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

BY YEWEN CHEN [†1], XIAOHUI CHANG [†2], BOHAI ZHANG[3] AND HUI HUANG[*4]

[1]*College of Public Health, University of Georgia, Yewen.Chen@uga.edu*

[2]*College of Business, Oregon State University, xiaohui.chang@oregonstate.edu*

[3]*Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science, BNU-HKBU United International College, bohaizhang@uic.edu.cn*

[4]*Center for Applied Statistics and School of Statistics, Renmin University of China, huangh89@mail.sysu.edu.cn*

In this document, we provide a comprehensive overview of the proposed HDCM, including the used 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 Beijing-Tianjin-Hebei (BTH) data in the summer of 2015, the Diebold-Mariano test, and the Laplace approximation, etc.

S1. Datasets from multiple sources. The effectiveness and efficiency of the proposed HDCM are demonstrated using two datasets.

(1) Moderately large datasets. These consist of $PM_{2.5}$ concentrations obtained from monitoring stations and the outputs of the Community Multiscale Air Quality (CMAQ) system for China's Beijing-Tianjin-Hebei (BTH) region. The dataset sizes are as follows: $68 \times 92 = 6{,}256$ spatio-temporal observations for monitoring stations and $5{,}587 \times 92 = 514{,}004$ raw spatio-temporal outputs for CMAQ, where the number of monitroing stations is 68, the number of CMAQ grid cells is $5{,}587$, and the time length is 92 days.

(2) Large datasets. These include the reanalysis of $PM_{2.5}$ outputs of the Nested Air Quality Prediction Modeling System (NAQPMS) and the raw $PM_{2.5}$ outputs of the CMAQ system. The dataset sizes are $6,382 \times 30 = 191{,}460$ spatio-temporal gridded outputs for NAQPMS and $16,093 \times 30 = 482,790$ raw gridded outputs for CMAQ, respectively.

S1.1. *Data feature in the BTH region.* As seen in Table S1, the variation of $PM_{2.5}$ concentration in most cities of the BTH region is significantly lower in summer than in winter. In general, the calibration for the winter outputs of the numerical model is more challenging than that for the summer outputs.

[†]Equal contribution.

[*]Corresponding author. Hui Huang, huangh89@mail.sysu.edu.cn.

TABLE S1

Mean, standard deviations (SD), and coefficient of variation (CV) for $PM_{2.5}$ concentration predictions ($\mu g/m^3$) calculated for 13 cities in the BTH region in both seasons of 2015.

City	Summer of 2015			Winter of 2015		
	Mean	SD	CV	Mean	SD	CV
Baoding	65.63	32.96	0.50	159.18	111.94	0.70
Beijing	56.18	37.87	0.67	112.80	103.69	0.92
Cangzhou	54.54	25.35	0.46	104.90	78.85	0.75
Chengde	33.49	25.81	0.77	56.38	37.22	0.66
Handan	73.90	28.73	0.39	125.25	90.66	0.72
Hengshui	74.30	28.41	0.38	151.35	113.24	0.75
Langfang	54.74	29.36	0.54	129.38	104.31	0.81
Qinhuangdao	36.36	26.58	0.73	50.81	42.70	0.84
Shijiazhuang	65.81	34.38	0.52	141.47	106.42	0.75
Tangshan	63.17	31.88	0.50	106.37	77.33	0.73
Tianjin	50.18	25.25	0.50	100.05	81.39	0.81
Xingtai	78.08	34.23	0.44	148.89	121.04	0.81
Zhangjiakou	32.45	21.89	0.67	34.40	25.16	0.73
Average	56.83	29.44	0.55	109.33	84.15	0.77

S1.1.1. *Imputing missing data.* In Section 2 of the manuscript, missing data are imputed. Because the missing rate is only 0.28%, this imputation will not affect the data pattern. Of course, some extreme values may be filtered out when we summarize hourly observations to the daily data. However, some of the sites were missing more than 24 hours of hourly observations on some days, and this counts for about 0.28% of our data. To reduce the impact of these missing data on the predictions, we summarize hourly data to daily data and compare the prediction performance of different methods. It is worth noting that the proposed approach is able to directly handle datasets with some hourly data that are completely missing, as it allows imputing the missing data within VB-spEnKS procedure.

Although the proposed approach can impute missing data through VB-spEnKS procedure, missing data is imputed at this stage rather than at the fitting HDCM stage to ensure a fair comparison between different methods by using the same imputed data as the input for all models. In the model comparison, several non-Bayesian methods have difficulties in automatically imputing missing data within their implementation procedures.

S1.1.2. *Data transformation.* In this section, we explore data distributions using histograms (Berrocal, Gelfand and Holland, 2010) and quantile-quantile (Q-Q) boxplots (Rodu and Kafadar, 2022). As seen in Figure S1 (a) and Figure S1 (d), the observed $PM_{2.5}$ concentrations show clear right skewness. To stabilize the variance, we carry out two transformations for the observations, namely the square root transformation and logarithmic transformation (Sahu, Gelfand and Holland, 2006). histograms presented in Figures S1 and Q-Q plots presented in Figures S2 show that the square root transformed data are closer to the normal distribution compared to the log-transformed data. Therefore, in this work, we use the square-root transformation.

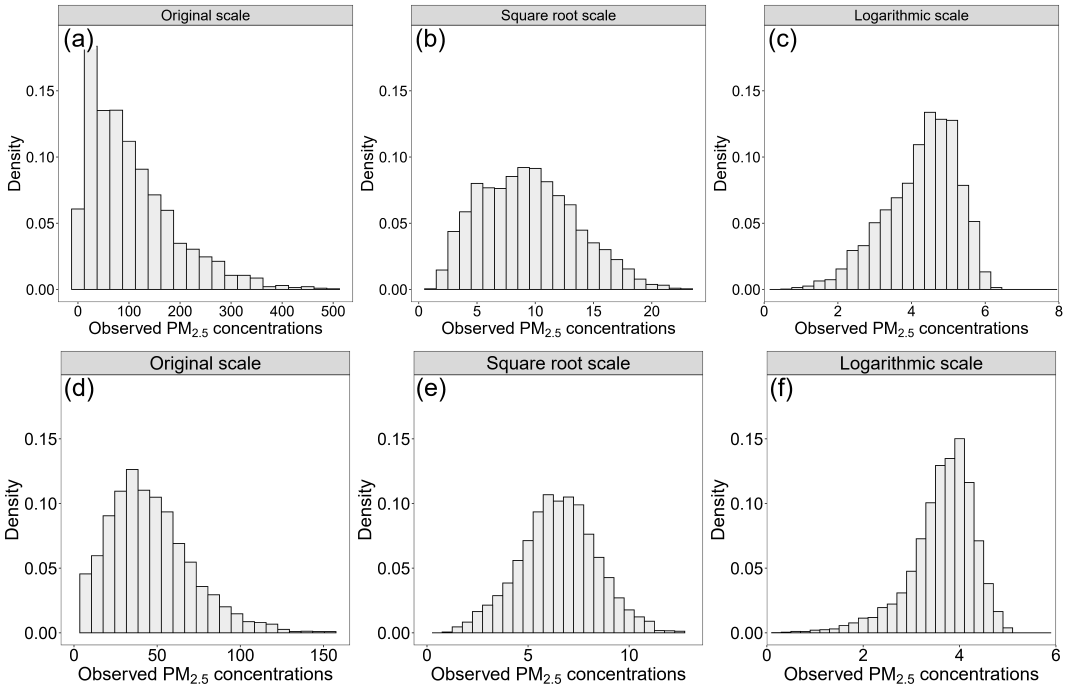

FIG S1. *Histograms of observed $PM_{2.5}$ concentrations at three scales, including the original, square root, and logarithmic scales. (a)-(c) In the winter of 2015. (d)-(f) In the summer of 2015.*

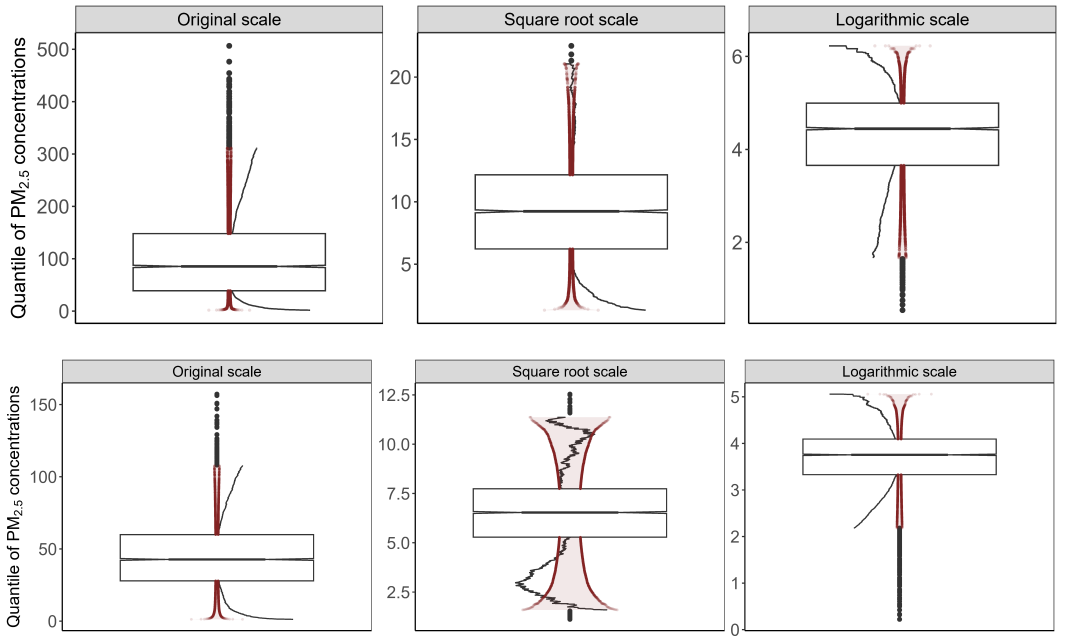

FIG S2. *Q-Q boxplots of observed $PM_{2.5}$ concentrations at three scales, including the original, square root, and logarithmic scales. (a)-(c) In the winter of 2015. (d)-(f) In the summer of 2015.*

S1.2. *Large datasets.* Figure S3 displays spatial distributions of the grid cells under different gridding systems, namely CMAQ and NAQPMS. Both datasets cover an area that is much larger than the BTH region.

