{2,(r)}\rangle \boldsymbol{I}_{m_r})$, then

for $t = 1, \ldots, N_t$ **do**

  for $e = 1, \ldots, N_e$ **do**

    Forecast step: **for** $r = 1, \ldots, R$ **do**

      Compute $\widetilde{\boldsymbol{v}}_{t|t-1}^{(r,e)} = \langle\theta_1^{(r)}\rangle \boldsymbol{\mathcal{M}}_{\langle\theta_2^{(r)}\rangle} \widetilde{\boldsymbol{v}}_{t-1|t-1}^{(r,e)} + \boldsymbol{\eta}_t^{(r,e)}$ where

      $\boldsymbol{\mathcal{M}}_{\langle\theta_2^{(r)}\rangle}$ is an $m_r \times m_r$ matrix with the (l, l')th element as $W(d_{l,l'}; \theta_2^{(r)})$,

      $\boldsymbol{\eta}_t^{(r,e)} \sim \mathcal{N}(\boldsymbol{0}, \langle \boldsymbol{Q}^{-1,(r)}\rangle)$ with $\langle \boldsymbol{Q}^{(r)}\rangle = \langle\tau^{2,(r)}\rangle(\boldsymbol{G}^{(r)} + \langle\zeta^{2,(r)}\rangle \boldsymbol{I}_{m_r})$.

    end

    Update step: **for** $t' = \max\{0, t - c_t\}, \ldots, t$ **do**

      (7) $$\boldsymbol{v}_{t'|t}^{(e)} = \boldsymbol{v}_{t'|t-1}^{(e)} + \widehat{\boldsymbol{K}}_{t',t}(\boldsymbol{z}_t - \widetilde{\boldsymbol{z}}_t^{(e)}),$$

      where $\widehat{\boldsymbol{K}}_{t',t}$ denotes an estimate of matrix $\boldsymbol{K}$ called Kalman gain, $\boldsymbol{z}_t = \boldsymbol{y}_t - \boldsymbol{X}_t\langle\boldsymbol{\beta}\rangle$ and $\widetilde{\boldsymbol{z}}_t^{(e)} = \boldsymbol{H}\boldsymbol{v}_{t|t-1}^{(e)} + \boldsymbol{\varepsilon}_t^{(e)}$ is the pseudo-observation with $\boldsymbol{\varepsilon}_t^{(e)} \sim \mathcal{N}(\boldsymbol{0}, \langle\sigma^2\rangle \boldsymbol{I}_n)$. Particularly, $\widehat{\boldsymbol{K}}_{t',t}$ is given by

      (8) $$\widehat{\boldsymbol{K}}_{t',t} = \widehat{\boldsymbol{\Sigma}}_{t',t|t-1}\boldsymbol{H}^T\big(\boldsymbol{H}\widehat{\boldsymbol{\Sigma}}_{t,t|t-1}\boldsymbol{H}^T + \langle\sigma^2\rangle \boldsymbol{I}_n\big)^{-1},$$

      where $\widehat{\boldsymbol{\Sigma}}_{t',t|t-1} = \mathrm{diag}(\widehat{\boldsymbol{\Sigma}}_{t',t|t-1}^{(1)}, \ldots, \widehat{\boldsymbol{\Sigma}}_{t',t|t-1}^{(R)})$ is a block-diagonal matrix, here $\widehat{\boldsymbol{\Sigma}}_{t',t|t-1}^{(r)}$ is a regularized version of the sample cross-covariance matrix $\widetilde{\boldsymbol{\Sigma}}_{t',t|t-1}^{(r)}$ of $\boldsymbol{v}_{t'|t-1}^{(r,1:N_e)}$ and $\boldsymbol{v}_{t|t-1}^{(r,1:N_e)}$, that is

      (9) $$\widehat{\boldsymbol{\Sigma}}_{t',t|t-1}^{(r)} = \gamma_{c_t}(t', t)\widetilde{\boldsymbol{\Sigma}}_{t',t|t-1}^{(r)} \circ \boldsymbol{\Gamma}_{c_s}^{(r)},$$

      where $\gamma_{c_t}(\cdot)$ with a tuning parameter c_t is a temporal tapering function, "$\circ$" denotes the element-wise multiplication, and $\boldsymbol{\Gamma}_{c_s}^{(r)}$ is an m_r by m_r spatial tapering matrix with the (l, l')th element as $\Gamma_{c_s}^{(r)}(l, l')$ that depends on a tuning parameter c_s.

    end

  end

end

3.3. *Tuning parameters selection*. The proposed approach involves the selection of four tuning parameters: (1) the number of ensembles N_e, (2) $c_h \in (0, 1]$ related to the mapping matrix $\boldsymbol{H}$, (3) $c_s \in (0, 1]$ for the covariance tapering in space, and (4) $c_t \in \{1, 2, \ldots, N_t\}$ for the covariance tapering in time. In this work we adopt the commonly used ensemble size in the literature (Mitchell, Houtekamer and Pellerin (2002), Houtekamer and Zhang (2016)) and let $N_e = 100$. The other three parameters are obtained through a cross-validation procedure presented in Section 4. Under the setting of HDCM, c_h and c_t generally depend on the spatial range and temporal range, respectively, and we use empirical covariograms to facilitate the grid search of the two parameters in the cross-validation. For c_s, Kirchgessner, Nerger and Bunse-Gerstner (2014) showed $c_s = 8\sqrt{\frac{N_e}{40}}\,dx$ in dense observations from regular grids,

Algorithm 2: A variational Bayes implementation with a space-partitioning-based EnKS (VB-spEnKS) for HDCM

Input: Initialize $\boldsymbol{\Theta}^{(0)}$.

Iterate the following steps until convergence: **for** $\ell = 1, 2, \ldots$ **do**

a. Obtain ensemble members $\boldsymbol{v}_{1:N_t|N_t}^{(1:N_e;\ell)}$ from $p(\boldsymbol{v}_{1:N_t}|\boldsymbol{y}, \boldsymbol{\Theta}^{(\ell-1)})$ using spEnKS of Algorithm 1.

b. Estimate the expectation and variance of $\boldsymbol{\Theta}^{(\ell)}$ by $p(\boldsymbol{\Theta}|\boldsymbol{y}, \boldsymbol{v}_{1:N_t|N_t}^{(1:N_e;\ell)})$ using the VB procedure provided in Section S9.2 of the Supplementary Material.

end

where dx represents the grid spacing. Although this formula cannot be employed directly as our observations are not dense and the triangulated mesh is irregular, it reduces the domain of grid search for c_s; see Section S4 of the Supplementary Material for further details on the selections of the parameters, along with a sensitivity of HDCM with respect to the selection of the tuning parameters.

Based on the cross-validation, for the BTH data, we set $c_h = 0.23$ in the winter (corresponding to about 140 km), $c_h = 0.22$ in the summer (corresponding to about 130 km), $c_s = 0.3$ (corresponding to about 280 km), and $c_t = 1$ in both seasons. For the reanalysis data in Section 6.1, $c_h = 0.05$ (corresponding to about 80 km), $c_t = 1$, and different c_s values are used in various meshes (see Table 3).

3.4. *Calibration specification.* For a given CMAQ grid cell, say C, we denote its calibration as $\tilde{y}_t(C)$ that can be obtained by integrating the point-level posterior samples

$$\tilde{y}_t(C) = \frac{1}{N_e \int_C ds} \sum_{e=1}^{N_e} \int_C y_t^{(e)}(\boldsymbol{s})\, d\boldsymbol{s}, \tag{11}$$

where $y_t^{(e)}(s)$ is from its posterior distribution.