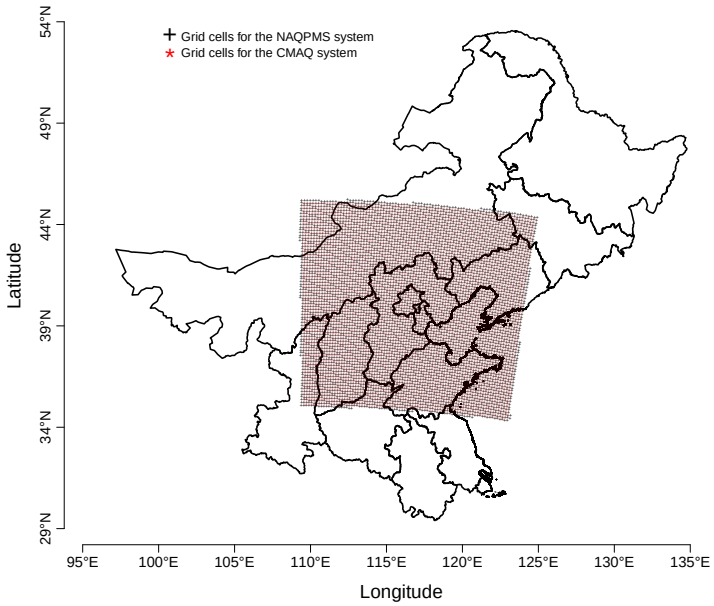

FIG S3. *Maps of grid cells of CMAQ and NAQPMS. The symbols "+" represent the centroids of 16, 093 9km CMAQ grids. The symbols "*" denote the centroids of* 6,382 *15km NAQPMS grids.*

As seen in Section 6 of the manuscript, both the proposed HDCM and STAR require triangulated meshes, and an equal area partition method is applied for HDCM. As an illustration, we present a triangulated mesh with triangle vertices of $m = 10{,}103$ and subregions of $R = 9$ in Figure S4.

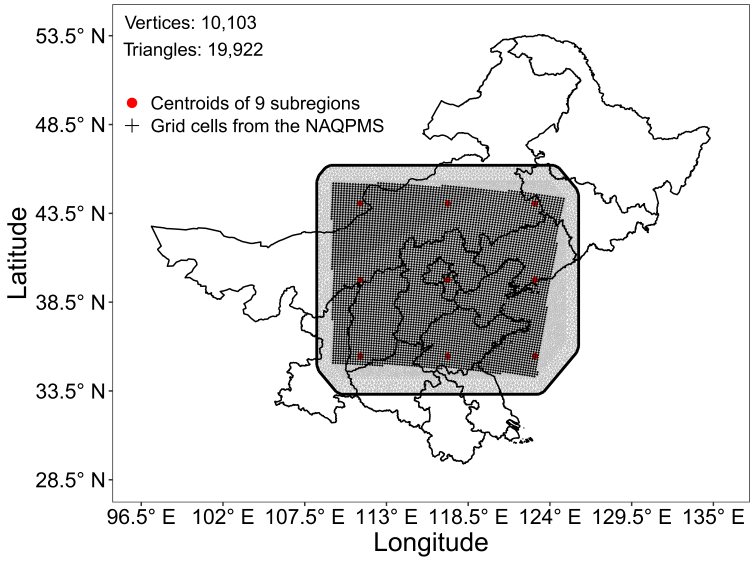

FIG S4. *Triangulated mesh for the reanalysis dataset with 9 subregions, whose centroids are marked with red dots.*

S2. Competing models. In this section, we describe how several competing models are implemented.

S2.1. *Spatially-varying coefficients (SVC) model.* The following version of SVC is used for the model comparison:

$$y_t(s) = \mu_t(s) + w_t(s) + \varepsilon_t(s). \tag{S1}$$

where the zero-mean spatial process $w_t(s)$ and the zero-mean white nose $\varepsilon_t(\boldsymbol{s})$ are mutually and serially independent.

This SVC model differs from the proposed HDCM model in the mean term $\mu_t(s)$ and the random effects term $w_t(s)$. For $\mu_t(s)$, we employ the weighted CMAQ $PM_{2.5}$ output as a single covariate to improve the performance of SVC model, i.e., $\mu_t(s) = \beta_0 + \beta_1 x_{t,1}(C_{\boldsymbol{s}})$ where $x_{t,1}(C_{\boldsymbol{s}})$ represents the square root of the weighted $PM_{2.5}$ output of CMAQ. This is similar to Berrocal et al. (2020) where SVC uses CMAQ output as a single covariate. Using SVC with the single covariate for the BTH data in the winter of 2015 improves the performance by 11.80% in average RMSE and 11.34% in average CRPS based on the leave-one-city-out cross-validation (LOCOCV) of Section 4 of the manuscript.

In addition, at each time t, SVC (S1) models $w_t(s)$ as spatial stationary Gaussian random processes with mean **0** and an exponential covariance function, $\text{Cov}\big(w_t(s_i), w_t(s_j)\big) = \tau_w^2 \exp\big(-\|s_i - s_j\|/\phi_w\big)$. The unknown parameters including τ_w^2, ϕ_w, and others appearing in (S1) can be estimated using the Markov chain Monte Carlo (MCMC) algorithm. We use the `spBayes` package (`https://cran.r-project.org/web/packages/spBayes/`) to fit SVC model.

S2.2. *Universal kriging (UK).* UK follows closely the model specification of SVC, where the prediction performance of UK is also improved by using the square root of the weighted $PM_{2.5}$ output of CMAQ as the single covariable. The difference is that parameters

other than the covariate coefficient vector $\boldsymbol{\beta} = \big(\beta_0, \beta_1\big)$ in the term $\mu_t(s) = \beta_0 + \beta_1 x_{t,1}(C_s)$ are first estimated using the restricted maximum likelihood method (details can be referred to the `geoR` package documentation https://cran.r-project.org/web/packages/geoR/). We then estimate $\boldsymbol{\beta}$ and implement UK interpolation using the `gstat` package (https://cran.r-project.org/web/packages/gstat/).

S2.3. *Spatiotemporally-varying coefficient (STVC) model.* Following Section 3.2 of Berrocal, Gelfand and Holland (2010), the spatial SVC (S1) can be extended to a spatio-temporal downscaler model by allowing spatiotemporally varying coefficients (STVC), where both the intercept and the slope of CMAQ variable vary with respect to time. STVC is given by

$$\text{(S2)} \qquad \begin{cases} y_t(\boldsymbol{s}) = \beta_0 + \beta_{0,t} + (\beta_1 + \beta_{1,t}) x_{t,1}(C_s) + \sum_{k=2}^{4} \beta_k x_{t,k}(C_s) + w_t(s) + \varepsilon_t(s) \\ \beta_{l,t} = \rho_l \beta_{l,t-1} + \eta_{l,t}, \text{ for } l = 0, 1, \end{cases}$$

where $x_{t,1}(C_{\boldsymbol{s}}), \cdots, x_{t,4}(C_{\boldsymbol{s}})$ represent the square root of the weighted average CMAQ $PM_{2.5}$ output, the weighted average surface temperature, the weighted average surface pressure, the weighted average northern and eastern cumulative wind powers at location $\boldsymbol{s}$ and time point t, respectively. The random effects $\beta_{l,t}$ and the spatial process $w_t(s)$ are assumed to be mutually and serially independent, the same for $\beta_{l,t}$ and $\varepsilon_t(s)$, for $w_t(s)$ and $\varepsilon_t(s)$, for $\beta_{l,t-1}$ and $\eta_{l,t}$, and for $\eta_{l,t}$ and $\varepsilon_t(s)$.

The definition of $w_t(s)$ is the same to SVC (S1), and $\boldsymbol{\eta}_{l,t} = \big(\eta_{l,1}, \cdots, \eta_{l,N_t}\big)^T \sim \mathcal{N}\big(0, \delta_l^2 \boldsymbol{I}_{N_t}\big)$. Especially, the performance of (S2) for the BTH data can be improved by assuming $\beta_{1,t} = 0$ based on LOCOCV, and thus the setting of $\beta_{1,t} = 0$ is used in STVC. STVC is estimated based on MCMC algorithms which are available through the `R` package `spTDyn` (Bakar, Kokic and Jin, 2016).

S2.4. *First-order spatiotemporal autoregression (STAR) model.* We follow Section 7.2 of Blangiardo and Cameletti (2015) and use STAR model of the form:

$$\text{(S3)} \qquad \begin{cases} y(\boldsymbol{s}) = \beta_0 + \sum_{k=1}^{4} \beta_k x_{t,k}(C_s) + w_t(\boldsymbol{s}) + \varepsilon_t(\boldsymbol{s}) \\ w_t(\boldsymbol{s}) = \rho w_{t-1}(\boldsymbol{s}) + \eta_t(\boldsymbol{s}), \end{cases}$$

where the zero-mean spatio-temporal process $w_t(s)$ and the zero-mean white nose $\varepsilon_t(\boldsymbol{s})$ are mutually and serially independent, the same for $w_{t-1}(s)$ and $\eta_t(\boldsymbol{s})$, and for $\eta_t(\boldsymbol{s})$ and $\varepsilon_t(\boldsymbol{s})$.

The initial Gaussian process $w_0(s) = \big(w_t(s_1), \cdots, w_t(s_n)\big)^T \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, and the Gaussian random process $\boldsymbol{\eta}_t = \big(\eta_t(s_1), \cdots, \eta_t(s_n)\big)^T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\eta)$. In this work, STAR is implemented through the `R` package `INLA` (Lindgren and Rue, 2015), where a Gaussian process is linked to a Gaussian Markov random field (GMRF, Rue and Held (2005)) on a triangulated mesh via a stochastic partial differential equation (Lindgren, Rue and Lindström, 2011), and the parameters are then estimated by the integrated nested Laplace approximation (INLA, Rue, Martino and Chopin (2009)). This a powerful Bayesian approach for handling large spatio-temporal datasets (Rue et al., 2017).

S2.5. *Random forest (RF).* For the random forest model, we estimate an unknown function $f(\cdot)$ by minimizing the out-of-bag error:

$$\sum_{t=1}^{T} \sum_{s=1}^{n} \text{Loss}\Big\{ y_t(s), f\big(x_{t,1}(C_{\boldsymbol{s}}), \cdots, x_{t,4}(C_{\boldsymbol{s}})\big) \Big\}.$$

The random forest recursively partitions the covariate space to build the trees and uses the weighted average of the predictions across all trees as the final prediction. In this work, we use 500 trees to build the predictive model that is implemented through the `rfinterval` package (https://github.com/haozhestat/rfinterval).

S3. HDCM.

S3.1. *Mapping matrix for HDCM and precision matrices of IDE models.* In Lindgren, Rue and Lindström (2011), the construction of the matrix $\boldsymbol{G}$ depends on a piecewise linear basis $\{\psi_k(s)\}_{k=1}^m$ on the mesh, and $G_{k,k'} = \langle \nabla\psi_k \nabla\psi_{k'} \rangle$, where $G_{k,k'}$ is the (k,k')th element of $\boldsymbol{G}$, $\langle\cdot\rangle$ is defined in Equation (7) of Lindgren, Rue and Lindström (2011), and the vertex index $k, k' = 1, \cdots, m$. In our work, $h(i,k)$ plays an equivalent role to $\psi_k(\boldsymbol{s}_i)$ in their method, where $h(i,k)$ is the (i,k)th element of $\boldsymbol{H}$ and is associated with the location $\boldsymbol{s}_i$ and the vertex B_k. We did not construct our matrix $\boldsymbol{G}$ as a function of the triangle geometry using the definition of $G_{k,k'}$ described earlier.

This is because the piecewise linear assumption of the bases is not suitable for our data based on additional numerical experiments, which may be related to the distribution of data. Especially, because the BTH data are sparsely observed in space and clustered in urban areas, it would be difficult to identify $\boldsymbol{v}_t$ located in rural areas with no stations using piecewise linear bases, and the cross-validation also shows that the piece-wise linear basis for the BTH data does not work well.

Moreover, we require the matrix $\boldsymbol{G}$ to be sparse. Under the setting of Lindgren, Rue and Lindström (2011) (i.e., $G_{k,k'} = \langle \nabla\psi_k \nabla\psi_{k'} \rangle$), it is usually difficult to ensure the sparsity of $\boldsymbol{G}$ when the piecewise linear basis ψ is replaced by other functions such as the Wendland function used in this work. Therefore, in this work, we employ a Laplacian matrix to define $\boldsymbol{G}$ and specify the corresponding precision matrix as a function of $\boldsymbol{G}$ as seen in Section 3.2.2 of the manuscript; a similar definition for the precision matrix of Gaussian random field is also adopted in Bolin, Wallin and Lindgren (2019).

S3.2. *Prior distributions for HDCM.* The prior distributions used in this work are provided in Table S2.

TABLE S2
Prior distributions of parameters in $\boldsymbol{\Theta}$.

Parameter	Prior distribution
$\boldsymbol{\beta}$	$\mathcal{N}(\boldsymbol{0}, 10^5 \boldsymbol{I}_5)$
σ^2	IG$(2,1)$
$\tau^{2,(r)}$	Gamma$(2,1)$
$\tau_0^{2,(r)}$	Gamma$(2,1)$
$\zeta^{2,(r)}$	Uniform$(0,400)$
$\zeta_0^{2,(r)}$	Uniform$(0,400)$
$\theta_1^{(r)}$	$\mathcal{N}(0.0001, 10^5)$
$\theta_2^{(r)}$	Uniform$(0.001,1)$

For all $r = 1, \cdots, R$.

S4. Selection of tuning parameters. Our approach needs to select 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) $c_t \in \{1,2,\cdots,N_t\}$ 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 LOCOCV of Section 4 of the manuscript.

S4.1. *Cross-validation for tuning parameters.* In LOCOCV, considering J combinations for given c_h, c_s, and c_t, an average RMSE across all 13 cities can be calculated with the predictions for each of the combinations, denoted by avgRMSE$_j$, where $j = 1, 2, \cdots, J_0$. Assuming $J_0 = \min_j\{\text{avgRMSE}_j\}$, then the J_0th combination of tuning parameters is specified as the choice of these three parameters.

As seen in Kirchgessner, Nerger and Bunse-Gerstner (2014), the tuning parameters usually depend on the distributions of observations, and the parameters themselves could also be interrelated.

In the setting of HDCM, c_h and c_t, in general, depend on the spatial range and temporal range, respectively, which can be utilized to facilitate the grid search in the cross-validation. The following strategies are adopted in LOCOCV:

a. For $c_h \in (0, 1]$, the length of this interval can be further reduced to improve the grid search in cross-validation by covariograms (or variograms). Based on covariograms, for a given c_t, the interval $(0, 1]$ can be reduced to $\left[c_h^{\min}, c_h^{\max}\right]$ by setting $c_h^{\min} d_{\max}^H = \phi_s^{(1)}$ and $c_h^{\max} d_{\max}^H = \phi_s^{(1)}$, where $d_{\max}^H$ is defined in Section 3.2.1 of the manuscript, $\phi_s^{(1)}$ and $\phi_s^{(2)}$ correspond to the distances where the spatial correlation coefficients are equal to 0.9 and 0.05, respectively. Because $0 < \phi_s^{(1)} \leq d_{\max}^H$ and $0 < \phi_s^{(2)} \leq d_{\max}^H$, the length of $\left[c_h^{\min}, c_h^{\max}\right] = \left[\phi_s^{(1)}/d_{\max}^H, \phi_s^{(2)}/d_{\max}^H\right]$ is smaller than that of $(0, 1]$. In practice, although these two values of $\phi_s^{(1)}$ and $\phi_s^{(2)}$ are usually not directly obtained from covariograms, we can fit a parametric spatio-temporal covariance model to an empirical covariogram such as Figure 2 of the manuscript and then estimate them using the fitted parametric covariance model. In the data analysis, we used $c_h \in \left[0.1, 0.31\right]$ for the BTH data and $c_h \in \left(0.03, 0.17\right]$ for the reanalysis data.

b. For c_t, the empirical covariograms can be also used to narrow down the grid search in LOCOCV. For example, Figure 2 shows the temporal range is likely to be less than 5 days, and we can select a c_t value from $\{1, 2, \cdots, 5\}$ other than $\{1, 2, \cdots, N_t\}$, where N_t is the length of time points.

c. For the covariance tapering parameter in space c_s, Kirchgessner, Nerger and Bunse-Gerstner (2014) showed $c_s = 8\sqrt{\frac{N_e}{40}}dx$ in dense observations from regular grids, where dx represents the grid spacing. Although the formula does not provide an optimal value of c_s for the triangle mesh in the proposed HDCM, we can use this definition to construct an interval such that $c_s \in \left[8\sqrt{\frac{N_e}{40}}dx_{(1)}/d_{\max}^B, 8\sqrt{\frac{N_e}{40}}dx_{(2)}/d_{\max}^B\right]$, where $d_{\max}^B = \max_{l,l'}\left\{\|B_l^{(r)} - B_{l'}^{(r)}\|\right\}$. The length of this interval is smaller than that of $(0, 1]$. Based on numerical experiments, the predictive performance of HDCM is rather competitive when $dx_{(1)}$ is set as 2.5% quantile of $\left\{\|B_l^{(r)} - B_{l'}^{(r)}\|\right\}_{r,r'=1}^m$, and $dx_{(1)}$ as 95% quantile of $\left\{\|B_l^{(r)} - B_{l'}^{(r)}\|\right\}_{l,l'=1}^m$.

S4.2. *Sensitivity analysis with respect to tuning parameters.* In this section, we investigate the proposed HDCM with respect to the sensitivity of the selection of several tuning parameters. The sensitivity analysis is carried out based on the cross-validation procedure described above (i.e., LOCOCV). Because the sensitivity analysis results for all 13 cities are similar, here we choose Zhangjiakou as an illustration. Specifically, we consider all monitoring data from Zhangjiakou in the winter as the test set and all data from the remaining twelve cities in the winter as the training set. We evaluate the predictive performance of HDCM in terms of RMSE with respect to various tuning parameters.

Although the tuning parameter c_h can be selected from the interval $(0, 1]$, we found that the performance of model is better when $c_h \in (0, 0.5]$ than when $c_h \in (0.5, 1]$. Therefore, we consider $c_h = \{0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5\}$ in the sensitivity analysis. We also set the two covariance tapering parameters to $c_s = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$ and $c_t = \{1, 2, 3, 4, 5\}$. The size of ensemble members $N_e = \{30, 50, 70, 100\}$.

RMSEs for various combinations of the tuning parameters for the training data and test data are presented in Figure S5 and Figure S6, respectively. Below are the results:

(a) For c_h, RMSE increases as c_h gets closer to 0.5 while the predictions are robust when $c_h \in (0.1, 0.3]$.
(b) For c_s, when $c_h < 0.3$, the predictions with $c_s < 0.5$ in the test dataset are more accurate than those with $c_s > 0.5$ while the predictions in the training dataset are robust for all c_s.
(c) For c_t, the predictions in the training dataset are usually robust in almost all cases while the prediction performance in the test dataset is better for a smaller c_t. This suggests $c_t = 1$ is an optimal choice.
(d) For N_e, the predictions in almost all cases tend to be more accurate for a larger N_e.

In addition, the computation becomes more expensive when c_s is closer to 1 and/or c_t is closer to 5. More specifically, the computation time when $c_s = 1$ and $c_t = 5$ is nearly 3 times of that when $c_s = 0.1$ and $c_t = 1$, while these two parameters are increased by a factor of 10 and 5, respectively; see Figure S7. In summary, under the current setting, the computational cost of the proposed approach is almost linear with respect to these two parameters.

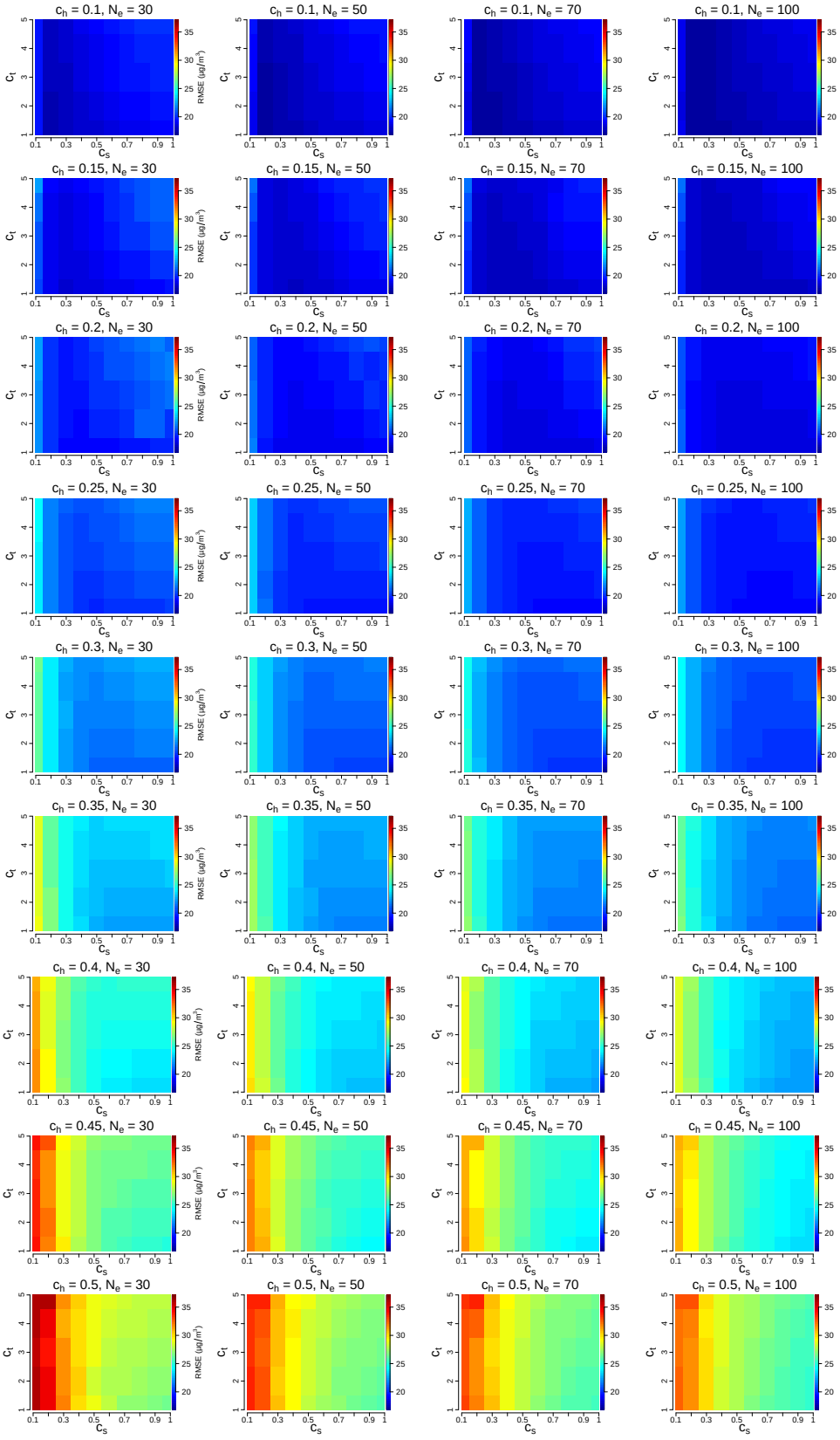

FIG S5. *Root mean squared error (RMSE) of the training dataset for various tuning parameters.*