To approximate the integration in (11), we use a bivariate uniform distribution to randomly sample $n_c = 50$ spatial points from each CMAQ grid cell and denote them for the grid C_0 as $\{s_0^{i'}\}_{i'=1}^{n_c}$. At the time t, we have their posterior samples:

$$y_t^{(e)}(s_0^{i'}) = \boldsymbol{X}_t(s_0^{i'})\boldsymbol{\beta}_t^{(e)} + \boldsymbol{h}_{s_0^{i'}}^T \boldsymbol{v}_{t|T}^{(e)} + \varepsilon_t^{(e)}(s_0^{i'}), \tag{12}$$

where $\boldsymbol{\beta}_t^{(e)}$ and $\varepsilon_t^{(e)}(s_0^{i'})$ are random samples from the posterior distribution $\mathcal{N}(\langle\boldsymbol{\beta}\rangle, \boldsymbol{D}_\beta)$ defined in equation (S15) of the Supplementary Material and $\mathcal{N}(0, \langle\sigma^2\rangle)$, for $e = 1, \ldots, N_e$, respectively.

The expectation and variance of CMAQ model outputs calibrated using (12) at the grid cell C_0 can be written as follows:

$$\mathrm{E}\{y_t(C_0)\} = \frac{1}{N_e n_c} \sum_{e=1}^{N_e} \sum_{i'=1}^{n_c} y_t^{(e)}(s_0^{i'}),$$

$$\mathrm{Var}\{y_t(C_0)\} = \frac{1}{N_e n_c - 1} \sum_{e=1}^{N_e} \sum_{i'=1}^{n_c} [y_t^{(e)}(s_0^{i'}) - \mathrm{E}\{y_t(C_0)\}]^2.$$

4. Regional cross-validation. To assess the calibration performance of HDCM, we carry out a regional cross-validation method to compare the out-of-sample model predictions against the observational data. Each time, we consider all monitoring data from one city as the test set and data from the remaining 12 cities as the training set. More specifically, for a given city, after leaving out the time series data collected from the stations in that city, we fit our models using the data from all other cities and predict the data at the stations in that city. The model performance is evaluated by comparing the predictions $\{\widehat{\boldsymbol{y}}_t\}_{t=1}^{N_t}$ against the observations $\{\boldsymbol{y}_t\}_{t=1}^{N_t}$ in the city, according to some criteria. We call this procedure "leave-one-city-out cross-validation."

For our data analysis, we use four widely accepted criteria: root mean squared error (RMSE), continuous rank probability score (CRPS) (Gneiting and Raftery (2007)), mean absolute error (MAE), and the fraction of predictions within a factor of two (FAC2) (Chang and Hanna (2004)). In general, a small RMSE, MAE, or CRPS indicates accurate, precise, or sharp predictions, and a FAC2 close to 1 represents robust predictions.

We compare the proposed HDCM with five popular models including universal kriging (UK), a random-forest-based method (RF) advocated by atmospheric scientists (Wang et al. (2019), Zhao et al. (2021), Berrocal et al. (2020)), a Bayesian downscaling model that allows for spatially varying coefficients (SVC) (Gelfand et al. (2003)), a Bayesian spatiotemporal downscaler model with spatiotemporally varying coefficients (STVC) (Section 3.2 of Berrocal, Gelfand and Holland (2010)), and a Bayesian spatiotemporal model with a first-order spatiotemporal autoregression (STAR) (Section 7.2 of Blangiardo and Cameletti (2015)). Note that UK is basically a linear interpolation method (Cressie and Wikle (2011)). RF is one of the most popular machine learning techniques. SVC models have been widely used in modeling air pollution in recent years, and their applications in numerical model calibrations have been considerably successful (Jiang and Yoo (2019), Berrocal et al. (2020)). The details of the five competitive models are presented in Section S2 of the Supplementary Material. In the data analysis, the first three are implemented using the source codes provided by Berrocal et al. (2020) (https://github.com/yawenguan/DataFusion); STVC and STAR are implemented through the `R` package `spTDyn` (Bakar, Kokic and Jin (2016)) and `INLA` (Lindgren and Rue (2015)), respectively.

4.1. *Numerical results*. For the BTH data in the winter of 2015, cross-validation results are reported in Tables 1–2. The cross-validation results of the summer of 2015 are shown in Tables S4–S5 of the Supplementary Material. Other than the proposed HDCM, we also consider a space-partitioning-based HDCM with two subregions (HDCM_2), and HDCM_2 performance is reported in Table S6 of the Supplementary Material; see Figure 4(b) for the distribution of two subregions. A modified Diebold–Mariano test (DM test, see Harvey, Leybourne and Newbold (1997)) is used to determine whether the difference between any two models is significant in terms of predictive accuracy. DM tests with a significance level of 5% for both the winter and summer seasons are provided in Table S7 and Table S8 of the Supplementary Material, respectively. Here are the results:

a. The proposed HDCM in all cases outperforms all six competing models using the average value of the different metrics across all cities, including average RMSE, average CRPS, average MAE, and average FAC2, as seen in the last rows of Tables 1–2 and Tables S4–S5.

b. As reported in Tables 1–2 for the winter, considering all cities individually, HDCM is either the best or the second best model for *nine* of the 13 cities using RMSE, CRPS, or MAE, and for 10 cities using FAC2. Based on Table S7 for DM test, HDCM significantly outperforms other methods for at least six cities, while other methods significantly outperform HDCM for no more than two cities. For example, the predictions of HDCM are significantly

TABLE 1

Averaged RMSE (Root Mean Squared Error) and CRPS (Continuous Rank Probability Score) for $PM_{2.5}$ *concentration predictions* ($\mu g/m^3$) *calculated for* 13 *cities in the BTH region using leave-one-city-out cross-validation from seven models*: (1) *CMAQ numerical model output*, (2) *universal kriging* (*UK*), (3) *random forest model* (*RF*), (4) *spatially-varying coefficient downscaling model* (*SVC*), (5) *spatiotemporally-varying coefficient downscaling model* (*STVC*), (6) *first-order spatiotemporal autoregression* (*STAR*), *and* (7) *the proposed hierarchical dynamic calibration model* (*HDCM*). *The smallest RMSE and CRPS are in bold, and the second smallest ones are underlined. Differences between methods at a* 5% *significance level are detailed in Table S*7 *of the Supplementary Material. Daily data from November* 1, 2015 *to January* 31, 2016 *are considered*

	RMSE						
City	CMAQ	UK	RF	SVC	STVC	STAR	HDCM
Baoding	94.64	77.40	77.37	<u>75.51</u>	78.38	76.58	**73.12**
Beijing	92.52	67.35	70.74	<u>61.92</u>	66.69	62.92	**57.65**
Cangzhou	60.89	37.04	41.41	<u>35.44</u>	36.69	**33.76**	39.54
Chengde	41.62	30.10	32.14	34.83	27.91	<u>25.08</u>	**24.14**
Handan	66.09	60.37	52.31	58.16	<u>51.86</u>	57.74	**44.23**
Hengshui	91.19	70.89	72.44	58.15	68.48	<u>57.46</u>	**51.82**
Langfang	85.44	43.85	55.74	**34.88**	45.88	38.95	<u>37.14</u>
Qinhuangdao	46.19	37.91	<u>34.85</u>	35.12	39.67	42.38	**27.74**
Shijiazhuang	87.72	<u>47.83</u>	54.50	49.19	50.62	**46.40**	54.09
Tangshan	95.75	44.07	47.58	<u>32.61</u>	40.68	**28.71**	34.16
Tianjin	70.57	38.37	42.70	**32.55**	38.91	<u>33.14</u>	34.64
Xingtai	85.69	58.21	59.42	<u>53.17</u>	58.53	53.42	**49.20**
Zhangjiakou	64.42	58.40	<u>37.64</u>	78.96	52.72	69.78	**29.94**
Average	75.60	51.68	52.22	49.27	50.54	<u>48.18</u>	**42.88**