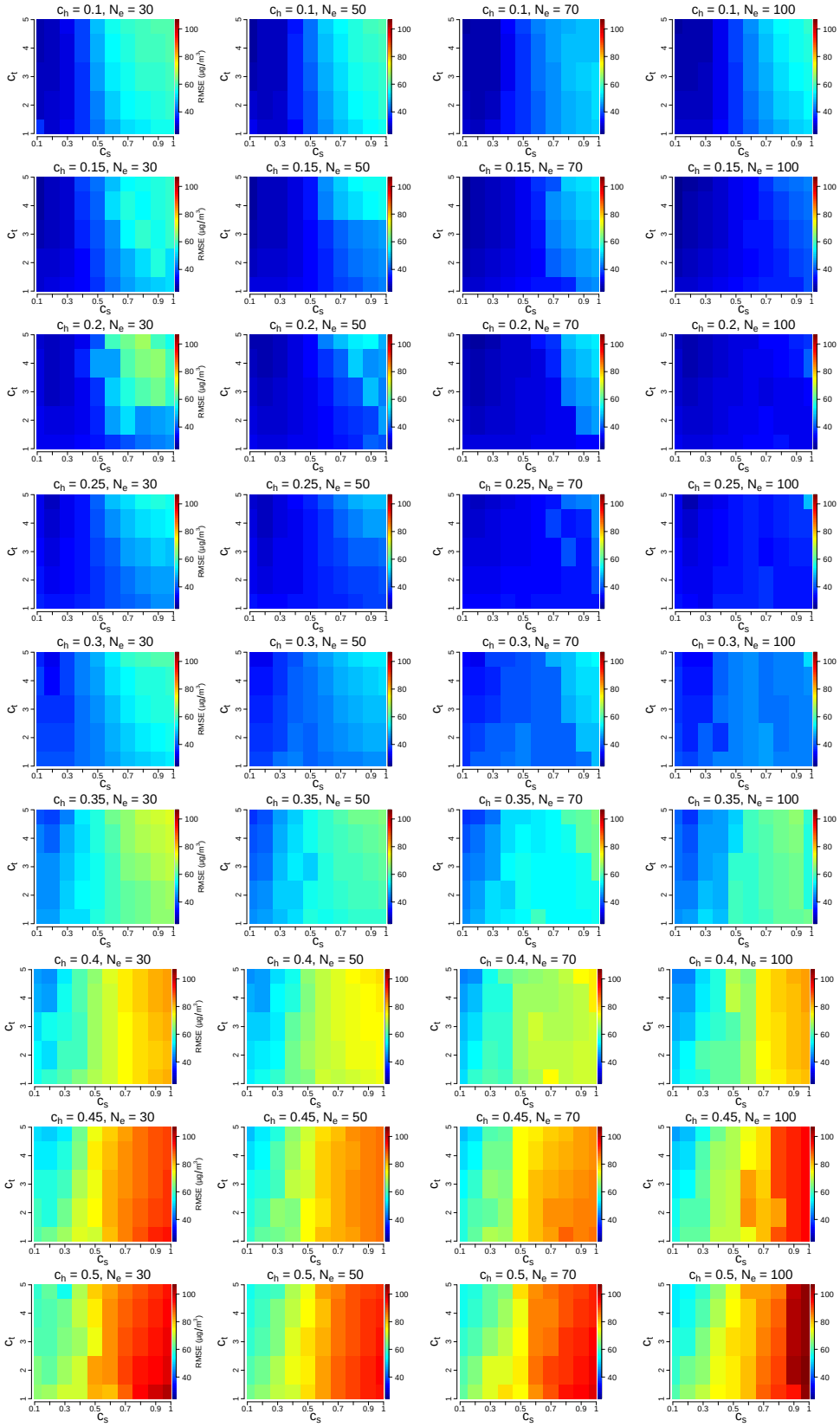

FIG S6. *Root mean squared error (RMSE) of the test dataset for various tuning parameters.*

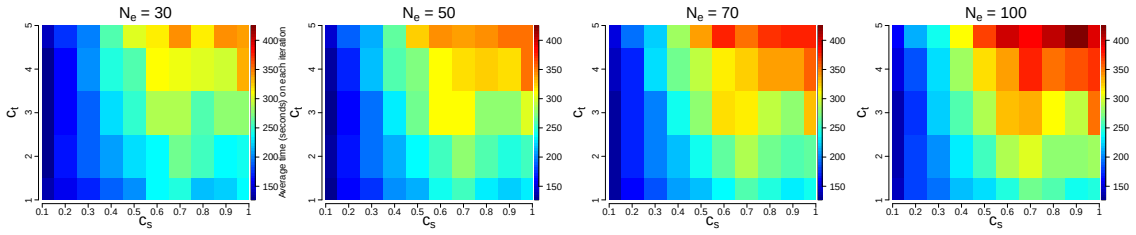

FIG S7. *Average running time (seconds) for each iteration.*

S5. Comparing VB-spEnKS with MCMC-KS. VB and spEnKS (VB-spEnKS) is the approximation of MCMC and KS (MCMC-KS). We first apply both VB-spEnKS and MCMC-KS to HDCM on the simulated data, and then compare their performances.

S5.1. *Simulation design.* We set the spatial domain $\mathcal{D}'$ to be the BTH region with 2,499 equally spaced grids, which are the regular CMAQ grids as shown in Figure 1 (b) of the manuscript and denoted as $\{C_{i''}\}_{i''=1}^{2,499}$. The time domain $\mathcal{T}'$ is set to be $[0,1]$ with 20 equally spaced time points. For any $\boldsymbol{s} \in \mathcal{D}'$ at time point $t \in \mathcal{T}'$, we use the following model to generate data,

$$y_t(\boldsymbol{s}) = \beta_0 + \beta_1 x_{1,t}(\boldsymbol{s}) + \beta_2 x_{2,t}(\boldsymbol{s}) + w_t(\boldsymbol{s}) + \varepsilon_t(\boldsymbol{s}), \tag{S4}$$

where the covariate $x_{1,t}(\boldsymbol{s})$ is generated from the standard normal distribution, $\boldsymbol{x}_{2,t}(\boldsymbol{s})$ is from a spatial Gaussian random field with mean $\mathbf{0}$ and an isotropic exponential covariance kernel, i.e., for location $\boldsymbol{s}$ and $\boldsymbol{s}'$, $\mathrm{Cov}\big(x_{2,t}(\boldsymbol{s}), x_{2,t}(\boldsymbol{s}')\big) = \exp\left(-10\|\boldsymbol{s}-\boldsymbol{s}'\|/d^{\boldsymbol{s}}_{\max}\right)$ with $d^{\boldsymbol{s}}_{\max} = \max_{\boldsymbol{s},\boldsymbol{s}'}\left\{\|\boldsymbol{s}-\boldsymbol{s}'\|\right\}$, the spatio-temporal random process $\boldsymbol{w}_t(\boldsymbol{s})$ is mutually and serially independent of the error term $\boldsymbol{\varepsilon}_t(\boldsymbol{s})$, and the error term $\varepsilon_t(\boldsymbol{s}) \sim N(0,\sigma^2)$. The random process $\boldsymbol{w}_t(\boldsymbol{s})$ follows a zero-mean stationary space-time Gaussian process with a kernel specified in the Gneiting class (Gneiting, 2002), i.e., for any $(\boldsymbol{s},t)$ and $(\boldsymbol{s}',t')$, the covariance between $\boldsymbol{w}_t(\boldsymbol{s})$ and $\boldsymbol{w}_{t'}(\boldsymbol{s}')$ is given by

$$\mathrm{cov}\big(\boldsymbol{w}_t(\boldsymbol{s}), \boldsymbol{w}_{t'}(\boldsymbol{s}')\big) = \sigma_w^2\big(1+\Psi_t(\tau)\big)^{-\delta/2}\Psi_s\Big(u/\sqrt{\Psi_t(\tau)}\Big), \tag{S5}$$

where $\tau = |t-t'|, \sigma_w > 0, \delta \geq 2, \Psi_t(\tau) = \big(|\tau/\phi_t|^a + 1\big)^{-b/a}$ for $\phi_t > 0, a \in (0,2]$ and $b > 0$, $u = \|\boldsymbol{s}-\boldsymbol{s}'\|, \Psi_s(u) = 2^{1-\nu}\Gamma^{-1}(\nu)\big(\sqrt{2\nu}u/\phi_s\big)^{\nu}K_\nu\big(\sqrt{2\nu}u/\phi_s\big)$ for $r \geq 0, \phi_s > 0, \nu > 0$, $\Gamma(\cdot)$ is the Gamma function, and K_ν is the modified Bessel function of the second kind.

We set model parameters of (S4) to $\boldsymbol{\beta} = (15,1,1)$, $\sigma_w^2 = 1$, $\delta = 2, a = 1, b = 1.5$, $\phi_t = 0.2$ for Ψ_t, and $\nu = 3$, $\phi_s = 100$ for Ψ_s.

The simulation consists of the following steps:

Step 1: Sample $w_t(\boldsymbol{s})$ using (S5), where $\boldsymbol{s} \in \mathcal{S} = \{C_1, \cdots, C_{2499}\}$, and $t = 0, 1/19, \cdots, 1$.
Step 2: Generate synthetic data $y_t(s)$ using (S4).
Step 3: Sample n spatial points from $\mathcal{S}$ and denote as $\mathcal{S}^{(1)} = \{\boldsymbol{s}'_1, \cdots, \boldsymbol{s}'_n\}$. Let $\mathcal{S}^{(2)} = \mathcal{S}\backslash\mathcal{S}^{(1)}$. The training set is from $\mathcal{S}^{(1)}$ while the test set is from $\mathcal{S}^{(2)}$.
Step 4: Fit HDCM using the training set through VB-spEnKS and MCMC-KS algorithms, and generate predictions of the test data $y_t(s)$ in $\mathcal{S}^{(2)}$, denoted by $\widehat{y}_t(s)$.
Step 5: Repeat Steps $2-4$ to assess the estimates of the parameters (e.g., $\boldsymbol{\beta}$), the predictions $\widehat{y}_t(s)$, and the recovered surface for the random process $w_t(s)$.

In step 4, we set four tuning parameters to $c_h = 0.23, c_s = 0.3, c_t = 1$, and $N_e = 100$. In step 5, the parameter estimates are evaluated using standard deviation (SD) and mean squared error (MSE). Under each setting, we repeat the procedure 40 times and report the results in Table S3.

S5.2. *Simulation results.* Table S3 shows the parameter estimates and prediction performance of VB-spEnKS are very similar to that of MCMC-KS. VB-spEnKS in some cases produces even smaller MSEs than MCMC-KS, indicating that VB-spEnKS can approximate MCMC-KS very well. Moreover, VB-spEnKS has higher computational efficiency than MCMC-KS as it took MCMC-KS more than two weeks to complete the 40 independent simulations but less than 5 hours for the same task on the same workstation. On average, VB-spEnKS converges after 15 iterations. In Table S3, VB-spEnKS results are the summaries after about 15 iterations and MCMC-KS results are from $2{,}500$ samples after a burn-in of $2{,}500$. All these results demonstrate that VB-spEnKS can provide a competitive performance within dozens of iterations.

TABLE S3

The estimates, standard deviations (SD), and mean squared errors (MSEs) of the regression coefficients $\boldsymbol{\beta}$ and the nugget effect σ^2 when the training spatial sample size $n = 100$. The results are based on 40 independent simulations and obtained by fitting HDCM model with two algorithms: (1) Markov chain Monte Carlo and Kalman smoother (MCMC-KS), (2) Variational Bayes and spatial-partitioning-based Ensemble Kalman smoother (VB-spEnKS) with a subregion.

n	Algorithm	Criteria	$\beta_0(=15)$	$\beta_1(=1)$	$\beta_2(=1)$	$\sigma^2(=0.1)$	Prediction
100	MCMC-KS	Estimate	14.8558	0.9994	0.9997	0.1235	-
		SD	0.0946	0.0095	0.0166	0.0055	-
		MSE	0.0297	0.0001	0.0003	0.0006	0.1678
	VB-spEnKS	Estimate	14.9769	0.9997	0.9998	0.1452	-
		SD	0.0726	0.0094	0.0169	0.0061	-
		MSE	0.0059	0.0001	0.0003	0.0021	0.1866

In addition, our proposed model (2) for $w_t(\boldsymbol{s})$ in the manuscript successfully captures most of the spatio-temporal patterns in the true spatio-temporal process. As an illustration, Figure S8 demonstrates the excellent performance of HDCM in recovering the true $w_t(s)$ from a simulated diffusion process for $n = 100$.

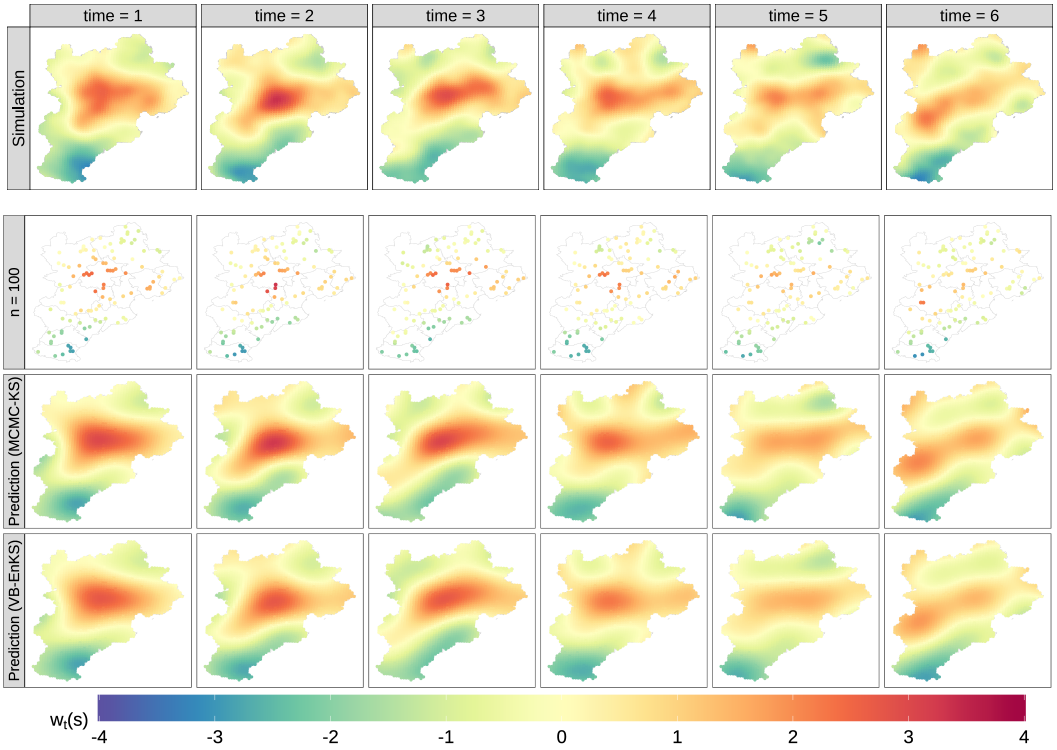

FIG S8. Maps of the random process $w_t(s)$ in the Gneiting space-time covariance model. From top to bottom: simulated $w_t(s)$ map, map of 100 sampling locations, predicted $w_t(s)$ map using MCMC-KS based on the 100 locations, and predicted $w_t(s)$ map using VB-spEnKS based on the 100 locations. The time horizon is from time $= 15$ to time $= 20$.

S6. Cross validation.

S6.1. *Model comparisons for the summer data.* In this section, we consider the data with the square root scale in the summer of 2015, and the results of LOCOCV are reported in Table S4 and Table S5.

TABLE S4

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 spatio-temporal 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 S8. Daily data from June 1, 2015, and August 31, 2015, are considered.

City	RMSE						
	CMAQ	UK	RF	SVC	STVC	STAR	HDCM
Baoding	36.08	24.92	25.41	**22.17**	23.96	22.96	22.74
Beijing	36.87	22.00	27.17	**21.57**	24.19	22.77	22.50
Cangzhou	34.41	20.92	20.32	17.70	18.42	**16.63**	19.07
Chengde	25.04	18.74	19.53	**16.68**	18.12	17.59	18.82
Handan	34.33	27.21	23.33	19.12	26.79	20.12	**19.10**
Hengshui	44.14	30.53	25.37	21.18	24.69	**20.47**	21.19
Langfang	27.52	15.48	17.56	**13.28**	16.08	13.61	13.78
Qinhuangdao	24.16	28.11	23.47	24.44	23.20	23.81	**20.93**
Shijiazhuang	34.23	27.57	28.43	25.90	28.44	**25.24**	26.23
Tangshan	43.65	36.51	25.39	27.61	25.50	**24.42**	25.65
Tianjin	24.22	15.04	17.50	**14.01**	18.68	14.30	14.69
Xingtai	40.02	28.73	26.48	**20.77**	28.20	21.87	21.09
Zhangjiakou	23.98	23.81	19.35	27.39	17.53	23.57	**16.37**
Average	32.97	24.58	23.02	20.91	22.60	20.57	**20.17**

City	CRPS						
	CMAQ	UK	RF	SVC	STVC	STAR	HDCM
Baoding	21.27	13.19	13.43	**11.97**	12.52	12.06	12.02
Beijing	20.51	11.87	14.48	**11.76**	12.93	12.24	12.10
Cangzhou	23.13	11.49	11.26	9.78	10.67	**9.44**	10.44
Chengde	12.56	10.41	10.10	9.51	**9.37**	9.60	9.61
Handan	20.29	15.37	12.99	11.34	15.01	11.39	**11.00**
Hengshui	30.35	17.85	13.90	11.80	13.74	**11.27**	11.79
Langfang	15.94	9.35	10.32	9.29	9.53	**8.91**	9.61
Qinhuangdao	12.77	15.66	12.82	13.39	12.62	13.02	**11.46**
Shijiazhuang	19.27	14.94	15.69	14.46	15.50	**13.67**	14.38
Tangshan	25.75	20.75	14.43	15.61	14.64	**13.98**	14.72
Tianjin	13.75	8.81	9.99	**8.55**	10.70	8.61	8.84
Xingtai	23.55	16.09	14.53	**11.62**	15.56	12.01	11.96
Zhangjiakou	13.32	13.28	10.61	15.19	9.74	13.19	**9.06**
Average	19.42	13.77	12.66	11.86	12.50	11.49	**11.30**

TABLE S5

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 eight 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 spatio-temporal 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 June 1, 2015, and August 31, 2015, are considered.

City	MAE						
	CMAQ	UK	RF	SVC	STVC	STAR	HDCM
Baoding	27.76	17.86	18.11	**15.77**	16.71	16.28	16.37
Beijing	26.37	15.38	19.49	**14.84**	17.44	16.14	15.19
Cangzhou	28.16	15.59	15.615	13.10	14.26	**12.53**	14.22
Chengde	15.83	13.71	13.23	**11.74**	12.06	12.56	12.56
Handan	26.71	21.76	18.02	15.01	20.76	15.60	**14.95**
Hengshui	36.56	24.90	19.20	16.30	19.27	**15.20**	16.07
Langfang	21.45	12.39	14.23	**9.90**	13.21	10.21	10.05
Qinhuangdao	17.18	21.63	17.77	18.01	17.26	17.88	**15.58**
Shijiazhuang	25.85	20.44	22.12	20.53	21.49	**19.03**	19.94
Tangshan	35.01	29.71	20.55	22.10	21.09	**20.01**	20.99
Tianjin	18.29	11.38	13.90	**10.69**	14.51	10.89	11.26
Xingtai	31.02	22.43	19.88	**15.52**	21.56	16.28	16.05
Zhangjiakou	18.23	18.42	14.41	20.48	13.29	18.25	**12.62**
Average	25.26	18.89	17.43	15.69	17.15	15.45	**15.07**

City	FAC2						
	CMAQ	UK	RF	SVC	STVC	STAR	HDCM
Baoding	0.714	0.937	0.909	**0.960**	0.951	**0.960**	0.951
Beijing	0.644	0.955	0.824	0.953	0.918	**0.956**	0.944
Cangzhou	0.500	0.931	0.917	**0.967**	0.960	**0.967**	0.938
Chengde	0.630	0.829	0.851	**0.891**	0.777	0.832	0.861
Handan	0.823	0.889	0.957	0.981	0.910	0.984	**0.986**
Hengshui	0.536	0.862	0.953	0.964	0.924	**0.967**	0.949
Langfang	0.788	0.959	0.916	**0.992**	0.957	**0.992**	**0.992**
Qinhuangdao	0.674	0.624	0.694	0.715	0.720	0.726	**0.744**
Shijiazhuang	0.788	0.897	0.868	0.911	0.866	**0.918**	0.893
Tangshan	0.726	0.736	0.846	0.781	0.802	**0.879**	0.851
Tianjin	0.827	0.970	0.938	**0.981**	0.940	0.980	0.974
Xingtai	0.739	0.924	0.957	0.995	0.929	**0.997**	0.984
Zhangjiakou	0.636	0.679	0.802	0.674	0.750	0.677	**0.815**
Average	0.694	0.861	0.879	0.905	0.877	0.910	**0.914**

S6.2. *The performance of the proposed HDCM with spatial partitioning.* In this section, to illustrate that spatial partitioning works for the HDCM, we performed a cross-validation for the BTH data in both seasons of 2015 by considering the space-partitioning-based HDCM with two subregions ($HDCM_2$).

TABLE S6

Averaged RMSE (Root Mean Squared Error), CRPS (Continuous Rank Probability Score), 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 the the space-partitioning-based HDCM with two subregions ($HDCM_2$). Daily data from the winter and summer of 2015 are considered.