	CRPS						
City	CMAQ	UK	RF	SVC	STVC	STAR	HDCM
Baoding	50.35	42.31	41.43	41.24	41.95	<u>41.18</u>	**39.51**
Beijing	45.82	33.18	35.00	**30.43**	32.61	31.05	<u>30.54</u>
Cangzhou	31.78	24.11	26.05	25.54	24.13	<u>23.90</u>	**22.05**
Chengde	22.66	16.87	17.09	19.47	16.18	<u>15.48</u>	**13.73**
Handan	37.58	36.91	32.05	37.21	<u>31.83</u>	36.61	**29.15**
Hengshui	46.21	36.64	35.30	32.32	35.14	**31.33**	<u>31.87</u>
Langfang	46.17	28.45	32.18	<u>26.97</u>	27.73	**26.40**	28.80
Qinhuangdao	24.90	21.91	<u>20.31</u>	20.70	23.07	24.46	**14.69**
Shijiazhuang	48.18	31.99	32.30	33.01	<u>30.95</u>	**30.18**	33.98
Tangshan	52.20	28.24	28.03	21.44	26.29	**19.62**	<u>20.94</u>
Tianjin	38.00	26.08	27.16	<u>24.55</u>	25.95	**24.08**	25.76
Xingtai	45.46	34.24	33.16	<u>31.38</u>	33.32	**30.84**	32.20
Zhangjiakou	33.94	31.27	<u>19.69</u>	43.29	28.39	38.86	**16.18**
Average	40.25	30.17	29.21	29.81	29.04	<u>28.77</u>	**26.11**

more accurate than SVC in *nine* of the 13 cities, while SVC outperforms HDCM in only *two* cities. Furthermore, HDCM is not merely marginally superior but considerably better than its six competitors in Handan, Qinhuangdao, and Zhangjiakou; see especially Zhangjiakou in Table 1, where five of the six competitors almost double the RMSE.

c. As reported in Tables S4–S5 for the summer, the proposed HDCM is either the best or the second best model for *five* of 13 cities using RMSE, *seven* cities using CRPS, *eight* cities using MAE, and *six* cities using FAC2. Based on Table S8 for DM test, HDCM significantly outperforms other methods for at least two cities while other methods significantly outperform HDCM for no more than two cities. Specifically, although SVC has better predictions

TABLE 2

Averaged MAE (Mean Absolute Error) and FAC2 (Fraction of Predictions within a factor of 2) for $PM_{2.5}$ *concentration predictions* ($\mu g/m^3$) *calculated for* 13 *cities in the BTH region using leave-one-city-out cross-validation from seven models*: (1) *CMAQ numerical model output*, (2) *universal kriging* (*UK*), (3) *random forest model* (*RF*), (4) *spatially-varying coefficient downscaling model* (*SVC*), (5) *spatiotemporally-varying coefficient downscaling model* (*STVC*), (6) *first-order spatiotemporal autoregression* (*STAR*), *and* (7) *the proposed hierarchical dynamic calibration model* (*HDCM*). *The smallest MAE and FAC2 are in bold, and the second smallest ones are underlined. Daily data from November* 1, 2015 *to January* 31, 2016 *are considered*

	MAE						
City	CMAQ	UK	RF	SVC	STVC	STAR	HDCM
Baoding	65.42	54.20	53.64	52.95	53.68	52.64	**51.34**
Beijing	56.90	38.34	41.93	34.35	38.33	37.36	**31.56**
Cangzhou	41.86	25.24	30.46	24.14	25.23	**23.66**	25.44
Chengde	31.05	21.39	22.63	21.90	20.66	17.73	**17.71**
Handan	50.90	43.30	38.96	37.37	36.96	39.88	**30.90**
Hengshui	58.78	42.91	42.64	34.03	41.16	**33.30**	34.03
Langfang	62.46	30.45	37.62	**23.07**	30.41	25.53	23.55
Qinhuangdao	31.23	27.46	26.60	26.74	27.24	34.05	**19.01**
Shijiazhuang	62.63	32.87	39.69	34.55	33.95	**32.76**	37.16
Tangshan	69.45	33.24	33.05	24.00	30.93	**20.12**	24.17
Tianjin	50.68	28.57	29.32	**21.83**	29.25	24.87	22.12
Xingtai	61.18	37.15	38.98	33.24	38.06	**32.80**	33.17
Zhangjiakou	43.90	40.55	26.40	57.08	36.98	51.75	**21.51**
Average	52.80	35.05	35.53	32.71	34.06	32.80	**28.59**

	FAC2						
City	CMAQ	UK	RF	SVC	STVC	STAR	HDCM
Baoding	0.70	0.86	0.85	0.86	**0.90**	0.88	**0.90**
Beijing	0.62	0.85	0.77	0.89	0.84	0.83	**0.90**
Cangzhou	0.79	**0.95**	0.89	**0.95**	**0.95**	**0.95**	0.93
Chengde	0.43	0.65	0.76	0.82	0.68	0.84	**0.87**
Handan	0.75	0.92	0.87	**0.94**	0.92	0.91	0.93
Hengshui	0.77	0.88	0.92	0.97	0.91	0.98	**0.99**
Langfang	0.67	0.93	0.91	**0.99**	0.94	0.98	0.98
Qinhuangdao	0.60	0.68	0.68	0.68	0.79	0.62	**0.80**
Shijiazhuang	0.71	0.95	0.90	0.95	0.95	**0.97**	0.95
Tangshan	0.60	0.89	0.89	0.92	0.91	**0.98**	0.96
Tianjin	0.64	0.89	0.88	0.96	0.90	0.94	**0.97**
Xingtai	0.74	0.95	0.94	**0.98**	0.93	**0.98**	0.96
Zhangjiakou	0.45	0.54	0.64	0.40	0.54	0.42	**0.69**
Average	0.65	0.84	0.84	0.87	0.86	0.87	**0.91**

than HDCM in *seven* cities in terms of RMSE as seen in Table S4, DM test shows SVC outperforms HDCM in only two of the seven cities while HDCM outperforms SVC in *three* other cities.

d. Compared to the best-performing HDCM, the predictive performance of $HDCM_2$ is about 1% worse using the average value of different metrics across all cities. This is not surprising because $HDCM_2$ for IDE (4) does not consider spatial interactions of the processes between two subregions. Nevertheless, $HDCM_2$ still outperforms other competing models in terms of the average RMSE, average CRPS, average MAE, and average FAC2 in both seasons of 2015.

In summary, CMAQ outputs have the worst performance among all models, suggesting an urgent need for calibration of the numerical model outputs. This is especially true for winters, as winter predictions are generally worse than summer predictions, mostly due to the large fluctuations in the winter air quality of the BTH region (see also Table S1 of the Supplementary Material). Compared with the raw CMAQ outputs, HDCM significantly improves average RMSE by 43.28%, average CRPS by 35.13%, average MAE by 45.85%, and average FAC2 by 40.17% in the winter, and 38.84%, 41.79%, 40.37%, and 31.64% in the summer, respectively. While SVC model generally performs better than UK and RF models, as seen in the comparison study done by Berrocal et al. (2020), HDCM further improves SVC model performance in average RMSE by 12.97%, average CRPS by 12.42%, average MAE by 12.59%, and average FAC2 by 5.02% in the winter, and 3.54%, 4.72%, 3.99%, and 1.01% in the summer, respectively. In addition, HDCM also improves the performance of two spatiotemporal models. For example, in comparison with STAR, HDCM improves average RMSE by 11.00%, average CRPS by 9.25%, average MAE by 12.85%, and average FAC2 by 5.08% in the winter, and 1.94%, 1.63%, 2.49%, and 0.42% in the summer.

4.2. *Graphical comparisons*. We also present several diagnostic graphs to compare the prediction performance of different models. We define a prediction error as $\bar{\epsilon}_t(s) = \bar{y}_t(s) - y_t(s)$, where $\bar{y}_t(s)$ and $y_t(s)$ are the predicted and the observed $PM_{2.5}$ concentrations at site s, respectively. The empirical distributions of the prediction errors at all monitoring stations from five Bayesian methods are illustrated in Figure 5, along with CMAQ. It is clear that CMAQ, SVC, STVC, and STAR are biased in their predictions, while the predictions from HDCM and $HDCM_2$ are centered around 0. This implies that only HDCM and $HDCM_2$ produce calibrated data around the true values of $PM_{2.5}$ while the others fail to do the same. Furthermore, compared with other models, the prediction errors obtained from HDCM and $HDCM_2$ are more concentrated in the center with thinner tails, implying that the predictions of HDCM and $HDCM_2$ are also more precise than others.