City	Winter of 2015				Summer of 2015			
	RMSE	CRPS	MAE	FAC2	RMSE	CRPS	MAE	FAC2
Baoding	70.03	37.99	48.44	0.90	22.58	11.92	16.34	0.95
Beijing	58.71	30.84	31.78	0.89	22.72	12.14	15.27	0.94
Cangzhou	41.05	22.34	26.27	0.95	18.34	10.08	13.85	0.94
Chengde	29.58	15.37	19.86	0.79	19.58	9.88	12.95	0.85
Handan	44.33	29.11	30.93	0.94	19.23	11.08	15.01	0.99
Hengshui	49.14	31.65	33.31	0.99	20.44	11.50	15.78	0.96
Langfang	37.67	29.03	23.61	0.99	13.83	9.63	9.96	0.99
Qinhuangdao	27.92	14.72	19.60	0.77	21.60	12.09	16.74	0.73
Shijiazhuang	53.89	33.30	38.77	0.94	26.16	14.44	20.28	0.90
Tangshan	38.28	22.48	27.03	0.95	25.88	14.94	21.28	0.87
Tianjin	36.70	26.41	23.81	0.96	14.84	8.94	11.35	0.97
Xingtai	47.46	32.48	32.46	0.96	21.36	12.01	16.01	0.98
Zhangjiakou	27.64	14.69	19.32	0.71	17.84	9.94	14.15	0.79
Average	43.26	26.19	28.86	0.90	20.34	11.43	15.31	0.91

S7. A modified Diebold-Mariano test for predictive accuracy among methods. To determine whether the difference between any two models is significant in terms of predictive accuracy, we carry out a modified Diebold-Mariano (DM) test (Harvey, Leybourne and Newbold, 1997).

Let $\{r_t^{(1)}(\boldsymbol{s})\}_{t=1}^{N_t}$ and $\{r_t^{(2)}(\boldsymbol{s})\}_{t=1}^{N_t}$ be the residual series at location $\boldsymbol{s}$ obtained from Model 1 and Model 2 respectively. More specifically, $r_t^{(1)}(\boldsymbol{s}) = \widehat{y}_t^{(1)}(\boldsymbol{s}) - y_t(\boldsymbol{s})$ and $r_t^{(2)}(\boldsymbol{s}) = \widehat{y}_t^{(2)}(\boldsymbol{s}) - y_t(\boldsymbol{s})$. Denote the loss-differential as $d_{\boldsymbol{s},t} = g\big(r_t^{(1)}(\boldsymbol{s})\big) - g\big(r_t^{(2)}(\boldsymbol{s})\big)$. Let $\mu = E(d_{\boldsymbol{s},t})$ for a specific function $g(\cdot)$, where $g(e) = e^2$ are used for this work. Under the null hypothesis $\mu = 0$ (i.e., no difference in the accuracy of two sets of predictions), the DM test statistic (Diebold and Mariano, 1995) is given by

$$\text{DM} = \frac{\bar{d}_s}{\sigma_{d_s}} \to \mathcal{N}(0,1), \tag{S6}$$

where $\bar{d}_s = \frac{1}{N_t}\sum_{t=1}^{N_t} d_{\boldsymbol{s},t}$ and σ_{d_s} is the standard deviation of $d_{\boldsymbol{s},t}$.

In practice, σ_{d_s} is replaced by its consistent estimate. More specifically, a consistent estimate of σ_{d_s} can be asymptotically represented by $\sqrt{\{\widehat{\gamma}_s(0) - 2\sum_{\tau=1}^{h-1}\widehat{\gamma}_s(\tau)\}/N_t}$, here $\widehat{\gamma}_s(\tau)$ is obtained by replacing μ with $\bar{d}_s$ into (S7):

$$\gamma_s(\tau) = \frac{1}{N_t}\sum_{t=\tau+1}^{N_t}\big(d_{\boldsymbol{s},t} - \mu\big)\big(d_{\boldsymbol{s},t-\tau} - \mu\big). \tag{S7}$$

where $\tau = 0, 1, 2, \cdots, N_t$. (S7) defines the autocovariance at lag τ.

Because the DM test (S6) tends to be oversized, a modified DM test is proposed by Harvey, Leybourne and Newbold (1997):

$$\text{DM}^* = \sqrt{\frac{N_t + 1 - 2h + h(h-1)/N_t}{N_t}}\text{DM} \to Student(N_t - 1), \tag{S8}$$

where $h = N_t^{1/3} + 1$ is often used. In this work, we used the modified DM test (S8).

Because the dependence of the residuals of predictions is usually much weaker in space than in time, we assume the residual series, $\{r_t^{(1)}(\boldsymbol{s})\}_{t=1}^{N_t}$ and $\{r_t^{(2)}(\boldsymbol{s})\}_{t=1}^{N_t}$, at different sites to be independent of each other. In light of this, the statistic DM* used here is obtained by substituting the DM $= \frac{\sum_{s=1}^{n_0} \bar{d}_s}{\sqrt{\sum_{s=1}^{n_0} \widehat{\sigma}_{d_s}^2}}$ into (S8), where n_0 is the number of sites in the test set. Based on the predictions from LOCOCV, the HLN statistics and their corresponding p-values are reported in Tables S7 and S8. In these tests, we always represent $\{r_t^{(1)}(\boldsymbol{s})\}_{t=1}^{N_t}$ as the residuals of the proposed HDCM, and represent $\{r_t^{(2)}(\boldsymbol{s})\}_{t=1}^{N_t}$ as that of one of the other seven methods.

Occasionally, even when HDCM is dominated by other methods in some cities under certain criteria (e.g., RMSE for the summer of 2015), the difference in the predictive accuracy is mostly not significant at a 5% significance level; see Table S8 of the summer.

TABLE S7

The values of the modified Diebold-Mariano (DM) statistic and their corresponding p-values between the proposed HDCM and each of the other seven methods calculated for 13 cities in the BTH region using the results of leave-one-city-out cross-validation. A DM less than 0 indicates that the proposed HDCM performs better predictions than the method in the corresponding column, and the significant difference with the test level of 0.05 *is represented by adding "*" to the upper right corner of the statistic, and the corresponding p-value is in bold. A DM greater than 0 indicates that the method in the corresponding column performs better predictions than the proposed HDCM and is colored in gray, and the significant difference is represented by adding "†" to the upper right corner of the statistic. Daily data from November* 1, 2015 *to January* 31, 2016 *are considered.*

City	Value of the modified DM statistic							
	CMAQ	UK	RF	SVC	STVC	STAR	HDCM	$HDCM_2$
Baoding	-5.535*	-4.806*	-3.472*	-2.959*	-4.887*	-3.225*	-	5.151†
Beijing	-8.837*	-6.393*	-6.768*	-5.182*	-5.689*	-4.678*	-	-3.515*
Cangzhou	-4.623*	0.813	-0.494	0.982	1.016	1.568	-	-1.628
Chengde	-6.574*	-6.804*	-4.679*	-3.253*	-3.189*	-0.751	-	-3.491*
Handan	-8.199*	-5.736*	-3.262*	-4.060*	-3.405*	-4.985*	-	-0.182
Hengshui	-4.267*	-4.504*	-3.617*	-2.206*	-4.064*	-1.539	-	3.932†
Langfang	-7.963*	-2.973*	-6.706*	1.819	-3.313*	-0.945	-	-1.207
Qinhuangdao	-7.709*	-4.109*	-3.521*	-3.030*	-4.659*	-6.868*	-	-0.681
Shijiazhuang	-8.013*	2.449†	-0.174	3.080†	2.158†	3.485†	-	0.265
Tangshan	-9.542*	-4.135*	-5.992*	1.304	-3.105*	6.118†	-	-5.709*
Tianjin	-8.722*	-2.130*	-3.989*	2.142†	-2.253*	0.892	-	-4.568*
Xingtai	-4.873*	-3.388*	-3.901*	-2.315*	-3.232*	-2.029*	-	1.883
Zhangjiakou	-5.275*	-5.069*	-3.233*	-6.779*	-6.160*	-7.030*	-	2.146†

City	p-value of the modified DM test							
	CMAQ	UK	RF	SVC	STVC	STAR	HDCM	$HDCM_2$
Baoding	**0.000**	**0.000**	**0.001**	**0.004**	**0.000**	**0.002**	-	0.000
Beijing	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	-	**0.001**
Cangzhou	**0.000**	0.418	0.623	0.329	0.312	0.120	-	0.107
Chengde	**0.000**	**0.000**	**0.000**	**0.002**	**0.002**	0.455	-	**0.001**
Handan	**0.000**	**0.000**	**0.002**	**0.000**	**0.001**	**0.000**	-	0.856
Hengshui	**0.000**	**0.000**	**0.000**	**0.030**	**0.000**	0.127	-	0.000
Langfang	**0.000**	**0.004**	**0.000**	0.072	**0.001**	0.347	-	0.231
Qinhuangdao	**0.000**	**0.000**	**0.001**	**0.003**	**0.000**	**0.000**	-	0.498
Shijiazhuang	**0.000**	0.016	0.862	0.003	0.034	0.001	-	0.791
Tangshan	**0.000**	**0.000**	**0.000**	0.196	**0.002**	0.000	-	**0.000**
Tianjin	**0.000**	**0.036**	**0.000**	0.035	**0.027**	0.375	-	**0.000**
Xingtai	**0.000**	**0.001**	**0.000**	**0.023**	**0.002**	**0.045**	-	0.063
Zhangjiakou	**0.000**	**0.000**	**0.002**	**0.000**	**0.000**	**0.000**	-	0.035

The null hypothesis: Two predictions from two different models have the same accuracy.