We also adopt a diagnostic verification approach, the conditional quantile plots (see Chapter 8 of Wilks (2011)), to compare CMAQ predictions with the proposed method predictions. When the joint distribution of the predictions and observations are factorized into conditional distributions and marginal distributions, that is, $p(\bar{\mathbf{y}}, \mathbf{y}) = p(\mathbf{y}|\bar{\mathbf{y}})p(\bar{\mathbf{y}})$, the conditional distributions can be used to assess prediction quality. Conditional quantiles, in particular conditional medians, offer valuable information about conditional bias (or calibration). Any deviation of the conditional median from the 45° line indicates that the forecasts are biased, either

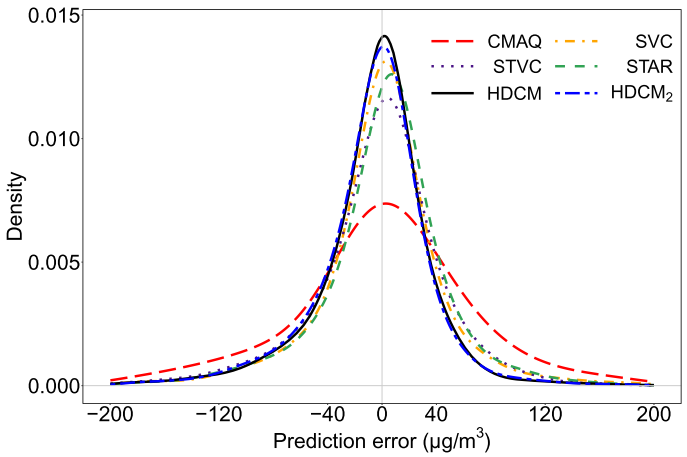
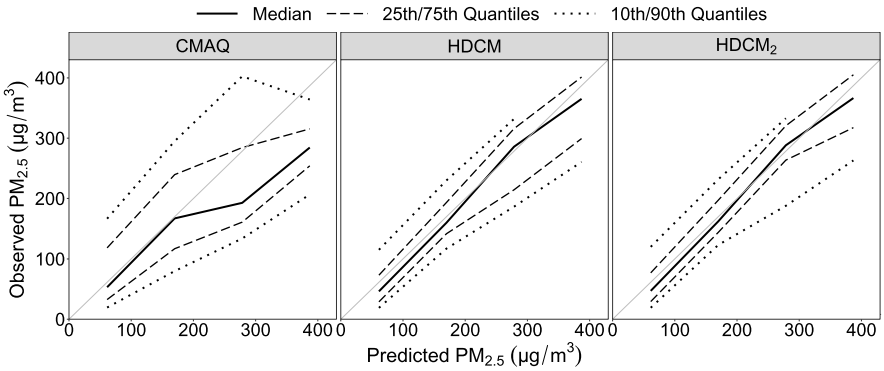


FIG. 5. *The empirical distribution of prediction errors of* $PM_{2.5}$ *concentrations was obtained using different methods at all monitoring sites from November* 1, 2015, *to January* 31, 2016.



(b) Conditional quantile plots from November 1, 2015, to January 31, 2016.

FIG. 6. *Conditional quantile plots of the observed* $PM_{2.5}$ *concentrations given predictions from CMAQ and predictions from HDCM and* $HDCM_2$ *at Langfang. The smoothed quantiles from the conditional distributions* $p(\boldsymbol{y}|\tilde{\boldsymbol{y}})$ *are drawn in relation to the* 45° *line.*

over- or underforecasting. As shown in Figure 6, compared with CMAQ that exhibits significant overfitting in the winter at Langfang, HDCM and $HDCM_2$ show much less conditional bias in their forecasts.

5. Calibrating CMAQ model outputs in the BTH region. Following the calibration specification in Section 3.4, we randomly sample 50 locations in each of CMAQ grid cells for a total of 124,950 spatial locations in the BTH region. We now perform a space-time calibration of CAMQ model outputs for the entire BTH region using the proposed HDCM.

5.1. *A dispersal process of heavy air pollution.* Figure 7 displays CMAQ outputs before and after calibration using HDCM from December 17 to December 22, 2015. During this period there was a heavy air pollution episode in the area. In each of the 13 cities, the average $PM_{2.5}$ concentration of all the stations in the city is marked using a solid square. The smoother the transition from cities to rural areas, the better the overall calibration results. It is evident that the before-calibration CMAQ outputs do not match well with most of the pollution data. After calibration the transition from the cities to their surrounding areas becomes much smoother, demonstrating the effectiveness of the proposed method. Figure 8

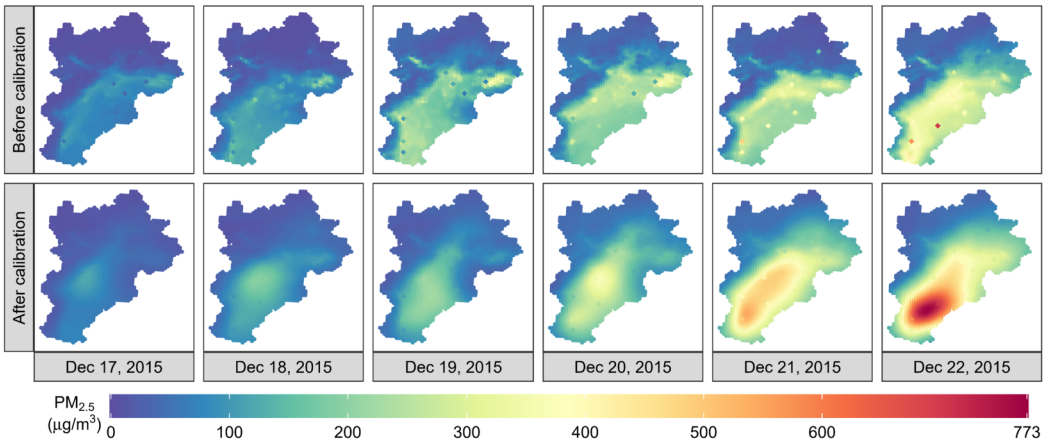


FIG. 7. *CMAQ numerical model outputs before and after HDCM calibration from December* 17 *to December* 22, 2015. *The solid dots represent the average* $PM_{2.5}$ *levels at the monitoring stations.*

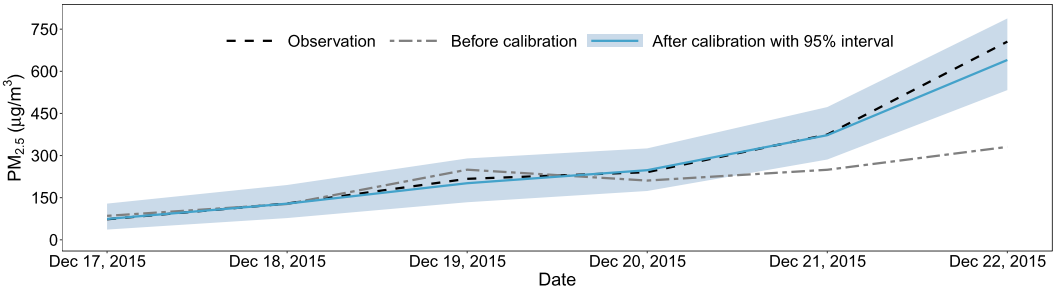


FIG. 8. *Time series plot of observed* $PM_{2.5}$ *concentrations, the precalibration CMAQ outputs, and the postcalibration CMAQ outputs (median and 95% credible intervals obtained using HDCM) at Hengshui, from December* 17 *to December* 22, 2015.

illustrates the difference between the before-calibration CMAQ outputs and after-calibration outputs with 95% credible intervals in Hengshui. In summary, HDCM proves to be an effective calibration method for CMAQ model outputs by providing accurate predictions, creating continuous spatial maps, and tracing the path of pollutants.