TABLE S8

The values of the modified Diebold-Mariano (DM) statistic and their corresponding p-values between the proposed HDCM and each of the other seven methods calculated for 13 cities in the BTH region using the results of leave-one-city-out cross-validation. A DM less than 0 indicates that the proposed HDCM performs better predictions than the method in the corresponding column, and the significant difference with the test level of 0.05 is represented by adding "" to the upper right corner of the statistic, and the corresponding p-value is in bold. A DM greater than 0 indicates that the method in the corresponding column performs better predictions than the proposed HDCM and is colored in gray, and the significant difference is represented by adding "†" to the upper right corner of the statistic. Daily data from June 1, 2015, and August 31, 2015, are considered.*

City	Value of the modified DM statistic							
	CMAQ	UK	RF	SVC	STVC	STAR	HDCM	$HDCM_2$
Baoding	-10.868*	-2.847*	-3.425*	1.869	-1.450	-0.375	-	0.547
Beijing	-12.224*	2.005†	-9.271*	3.192†	-5.222*	-1.503	-	-1.497
Cangzhou	-8.760*	-3.143*	-1.685	3.892†	0.587	3.648†	-	2.534†
Chengde	-4.398*	0.062	-1.432	1.355	1.450	1.160	-	-2.687*
Handan	-8.460*	-8.725*	-5.105*	-0.038	-9.187*	-1.628	-	-0.591
Hengshui	-11.840*	-6.984*	-3.810*	0.022	-3.994*	1.457	-	2.765†
Langfang	-8.617*	-2.812*	-5.907*	1.736	-4.460*	0.539	-	-0.353
Qinhuangdao	-4.483*	-5.211*	-4.361*	-3.910*	-3.817*	-2.587*	-	-2.213*
Shijiazhuang	-5.745*	-1.320	-3.457*	0.511	-1.990*	1.159	-	0.391
Tangshan	-5.208*	-6.814*	0.427	-2.970*	0.133	3.890†	-	-1.078
Tianjin	-8.202*	-0.720	-4.889*	1.786	-5.861*	0.667	-	-1.237
Xingtai	-9.456*	-7.077*	-6.951*	1.264	-4.469*	-1.397	-	-1.519
Zhangjiakou	-5.733*	-7.285*	-5.552*	-7.211*	-2.596*	-6.641*	-	-5.156*

City	p-value of the modified DM test							
	CMAQ	UK	RF	SVC	STVC	STAR	HDCM	$HDCM_2$
Baoding	**0.000**	**0.005**	**0.001**	0.065	0.151	0.709	-	0.586
Beijing	**0.000**	0.048	**0.000**	0.002	**0.000**	0.136	-	0.138
Cangzhou	**0.000**	**0.002**	0.095	0.000	0.559	0.000	-	0.013
Chengde	**0.000**	0.951	0.156	0.179	0.150	0.249	-	**0.009**
Handan	**0.000**	**0.000**	**0.000**	0.970	**0.000**	0.107	-	0.556
Hengshui	**0.000**	**0.000**	**0.000**	0.982	**0.000**	0.149	-	0.007
Langfang	**0.000**	**0.006**	**0.000**	0.086	**0.000**	0.592	-	0.725
Qinhuangdao	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.011**	-	**0.029**
Shijiazhuang	**0.000**	0.190	**0.001**	0.610	**0.050**	0.250	-	0.697
Tangshan	**0.000**	**0.000**	0.670	**0.004**	0.894	0.000	-	0.284
Tianjin	**0.000**	0.473	**0.000**	0.077	**0.000**	0.506	-	0.219
Xingtai	**0.000**	**0.000**	**0.000**	0.210	**0.000**	0.166	-	0.132
Zhangjiakou	**0.000**	**0.000**	**0.000**	**0.000**	**0.011**	**0.000**	-	**0.000**

S8. Robustness of the proposed approach to deviations from the normality and linearity assumptions.

S8.1. *Robustness to normality.* Robustness to normality is investigated through a cross validation for the raw data based on the proposed models (i.e., HDCM and $HDCM_2$). For the BTH $PM_{2.5}$ concentration data, the raw observations in the summer of 2015 tend to admit a skewed right distribution where the right tail (larger values) is much longer than the left tail (smaller values), while the distribution of observations on the square root scale from the summer is very close to a normal distribution; see Figure S1(d) and Figure S1(e). Therefore, we perform a cross-validation for the raw observations in the summer to verify whether our method is valid when the normality assumption is violated. LOCOCV results are reported in Table S9.

TABLE S9

Averaged RMSE (Root Mean Squared Error), CRPS (Continuous Rank Probability Score), 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 eight models: (1) CMAQ numerical model output, (2) universal kriging, (3) random forest model, (4) spatially-varying coefficient downscaling model (SVC), (5) spatiotemporally-varying coefficient downscaling model (STVC), (6) first-order spatiotemporal autoregression (STAR), (7) proposed hierarchical dynamic calibration model (HDCM), and (8) spatial-partitioning-based HDCM with two subregions ($HDCM_2$). The smallest RMSE, CRPS, MAE, and FAC2 are in bold and the second smallest ones are underlined. Daily raw data from June 1, 2015, and August 31, 2015, are used for all models.

Criterion	CMAQ	UK	RF	SVC	STVC	STAR	HDCM	$HDCM_2$
RMSE	32.97	24.67	23.09	22.03	22.79	20.46	**20.35**	<u>20.43</u>
CRPS	19.54	13.95	12.57	12.15	12.70	11.33	<u>11.31</u>	**11.25**
MAE	25.26	19.12	17.69	16.75	17.58	15.52	<u>15.51</u>	**15.41**
FAC2	0.64	0.86	0.87	0.89	0.85	<u>0.90</u>	0.89	**0.90**

S8.2. *Robustness to the linearity.* The $PM_{2.5}$ concentrations are often nonlinearly affected by different weather conditions such as temperature, wind power (Liang et al., 2015). We investigate the robustness of the proposed approach to the linearity by comparing the proposed models with the additive model (ADM). We use ADM of the form

$$y_t(\boldsymbol{s}) = \beta_0 + \beta_1 x_{t,1}(C_{\boldsymbol{s}}) + \sum_{\kappa=2}^{4} g_{\kappa-1}\big(x_{t,\kappa}(C_{\boldsymbol{s}})\big) + g_4\big(s^x, s^y, t\big)\big) + \varepsilon_t(s), \tag{S9}$$

where $x_{t,1}(C_{\boldsymbol{s}})$ is the weighted outputs of CMAQ at location $\boldsymbol{s}$ and at time point t, $\{x_{t,2}(C_{\boldsymbol{s}}), x_{t,3}(C_{\boldsymbol{s}}), x_{t,4}(C_{\boldsymbol{s}})\}$ represent the weighted surface temperature, northern cumulative wind power, and eastern cumulative wind power, respectively; and s^x and s^y denote longitude and latitude, respectively. The functions $\{g_k(\cdot)\}, k = 1, \cdots, 4$ are estimated by basis expansion $g_k\Big(z_{t,k}(\boldsymbol{s})\Big) = \sum_{j=1}^{J_k} \alpha_{k,j}\phi_{k,j}$ with $J_k = 5$ for univariate functions $\{g_k(\cdot)\}_{k=1}^3$ and $J_4 = 180$ for the three-dimensional smooth function $g_4(\cdot)$. We use cyclic cubic-spline basis functions for $\{\phi_{k,j}\}_{k=1}^3$ and thin plate splines for $\phi_{4,j}$. See Wood (2017) for more discussions on additive models. ADM is implemented using the `R` package `mgcv` (Wood, 2022), and the results of cross-validation with ADM (S9) are presented in Table S10.

ADM (S9) is used to fit the BTH $PM_{2.5}$ concentration data in the summer of 2015. Although relationships between the $PM_{2.5}$ readings and other meteorological variables are nonlinear, the average value of the different metrics across all cities shows that HDCM (or $HDCM_2$) using the linear trends outperforms ADM (S9) using the nonlinear trends; see Tables S4-S6 for the performance of HDCM and Table S10 for ADM. Therefore, the proposed approach is robust to the deviation from the linearity assumption on fixed effects.

TABLE S10

Averaged RMSE (Root Mean Squared Error), CRPS (Continuous Rank Probability Score), 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 the additive model (ADM). Daily data from November 1, 2015 to January 31, 2016 are considered. Daily data from June 1, 2015, and August 31, 2015, are considered.

City	RMSE	CRPS	MAE	FAC2
Baoding	24.24	12.97	17.60	0.95
Beijing	24.86	13.30	17.55	0.90
Cangzhou	18.69	10.62	14.65	0.94
Chengde	18.80	10.32	14.05	0.77
Handan	24.40	13.81	19.54	0.96
Hengshui	23.19	12.51	17.18	0.96
Langfang	17.85	10.25	13.74	0.95
Qinhuangdao	25.90	13.66	17.45	0.66
Shijiazhuang	27.47	15.33	21.90	0.87
Tangshan	25.39	14.60	21.04	0.89
Tianjin	19.03	10.92	15.35	0.90
Xingtai	24.13	13.25	18.22	0.97
Zhangjiakou	43.18	24.14	34.90	0.47
Average	24.39	13.51	18.70	0.86

S9. Variational Bayes for HDCM parameters.

S9.1. *Some additional results from spEnKS.* By Algorithm 1 of the Manuscript, we have $q_{v_t}(\boldsymbol{v}_t) \sim \mathcal{N}\left(\widehat{\boldsymbol{v}}_{t|N_t}, \widehat{\boldsymbol{\Sigma}}_{t,t|N_t}\right)$ for $t \leq N_t$. Furthermore, the following results are necessary to run the VB procedure from (S15) to (S22):

$$\boldsymbol{S}_{00}^{(r)} = \sum_{t=1}^{N_t} \langle \boldsymbol{v}_{t-1}^{(r)} \boldsymbol{v}_{t-1}^{T,(r)} \rangle = \sum_{t=1}^{N_t} \left(\widehat{\boldsymbol{v}}_{t-1|N_t}^{(r)} \widehat{\boldsymbol{v}}_{t-1|N_t}^{T,(r)} + \widehat{\boldsymbol{\Sigma}}_{t-1,t-1|N_t}^{(r)} \right), \tag{S10}$$

$$\boldsymbol{S}_{10}^{(r)} = \sum_{t=1}^{N_t} \langle \boldsymbol{v}_{t}^{(r)} \boldsymbol{v}_{t-1}^{T,(r)} \rangle = \sum_{t=1}^{N_t} \left(\widehat{\boldsymbol{v}}_{t|N_t}^{(r)} \widehat{\boldsymbol{v}}_{t-1|N_t}^{T,(r)} + \widehat{\boldsymbol{\Sigma}}_{t,t-1|N_t}^{(r)} \right), \tag{S11}$$

$$\boldsymbol{S}_{11}^{(r)} = \sum_{t=1}^{N_t} \langle \boldsymbol{v}_{t}^{(r)} \boldsymbol{v}_{t}^{T,(r)} \rangle = \sum_{t=1}^{N_t} \left(\widehat{\boldsymbol{v}}_{t|N_t}^{(r)} \widehat{\boldsymbol{v}}_{t|N_t}^{T,(r)} + \widehat{\boldsymbol{\Sigma}}_{t,t|N_t}^{(r)} \right), \tag{S12}$$

$$\boldsymbol{S}_{11} = \operatorname{diag}\left(\boldsymbol{S}_{11}^{(1)}, \cdots, \boldsymbol{S}_{11}^{(R)} \right). \tag{S13}$$

S9.2. *Coordinate ascent mean-field VB for HDCM.* Following Blei, Kucukelbir and McAuliffe (2017), we decompose $\log p(\boldsymbol{y})$ into the following two terms:

$$\begin{aligned}
\log p(\boldsymbol{y}) &= \log p(\boldsymbol{y}, \boldsymbol{\Theta}, \boldsymbol{v}) - \log p(\boldsymbol{\Theta}, \boldsymbol{v}|\boldsymbol{y}) \\
&= \log \left\{ \frac{p(\boldsymbol{y}, \boldsymbol{\Theta}, \boldsymbol{v})}{q(\boldsymbol{\Theta}, \boldsymbol{v})} \right\} - \log \left\{ \frac{p(\boldsymbol{\Theta}, \boldsymbol{v}|\boldsymbol{y})}{q(\boldsymbol{\Theta}, \boldsymbol{v})} \right\} \\
&= \int q(\boldsymbol{\Theta}, \boldsymbol{v}) \log \left\{ \frac{p(\boldsymbol{y}, \boldsymbol{\Theta}, \boldsymbol{v})}{q(\boldsymbol{\Theta}, \boldsymbol{v})} \right\} - \int q(\boldsymbol{\Theta}, \boldsymbol{v}) \log \left\{ \frac{p(\boldsymbol{\Theta}, \boldsymbol{v}|\boldsymbol{y})}{q(\boldsymbol{\Theta}, \boldsymbol{v})} \right\} \\
&= \mathcal{L}(q) + \mathtt{KL}[q(\boldsymbol{\Theta}, \boldsymbol{v}) \parallel p(\boldsymbol{\Theta}, \boldsymbol{v}|\boldsymbol{y})] \geq \mathcal{L}(q).
\end{aligned} \tag{S14}$$

We approximate the posterior using $q \equiv q(\boldsymbol{\Theta}, \boldsymbol{v})$. The objective of optimization is to minimize the Kullback-Leibler divergence $\mathtt{KL}[q(\boldsymbol{\Theta}, \boldsymbol{v}) | p(\boldsymbol{\Theta}, \boldsymbol{v} | \boldsymbol{y})]$, or equivalently, to maximize the function $\mathcal{L}(q)$ with respect to the function q.