5.2. *Model diagnostics*. We also investigate the goodness-of-fit of the standard CMAQ model and the proposed HDCM model using their respective residuals. Residuals are defined as $\hat{\epsilon}_t(s) = \hat{y}_t(s) - y_t(s)$, where $y_t(s)$ is the concentration of observed $PM_{2.5}$ and $\hat{y}_t(s)$ is the model output that could be from the standard CMAQ model or the output from the postcalibration model using HDCM. As shown in Figure 9, HDCM residuals are more concentrated around 0 with thinner tails than the residuals obtained from the standard CMAQ model in the winter of 2015, demonstrating that HDCM provides a much better fit to the pollution data than CMAQ model.

6. Scalability for large datasets. To illustrate the effectiveness and efficiency of the proposed method for large datasets, we use two other $PM_{2.5}$ concentration datasets from November 1, 2015 to November 30, 2015. The first data are CMAQ model outputs at a *nine*-km scale with a total of 16,093 grid cells, and the second data are a reanalysis dataset from NAQPMS with a total of 6382 grid cells (Kong et al. (2020)). Both datasets cover an area that is much larger than the BTH region; refer to Figure S3 of the Supplementary Material. In bias correction reanalysis datasets are often used as observations to calibrate raw model

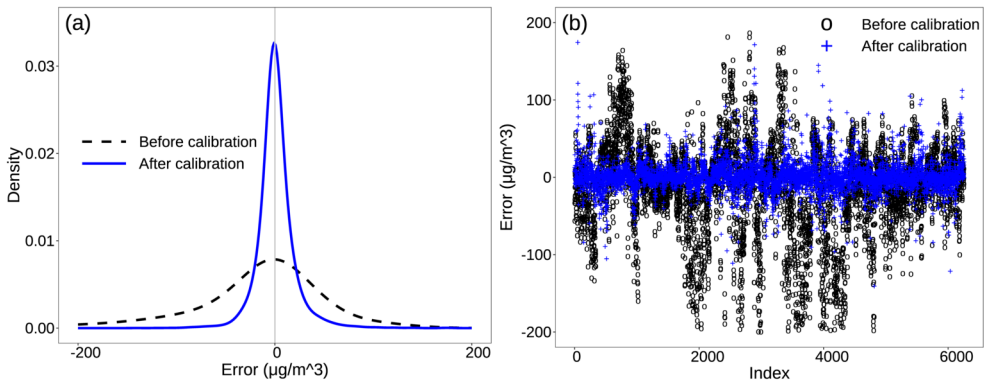


FIG. 9. (*a*) *Probability density plot of the residuals obtained from CMAQ* (*i.e., before calibration*) *and the residuals from HDCM* (*i.e., after calibration*) *and* (*b*) *scatter plot of the residuals between November* 1, 2015, *and January* 31, 2016, *at all monitoring stations.*

outputs; see He et al. (2019), Hersbach et al. (2020), Han et al. (2021). For this data analysis, the nearest grid cells of CMAQ are used to connect to the grids of the reanalysis data.

6.1. *Data preparation and model specification*. In this data analysis, the training data contain $N_t = 30$ time points from November 1, 2015 to November 30, 2015, and at each time point, there is the same set of $n = 5000$ spatial observations that are randomly selected from the 6382 reanalysis grid cells. The remaining observations are set as test data. For this large spatiotemporal dataset, our proposed models are compared against only STAR, which is generally more computationally efficient and produces better prediction performance than other methods, as seen in Section 4.

For HDCM, because observations are uniformly located in space, we partition the spatial domain into R subregions through an equal area partition method (Heaton et al. (2019)). Additional numerical experiments show that, when the observations are dense, HDCM is robust for the choice of R. We consider $R = \{9, 16\}$ here. We also set the number of triangle vertices in the construction of triangulated meshes to $m = 2042$, 3158, or 10,103. The same triangulated meshes are also used in STAR. As an illustration, Figure S4 of the Supplementary Material presents the triangulated mesh with $R = 9$ and $m = 10,103$.

6.2. *Performance comparison*. Table 3 shows that HDCM predictions are better than those from STAR using RMSE, CRPS, and MAE for $m = 2042$ or 3158. When the number of triangle vertices is large, for example, $m = 10,103$, STAR fails to produce sensible results using `INLA` (Lindgren and Rue (2015)), while HDCM achieves even better predictions compared to when the number of vertices is small. The RMSE, CRPS, and MAE decrease when the number of vertices increases, suggesting that the performance of HDCM and STAR improves for finer triangulated meshes.

6.3. *Computational comparison*. The computational efficiency of HDCM is assessed by comparing it to SVC and STAR. All models were run on a 10-core workstation with a 3.8 GHz Intel Core i9-10900KF processor and 128 GB memory.

The computational advantages of HDCM are evident in both the model fitting stage and the prediction stage. In each iteration of the model fitting stage, for $m > n$ or $n > m$, the proposed

TABLE 3

Averaged RMSE (*Root Mean Squared Error*), *CRPS* (*Continuous Rank Probability Score*), *MAE* (*Mean Absolute Error*), *and FAC*2 (*Fraction of Predictions within a factor of* 2) *for* $PM_{2.5}$ *concentration predictions calculated for the test data from two methods*: (1) *the first-order spatiotemporal autoregression* (*STAR*) *based on the integrated nested Laplace approximation* (*INLA*, *Rue, Martino and Chopin* (2009)) *and* (2) *space-partitioning-based HDCM with nine or* 16 *subregions based on the proposed VB-spEnKS. Daily data from November* 1, 2015 *to November* 31, 2015 *are considered*

n	m	Model	c_s	RMSE	CRPS	MAE	FAC2	Running time (seconds)
5000	2042	STAR	–	3.821	1.311	1.597	1.000	32,093
		$HDCM_9$	0.050	3.006	1.152	1.476	1.000	1582
		$HDCM_{16}$	0.080	2.876	1.122	1.424	1.000	1778
	3158	STAR	–	3.296	1.160	1.413	0.998	58,995
		$HDCM_9$	0.050	2.706	1.125	1.364	1.000	2702
		$HDCM_{16}$	0.080	2.670	1.099	1.358	1.000	2727
	10,103	STAR	–	–	–	–	–	–
		$HDCM_9$	0.050	2.559	1.018	1.287	1.000	23,370
		$HDCM_{16}$	0.080	2.519	0.994	1.250	1.000	12,677

HDCM inverts either an $n \times n$ matrix that is not too dense by using (8) or an $m \times m$ sparse matrix by using the Sherman–Morrison–Woodbury formula, while the spatial SVC needs to invert an $n \times n$ dense matrix. In the prediction stage, for each spatiotemporal point, HDCM requires no more than $O(c_h^2 m N_e)$ flops operations for producing N_e predictive samples with (12), whereas SVC requires $O(n^2 m_c)$ flops operations for generating m_c MCMC samples with kriging (Chen and Stein (2023)). Note that $c_h^2 m$ and N_e are usually much smaller than n^2 and m_c, respectively, since $c_h \in (0, 1]$, and m_c of SVC is much larger than N_e of our HDCM in almost all applications. For example, Berrocal et al. (2020) use $m_c = 5000$ after a burn-in of 5000 in SVC, while $N_e \leq 100$ is frequently seen in earlier studies (Mitchell, Houtekamer and Pellerin (2002), Houtekamer and Zhang (2016)).

Considering the two large datasets in this section with 5000×30 and 1382×30 spatiotemporal points, HDCM is almost 80 times faster than SVC and 10 times faster than STAR, as seen in Table 3 for $m = 2042$ or 3158. In the current procedure of VB-spEnKS, IDEs of different subregions are estimated by simple for-loops; thus, computational efficiency of HDCM could be further improved through parallel computing.

7. Concluding remarks. In this work we proposed a novel method to calibrate the raw outputs of numerical models using observations. Extensive numerical and graphical comparisons demonstrated that the proposed approach is, in general, more effective and efficient than other competing methods.

The proposed HDCM can be applied to a broad range of situations. It allows the spatial support of latent processes to be different from that of observations, which is particularly useful when n is very large. It can handle cases where the number of grid cells is very large. Through some appropriate redistribution kernels of IDE models, HDCM is capable of describing the dynamic evolution of spatial interactions. The triangulation scheme applied to IDE models enables the modeling of data variations in both space and time, even for very large datasets.

HDCM can be easily scaled up to massive datasets due to the well-designed VB-spEnKS. In geoscience research, as grids of numerical model forecasts are usually on a kilometric scale, the entire spatial field may contain several million grid points (Vannitsem et al. (2021)). Using China as an example, a common resolution for the whole of China is at 15 km (Kong et al. (2020)). The numerical model outputs at this scale contain nearly 150,000 grid cells, and the size of grid cells is $O(10^5)$, where the big "O" is the same as that in Houtekamer and Zhang (2016). The largest datasets we could obtain from the BTH region are moderately large (i.e., $O(10^4)$). Nevertheless, the proposed method can handle datasets that are $O(10^5)$ because its computational cost depends mostly on the number of triangle vertices rather than the number of spatial observations or the numerical model grid cells, as seen in Section 3.1 and Section 3.2.4.

Our work is subject to limitations that need further study. First, tuning parameters usually depend on the distributions of observations and the triangle vertices, and they could also be interrelated. Their relationships could be analyzed using an experimental method similar to Kirchgessner, Nerger and Bunse-Gerstner (2014), but they are not investigated in this work. Second, the normality and linearity assumptions adopted here may not be the most appropriate distributional assumptions for the air pollution data. The pollutant data in the BTH region appear to be heavy-tailed (Figure S1 of the Supplementary Material). There is room for improvement for HDCM, though it is rather robust to deviations from the assumptions as shown in Section S8 of the Supplementary Material.

The proposed framework can be extended to meet a wide variety of needs. Nonlinear IDEs can be included within the modeling framework to better approximate more complex physical processes. When HDCM is applied for forecasting purposes, the nonlinearity of IDEs can be

extremely beneficial for prediction performance; see Zammit-Mangion and Wikle (2020). Moreover, because $PM_{2.5}$ concentrations are often affected by meteorological conditions in a nonlinear fashion, for example, wind direction and speed (Liang et al. (2015)), HDCM can conveniently include nonlinear additive models to account for nonlinear relationships between air pollutants and other covariates. Another avenue of research is to consider more general frameworks such as joint calibration models of multiple pollutants.

Acknowledgments. We thank the anonymous reviewers and the Associate Editor for their invaluable comments. We are particularly grateful to the Editor, Professor Karen Kafadar (former Editor-in-Chief), for her detailed comments and insightful suggestions. Their input significantly improved the quality of this article.

The first and the second authors are equal contributors.

Hui Huang is the corresponding author.

Funding. This research is supported by the National Natural Science Foundation of China (Grant No. 12231017, No. 12292984, and No. 12161016), the MOE Project of Key Research Institute of Humanities and Social Sciences (Grant No. 22JJD910001), and the Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science (project code 2022B1212010006).

SUPPLEMENTARY MATERIAL

Supplementary material for the paper "Efficient and effective calibration of numerical model outputs using hierarchical dynamic models" (DOI: 10.1214/23-AOAS1823SUPPA; .pdf). We describe the details related to the proposed approach HDCM, including the datasets, the competitive models, the selection of the tuning parameters, the simulation study for the proposed VB-spEnKS, the sensitivity of the proposed HDCM with respect to the selection of the tuning parameters, the cross-validation for China's BTH data in the summer of 2015, the Diebold-Mariano test, coordinate ascent mean-field VB and the Laplace approximation, etc.

R code and data for the paper "Efficient and effective calibration of numerical model outputs using hierarchical dynamic models" (DOI: 10.1214/23-AOAS1823SUPPB; .zip). We provide the R code for fitting our hierarchical dynamic model with variational Bayes and space-partitioning-based ensemble Kalman smoother. We also include $PM_{2.5}$ concentration datasets in China's BTH region to illustrate our calibration method.

REFERENCES

APPEL, K. W., NAPELENOK, S. L., FOLEY, K. M., PYE, H. O. T., HOGREFE, C., LUECKEN, D. J., BASH, J. O., ROSELLE, S. J., PLEIM, J. E. et al. (2017). Description and evaluation of the Community Multiscale Air Quality (CMAQ) modeling system version 5.1. *Geosci. Model Dev.* **10** 1703–1732.

BAKAR, K. S., KOKIC, P. and JIN, H. (2016). Hierarchical spatially varying coefficient and temporal dynamic process models using spTDyn. *J. Stat. Comput. Simul.* **86** 820–840. MR3432519 https://doi.org/10.1080/00949655.2015.1038267

BANERJEE, S., GELFAND, A. E., FINLEY, A. O. and SANG, H. (2008). Gaussian predictive process models for large spatial data sets. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 825–848. MR2523906 https://doi.org/10.1111/j.1467-9868.2008.00663.x

BERROCAL, V. J., GELFAND, A. E. and HOLLAND, D. M. (2010). A spatio-temporal downscaler for output from numerical models. *J. Agric. Biol. Environ. Stat.* **15** 176–197. MR2787270 https://doi.org/10.1007/s13253-009-0004-z

BERROCAL, V. J., GELFAND, A. E. and HOLLAND, D. M. (2012). Space-time data fusion under error in computer model output: An application to modeling air quality. *Biometrics* **68** 837–848. MR3055188 https://doi.org/10.1111/j.1541-0420.2011.01725.x

BERROCAL, V. J., GUAN, Y., MUYSKENS, A., WANG, H., REICH, B. J., MULHOLLAND, J. A. and CHANG, H. H. (2020). A comparison of statistical and machine learning methods for creating national daily maps of ambient $PM_{2.5}$ concentration. *Atmos. Environ.* **222** 117130.

BLANGIARDO, M. and CAMELETTI, M. (2015). *Spatial and Spatio-Temporal Bayesian Models with R-INLA*. Wiley, Chichester. MR3364017

BLEI, D. M., KUCUKELBIR, A. and MCAULIFFE, J. D. (2017). Variational inference: A review for statisticians. *J. Amer. Statist. Assoc.* **112** 859–877. MR3671776 https://doi.org/10.1080/01621459.2017.1285773

BOLIN, D., WALLIN, J. and LINDGREN, F. (2019). Latent Gaussian random field mixture models. *Comput. Statist. Data Anal.* **130** 80–93. MR3860530 https://doi.org/10.1016/j.csda.2018.08.007

BYUN, D. and SCHERE, K. L. (2006). Review of the governing equations, computational algorithms, and other components of the Models-3 Community Multiscale Air Quality (CMAQ) modeling system. *Appl. Mech. Rev.* **59** 51–77.

CHANG, J. C. and HANNA, S. R. (2004). Air quality model performance evaluation. *Meteorol. Atmos. Phys.* **87** 167–196.

CHEN, J. and STEIN, M. L. (2023). Linear-cost covariance functions for Gaussian random fields. *J. Amer. Statist. Assoc.* **118** 147–164. MR4571113 https://doi.org/10.1080/01621459.2021.1919122

CHEN, Y., CHANG, X., LUO, F. and HUANG, H. (2023). Additive dynamic models for correcting numerical model outputs. *Comput. Statist. Data Anal.* **187** Paper No. 107799, 21 pp. MR4604820 https://doi.org/10.1016/j.csda.2023.107799

CHEN, Y., CHANG, X., ZHANG, B. and HUANG, H. (2024). Supplement to "Efficient and effective calibration of numerical model outputs using hierarchical dynamic models." https://doi.org/10.1214/23-AOAS1823SUPPA, https://doi.org/10.1214/23-AOAS1823SUPPB

CHINA'S STATE COUNCIL (2013). The action plan for air pollution prevention and control. Available at http://www.gov.cn/zwgk/2013-09/12/content_2486773.htm. In Chinese.

CHINA'S STATE COUNCIL (2018). The three-year action plan for winning the blue sky defense battle. Available at http://www.gov.cn/xinwen/2018-07/03/content_5303212.htm. In Chinese.

CHINA'S STATE COUNCIL (2021). The long-term "Beautiful China" targets through 2035. Available at http://www.gov.cn/xinwen/2021-02/22/content_5588304.htm. In Chinese.

CRESSIE, N. and JOHANNESSON, G. (2008). Fixed rank Kriging for very large spatial data sets. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 209–226. MR2412639 https://doi.org/10.1111/j.1467-9868.2007.00633.x

CRESSIE, N. and WIKLE, C. K. (2011). *Statistics for Spatio-Temporal Data. Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR2848400

DATTA, A., BANERJEE, S., FINLEY, A. O. and GELFAND, A. E. (2016). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *J. Amer. Statist. Assoc.* **111** 800–812. MR3538706 https://doi.org/10.1080/01621459.2015.1044091

EVENSEN, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res., Oceans* **99** 10143–10162.

EVENSEN, G. and VAN LEEUWEN, P. J. (2000). An ensemble Kalman smoother for nonlinear dynamics. *Mon. Weather Rev.* **128** 1852–1867.

FUENTES, M. and RAFTERY, A. E. (2005). Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. *Biometrics* **61** 36–45. MR2129199 https://doi.org/10.1111/j.0006-341X.2005.030821.x

FURRER, R., GENTON, M. G. and NYCHKA, D. (2006). Covariance tapering for interpolation of large spatial datasets. *J. Comput. Graph. Statist.* **15** 502–523. MR2291261 https://doi.org/10.1198/106186006X132178

GELFAND, A. E., KIM, H.-J., SIRMANS, C. F. and BANERJEE, S. (2003). Spatial modeling with spatially varying coefficient processes. *J. Amer. Statist. Assoc.* **98** 387–396. MR1995715 https://doi.org/10.1198/016214503000170

GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. MR2345548 https://doi.org/10.1198/016214506000001437

GUAN, Y., JOHNSON, M. C., KATZFUSS, M., MANNSHARDT, E., MESSIER, K. P., REICH, B. J. and SONG, J. J. (2020). Fine-scale spatiotemporal air pollution analysis using mobile monitors on Google Street View vehicles. *J. Amer. Statist. Assoc.* **115** 1111–1124. MR4143453 https://doi.org/10.1080/01621459.2019.1665526

GUILLAS, S., BAO, J., CHOI, Y. and WANG, Y. (2008). Statistical correction and downscaling of chemical transport model ozone forecasts over Atlanta. *Atmos. Environ.* **42** 1338–1348.

HAN, L., CHEN, M., CHEN, K., CHEN, H., ZHANG, Y., LU, B., SONG, L. and QIN, R. (2021). A deep learning method for bias correction of ECMWF 24–240 h forecasts. *Adv. Atmos. Sci.* **38** 1444–1459.

HARVEY, D., LEYBOURNE, S. and NEWBOLD, P. (1997). Testing the equality of prediction mean squared errors. *Int. J. Forecast.* **13** 281–291.

HE, D., ZHOU, Z., KANG, Z. and LIU, L. (2019). Numerical studies on forecast error correction of GRAPES model with variational approach. *Adv. Meteorol.* **2019** 1–13.

HEATON, M. J., CHRISTENSEN, W. F. and TERRES, M. A. (2017). Nonstationary Gaussian process models using spatial hierarchical clustering from finite differences. *Technometrics* **59** 93–101. MR3604192 https://doi.org/10.1080/00401706.2015.1102763

HEATON, M. J., DATTA, A., FINLEY, A. O. et al. (2019). A case study competition among methods for analyzing large spatial data. *J. Agric. Biol. Environ. Stat.* **24** 398–425. MR3996451 https://doi.org/10.1007/s13253-018-00348-w

HEINRICH, C., HELLTON, K. H., LENKOSKI, A. and THORARINSDOTTIR, T. L. (2021). Multivariate postprocessing methods for high-dimensional seasonal weather forecasts. *J. Amer. Statist. Assoc.* **116** 1048–1059. MR4309249 https://doi.org/10.1080/01621459.2020.1769634

HERSBACH, H., BELL, B., BERRISFORD, P., HIRAHARA, S., HORÁNYI, A., MUÑOZ-SABATER, J., NICOLAS, J., PEUBEY, C., RADU, R. et al. (2020). The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **146** 1999–2049.

HIGDON, D., GATTIKER, J., WILLIAMS, B. and RIGHTLEY, M. (2008). Computer model calibration using high-dimensional output. *J. Amer. Statist. Assoc.* **103** 570–583. MR2523994 https://doi.org/10.1198/016214507000000888

HOUTEKAMER, P. L., MITCHELL, H. L., PELLERIN, G., BUEHNER, M., CHARRON, M., SPACEK, L. and HANSEN, B. (2005). Atmospheric data assimilation with an ensemble Kalman filter: Results with real observations. *Mon. Weather Rev.* **133** 604–620.

HOUTEKAMER, P. L. and ZHANG, F. (2016). Review of the ensemble Kalman filter for atmospheric data assimilation. *Mon. Weather Rev.* **144** 4489–4532.

ISHIGURO, K., SATO, I. and UEDA, N. (2017). Averaged collapsed variational Bayes inference. *J. Mach. Learn. Res.* **18** Paper. No. 1, 29 pp. MR3625705

JIANG, X. and YOO, E.-H. E. (2019). Modeling wildland fire-specific $PM_{2.5}$ concentrations for uncertainty-aware health impact assessments. *Environ. Sci. Technol.* **53** 11828–11839. https://doi.org/10.1021/acs.est.9b02660

KATZFUSS, M., STROUD, J. R. and WIKLE, C. K. (2016). Understanding the ensemble Kalman filter. *Amer. Statist.* **70** 350–357. MR3574787 https://doi.org/10.1080/00031305.2016.1141709

KATZFUSS, M., STROUD, J. R. and WIKLE, C. K. (2020). Ensemble Kalman methods for high-dimensional hierarchical dynamic space-time models. *J. Amer. Statist. Assoc.* **115** 866–885. MR4107685 https://doi.org/10.1080/01621459.2019.1592753

KAUFMAN, C. G., SCHERVISH, M. J. and NYCHKA, D. W. (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *J. Amer. Statist. Assoc.* **103** 1545–1555. MR2504203 https://doi.org/10.1198/016214508000000959

KENNEDY, M. C. and O'HAGAN, A. (2001). Bayesian calibration of computer models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 425–464. MR1858398 https://doi.org/10.1111/1467-9868.00294

KIM, H.-M., MALLICK, B. K. and HOLMES, C. C. (2005). Analyzing nonstationary spatial data using piecewise Gaussian processes. *J. Amer. Statist. Assoc.* **100** 653–668. MR2160567 https://doi.org/10.1198/016214504000002014

KIRCHGESSNER, P., NERGER, L. and BUNSE-GERSTNER, A. (2014). On the choice of an optimal localization radius in ensemble Kalman filter methods. *Mon. Weather Rev.* **142** 2165–2175.

KONG, L., TANG, X., ZHU, J., WANG, Z., WU, H. and LI, J. (2020). Developing high-resolution air quality reanalysis dataset over China for years 2013–2018 based on ensemble Kalman filter and surface observations from CNEMC. In *EGU General Assembly Conference Abstracts* 6848.

KONOMI, B. A., SANG, H. and MALLICK, B. K. (2014). Adaptive Bayesian nonstationary modeling for large spatial datasets using covariance approximations. *J. Comput. Graph. Statist.* **23** 802–829. MR3224657 https://doi.org/10.1080/10618600.2013.812872

KOT, M., LEWIS, M. A. and VAN DEN DRIESSCHE, P. (1996). Dispersal data and the spread of invading organisms. *Ecology* **77** 2027–2042.

LIANG, D., ZHANG, H., CHANG, X. and HUANG, H. (2021). Modeling and regionalization of China's $PM_{2.5}$ using spatial-functional mixture models. *J. Amer. Statist. Assoc.* **116** 116–132. MR4227679 https://doi.org/10.1080/01621459.2020.1764363

LIANG, X., ZOU, T., GUO, B., LI, S., ZHANG, H., ZHANG, S., HUANG, H. and CHEN, S. X. (2015). Assessing Beijing's $PM_{2.5}$ pollution: Severity, weather impact, APEC and winter heating. *Proc. R. Soc. A, Math. Phys. Eng. Sci.* **471** 20150257.

LINDGREN, F. and RUE, H. (2015). Bayesian spatial modelling with R-INLA. *J. Stat. Softw.* **63** 1–25.

LINDGREN, F., RUE, H. and LINDSTRÖM, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 423–498. MR2853727 https://doi.org/10.1111/j.1467-9868.2011.00777.x

LU, X., ZHANG, S., XING, J., WANG, Y., CHEN, W., DING, D., WU, Y., WANG, S., DUAN, L. et al. (2020). Progress of air pollution control in China and its challenges and opportunities in the ecological civilization era. *Engineering* **6** 1423–1431.

MCMILLAN, N. J., HOLLAND, D. M., MORARA, M. and FENG, J. (2010). Combining numerical model output and particulate data using Bayesian space-time modeling. *Environmetrics* **21** 48–65. MR2842223 https://doi.org/10.1002/env.984

MITCHELL, H. L., HOUTEKAMER, P. L. and PELLERIN, G. (2002). Ensemble size, balance, and model-error representation in an ensemble Kalman filter. *Mon. Weather Rev.* **130** 2791–2808.

NYCHKA, D., BANDYOPADHYAY, S., HAMMERLING, D., LINDGREN, F. and SAIN, S. (2015). A multiresolution Gaussian process model for the analysis of large spatial datasets. *J. Comput. Graph. Statist.* **24** 579–599. MR3357396 https://doi.org/10.1080/10618600.2014.914946

QI, J., ZHENG, B., LI, M., YU, F., CHEN, C., LIU, F., ZHOU, X., YUAN, J., ZHANG, Q. et al. (2017). A high-resolution air pollutants emission inventory in 2013 for the Beijing–Tianjin–Hebei region, China. *Atmos. Environ.* **170** 156–168.

REN, Q., BANERJEE, S., FINLEY, A. O. and HODGES, J. S. (2011). Variational Bayesian methods for spatial data analysis. *Comput. Statist. Data Anal.* **55** 3197–3217. MR2825404 https://doi.org/10.1016/j.csda.2011.05.021

RICHARDSON, R., KOTTAS, A. and SANSÓ, B. (2017). Flexible integro-difference equation modeling for spatio-temporal data. *Comput. Statist. Data Anal.* **109** 182–198. MR3603648 https://doi.org/10.1016/j.csda.2016.11.011

RODU, J. and KAFADAR, K. (2022). The q-q boxplot. *J. Comput. Graph. Statist.* **31** 26–39. MR4387208 https://doi.org/10.1080/10618600.2021.1938586

RUE, H. and HELD, L. (2005). *Gaussian Markov Random Fields*: *Theory and Applications*. *Monographs on Statistics and Applied Probability* **104**. CRC Press/CRC, Boca Raton, FL. MR2130347 https://doi.org/10.1201/9780203492024

RUE, H., MARTINO, S. and CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 319–392. MR2649602 https://doi.org/10.1111/j.1467-9868.2008.00700.x

SAHU, S. K., GELFAND, A. E. and HOLLAND, D. M. (2006). Spatiotemporal modeling of fine particulate matter. *J. Agric. Biol. Environ. Stat.* **11** 61–86.

SALTER, J. M., WILLIAMSON, D. B., SCINOCCA, J. and KHARIN, V. (2019). Uncertainty quantification for computer models with spatial output using calibration-optimal bases. *J. Amer. Statist. Assoc.* **114** 1800–1814. MR4047301 https://doi.org/10.1080/01621459.2018.1514306

SANG, H., JUN, M. and HUANG, J. Z. (2011). Covariance approximation for large multivariate spatial data sets with an application to multiple climate model errors. *Ann. Appl. Stat.* **5** 2519–2548. MR2907125 https://doi.org/10.1214/11-AOAS478

SHADDICK, G., THOMAS, M. L., GREEN, A. et al. (2018). Data integration model for air quality: A hierarchical approach to the global estimation of exposures to ambient air pollution. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **67** 231–253. MR3758764 https://doi.org/10.1111/rssc.12227

STROUD, J. R., STEIN, M. L., LESHT, B. M., SCHWAB, D. J. and BELETSKY, D. (2010). An ensemble Kalman filter and smoother for satellite data assimilation. *J. Amer. Statist. Assoc.* **105** 978–990. MR2752594 https://doi.org/10.1198/jasa.2010.ap07636

TABOUY, T., BARBILLON, P. and CHIQUET, J. (2020). Variational inference for stochastic block models from sampled data. *J. Amer. Statist. Assoc.* **115** 455–466. MR4078475 https://doi.org/10.1080/01621459.2018.1562934

VANNITSEM, S., BREMNES, J. B., DEMAEYER, J., EVANS, G. R., FLOWERDEW, J., HEMRI, S., LERCH, S., ROBERTS, N., THEIS, S. et al. (2021). Statistical postprocessing for weather forecasts: Review, challenges, and avenues in a big data world. *Bull. Am. Meteorol. Soc.* **102** E681–E699.

WAN, Y., XU, M., HUANG, H. and CHEN, S. X. (2021). A spatio-temporal model for the analysis and prediction of fine particulate matter concentration in Beijing. *Environmetrics* **32** Paper No. e2648, 16 pp. MR4207558 https://doi.org/10.1002/env.2648

WANG, Y., DU, Y., WANG, J. and LI, T. (2019). Calibration of a low-cost $PM_{2.5}$ monitor using a random forest model. *Environ. Int.* **133** 105161.

WANG, Z. F., XIE, F. Y., WANG, X. Q., AN, J. and ZHU, J. (2006). Development and application of nested air quality prediction modeling system (in Chinese). *Chin. J. Atmos. Sci.* **30** 778–790.

WENDLAND, H. (1995). Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Adv. Comput. Math.* **4** 389–396. MR1366510 https://doi.org/10.1007/BF02123482

WIKLE, C. K. (2002). A kernel-based spectral model for non-Gaussian spatio-temporal processes. *Stat. Model.* **2** 299–314. MR1951587 https://doi.org/10.1191/1471082x02st036oa

WIKLE, C. K. and HOLAN, S. H. (2011). Polynomial nonlinear spatio-temporal integro-difference equation models. *J. Time Series Anal.* **32** 339–350. MR2841788 https://doi.org/10.1111/j.1467-9892.2011.00729.x

WIKLE, C. K., ZAMMIT-MANGION, A. and CRESSIE, N. (2019). *Spatiotemporal Statistics with R*. CRC Press/CRC, Boca Raton.

WILKS, D. S. (2011). *Statistical Methods in the Atmospheric Sciences*, 3rd ed. Academic Press, Waltham, MA.

XU, K., WIKLE, C. K. and FOX, N. I. (2005). A kernel-based spatio-temporal dynamical model for nowcasting weather radar reflectivities. *J. Amer. Statist. Assoc.* **100** 1133–1144. MR2236929 https://doi.org/10.1198/016214505000000682

ZAMMIT-MANGION, A. and WIKLE, C. K. (2020). Deep integro-difference equation models for spatio-temporal forecasting. *Spat. Stat.* **37** 100408, 20 pp. MR4109596 https://doi.org/10.1016/j.spasta.2020.100408

ZHANG, B., SANG, H. and HUANG, J. Z. (2015). Full-scale approximations of spatio-temporal covariance models for large datasets. *Statist. Sinica* **25** 99–114. MR3328805

ZHANG, L., SHAO, J., LU, X., ZHAO, Y., HU, Y., HENZE, D. K., LIAO, H., GONG, S. and ZHANG, Q. (2016). Sources and processes affecting fine particulate matter pollution over North China: An adjoint analysis of the Beijing APEC period. *Environ. Sci. Technol.* **50** 8731–8740.

ZHAO, C., WANG, Q., BAN, J., LIU, Z., ZHANG, Y., MA, R., LI, S. and LI, T. (2020). Estimating the daily $PM_{2.5}$ concentration in the Beijing–Tianjin–Hebei region using a random forest model with a $0.01^\circ \times 0.01^\circ$ spatial resolution. *Environ. Int.* **134** 105297.