Using the mean-field approximation method (Blei, Kucukelbir and McAuliffe, 2017), we assume $q(\boldsymbol{\Theta}, \boldsymbol{v})$ in (S14) is separable and obtain $q = q_\Theta(\boldsymbol{\Theta}) q_v(\boldsymbol{v})$ where $q_\Theta(\boldsymbol{\Theta}) = q_\beta(\boldsymbol{\beta}) q_{\sigma^2}(\sigma^2) q_{\boldsymbol{\tau}_0^2}(\boldsymbol{\tau}_0^2) q_{\boldsymbol{\tau}^2}(\boldsymbol{\tau}^2) q_{\zeta_0^2}(\zeta_0^2) q_{\zeta^2}(\zeta^2) q_{\boldsymbol{\theta}_1}(\boldsymbol{\theta}_1) q_{\boldsymbol{\theta}_2}(\boldsymbol{\theta}_2)$. Note that $q_v(\boldsymbol{v})$ is approximated using the smoothing ensemble from spEnKS of Section 3.2.4 of the Manuscript, and a complete coordinate ascent procedure for inferring $q_\Theta(\boldsymbol{\Theta})$ is described as follows:

a. For the regression coefficient $\boldsymbol{\beta}$, we set $\pi(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$ and have

$$q_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{D}_\beta \boldsymbol{\xi}_\beta, \boldsymbol{D}_\beta), \tag{S15}$$

where $\boldsymbol{D}_\beta = \left(\langle \frac{1}{\sigma^2} \rangle \sum_{t=1}^{N_t} \boldsymbol{X}_t^T \boldsymbol{X}_t + \boldsymbol{\Sigma}_\beta^{-1} \right)^{-1}$ and $\boldsymbol{\xi}_\beta = \langle \frac{1}{\sigma^2} \rangle \sum_{t=1}^{N_t} \boldsymbol{X}_t^T \left(\boldsymbol{y}_t - \boldsymbol{H} \widehat{\boldsymbol{v}}_{t|N_t} \right) + \boldsymbol{\Sigma}_\beta^{-1} \boldsymbol{\mu}_\beta$.

b. For variance of the noise σ^2, we set $\pi(\sigma^2) = IG(a_{\sigma^2}, b_{\sigma^2})$ with the shape a_{σ^2} and the scale b_{σ^2}, then

$$q_{\sigma^2}(\sigma^2) = IG\Big(a_{\sigma^2} + nN_t/2, b_{\sigma^2} + f_{\sigma^2}/2 \Big), \tag{S16}$$

where "IG" denotes an Inverse Gamma distribution, $f_{\sigma^2} = \sum_{t=1}^{N_t} \Big[\boldsymbol{y}_t^T \left(\boldsymbol{y}_t - 2 \boldsymbol{X}_t \langle \boldsymbol{\beta} \rangle \right) - 2 \widehat{\boldsymbol{v}}_t^T \boldsymbol{H}^T \left(\boldsymbol{y}_t - \boldsymbol{X}_t \langle \boldsymbol{\beta} \rangle \right) + \mathtt{Tr} \Big\{ \boldsymbol{X}_t^T \boldsymbol{X}_t \left(\boldsymbol{D}_\beta + \langle \boldsymbol{\beta} \boldsymbol{\beta}^T \rangle \right) \Big\} \Big] + \mathtt{Tr} \left(\boldsymbol{H}^T \boldsymbol{H} \boldsymbol{S}_{11} \right)$, and $\boldsymbol{S}_{11}$ is defined in (S13), here $\mathtt{Tr}(\boldsymbol{A})$ denotes the trace of the matrix $\boldsymbol{A}$.

c. For the process transition parameter $\theta_1^{(r)}$ of IDE (4) of the Manuscript at the rth subregion, we set $\pi(\theta_1^{(r)}) = \mathcal{N}\big(\mu_{\theta_1^{(r)}}, \sigma^2_{\theta_1^{(r)}}\big)$, then

$$q_{\theta_1^{(r)}}\big(\theta_1^{(r)}\big) = \mathcal{N}\Big(D_{\theta_1^{(r)}} \xi_{\theta_1^{(r)}}, D_{\theta_1^{(r)}} \Big), \tag{S17}$$

where $D_{\theta_1} = \Big\{ \mathtt{Tr}\Big(\boldsymbol{\mathcal{M}}_{\langle \theta_2^{(r)} \rangle} \langle \boldsymbol{Q}^{(r)} \rangle \boldsymbol{\mathcal{M}}_{\langle \theta_2^{(r)} \rangle} \boldsymbol{S}_{00}^{(r)} \Big) + \sigma^{-2}_{\theta_1^{(r)}} \Big\}^{-1}, \xi_{\theta_1} = \mathtt{Tr}\Big(\boldsymbol{\mathcal{M}}_{\langle \theta_2^{(r)} \rangle} \langle \boldsymbol{Q}^{(r)} \rangle \boldsymbol{S}_{10}^{(r)} \Big) + \mu_{\theta_1^{(r)}} \sigma^{-2}_{\theta_1^{(r)}}$, $\boldsymbol{S}_{00}^{(r)}$ and $\boldsymbol{S}_{10}^{(r)}$ are defined in (S10) and (S11), respectively.

d. For the scale parameter $\tau^{2,(r)}$ of IDE (4) of the Manuscript at the rth subregion, let $\pi(\tau^{2,(r)}) = \text{Gamma}\left(a_{\tau^{2,(r)}}, b_{\tau^{2,(r)}}\right)$ with the shape $a_{\tau^{2,(r)}}$ and the scale $b_{\tau^{2,(r)}}$. We have

$$q_{\tau^{2,(r)}}(\tau^{2,(r)}) = \text{Gamma}\big(\tilde{a}_{\tau^{2,(r)}}, \tilde{b}_{\tau^{2,(r)}} \big), \tag{S18}$$

with $\tilde{a}_{\tau^{2,(r)}} = a_{\tau^{2,(r)}} + m_r N_t/2$, $\tilde{b}_{\tau^{2,(r)}} = b_{\tau^{2,(r)}} + f_{\tau^{2,(r)}}/2$, where let $\langle \boldsymbol{\Lambda}^{(r)} \rangle = \boldsymbol{G}^{(r)} + \langle \zeta^{2,(r)} \rangle \boldsymbol{I}_{m_r}$ and

$$f_{\tau^{2,(r)}} = \mathtt{Tr}\Big(\langle \boldsymbol{\Lambda}^{(r)} \rangle \boldsymbol{S}_{11}^{(r)} - 2 \langle \widehat{\theta}_1^{(r)} \rangle \boldsymbol{\mathcal{M}}_{\langle \theta_2^{(r)} \rangle} \langle \boldsymbol{\Lambda}^{(r)} \rangle \boldsymbol{S}_{10}^{(r)} + \langle \widehat{\theta}_1^{2,(r)} \rangle \boldsymbol{\mathcal{M}}_{\langle \theta_2^{(r)} \rangle} \langle \boldsymbol{\Lambda}^{(r)} \rangle \boldsymbol{\mathcal{M}}_{\langle \theta_2^{(r)} \rangle} \boldsymbol{S}_{00}^{(r)} \Big).$$

e. For the scale parameter $\tau_0^{2,(r)}$ of the initial state of IDE (4) of the Manuscript at the rth subregion, we set $\pi(\tau_0^{2,(r)}) = \text{Gamma}\big(a_{\tau_0^{2,(r)}}, b_{\tau_0^{2,(r)}}\big)$ and have

$$q_{\tau_0^{2,(r)}}(\tau_0^{2,(r)}) = \text{Gamma}\big(\tilde{a}_{\tau_0^{2,(r)}}, \tilde{b}_{\tau_0^{2,(r)}} \big), \tag{S19}$$

with $\tilde{a}_{\tau_0^{2,(r)}} = a_{\tau_0^{2,(r)}} + m_r/2$, $\tilde{b}_{\tau_0^{2,(r)}} = b_{\tau_0^{2,(r)}} + \mathtt{Tr}\big\{ \langle \boldsymbol{\Lambda}_0^{(r)} \rangle \langle \boldsymbol{v}_0 \boldsymbol{v}_0^T \rangle \big\}/2$, here $\langle \boldsymbol{\Lambda}_0^{(r)} \rangle = \boldsymbol{G}^{(r)} + \langle \zeta_0^{2,(r)} \rangle \boldsymbol{I}_{m_r}$.

f. Although $q_{\theta_2^{(r)}}, q_{\zeta^{2,(r)}}$, and $q_{\zeta_0^{2,(r)}}$ do not have closed form expressions, they are proportional to:

(S20) $$q_{\theta_2}(\theta_2^{(r)}) \propto \exp\big(-f_{\theta_2^{(r)}}/2\big)\pi(\theta_2^{(r)}),$$

(S21) $$q_{\zeta^{2,(r)}}(\zeta^{2,(r)}) \propto |\boldsymbol{Q}^{(r)}|^{\frac{N_t}{2}} \exp\big(-f_{\zeta^{2,(r)}}/2\big)\pi(\zeta^{(r)}),$$

(S22) $$q_{\zeta_0^{2,(r)}}(\zeta_0^{2,(r)}) \propto |\boldsymbol{Q}_0^{(r)}|^{\frac{1}{2}} \exp\big(-f_{\zeta_0^{2,(r)}}/2\big)\pi(\zeta_0^{(r)}),$$

where

$$f_{\theta_2} = \mathtt{Tr}\left\{-2\langle\theta_1^{(r)}\rangle\boldsymbol{\mathcal{M}}_{\theta_2^{(r)}}\langle\boldsymbol{Q}^{(r)}\rangle\boldsymbol{S}_{10}^{(r)} + \big(\langle\theta_1^{2,(r)}\rangle + D_{\theta_1^{(r)}}\big)\boldsymbol{\mathcal{M}}_{\theta_2^{(r)}}\langle\boldsymbol{Q}^{(r)}\rangle\boldsymbol{\mathcal{M}}_{\theta_2^{(r)}}\boldsymbol{S}_{00}^{(r)}\right\},$$

$$f_{\zeta^{2,(r)}} = \mathtt{Tr}\left\{\boldsymbol{S}_{11}^{(r)} - 2\langle\theta_1^{(r)}\rangle\boldsymbol{\mathcal{M}}_{\langle\theta_2^{(r)}\rangle}\boldsymbol{S}_{10}^{(r)} + \big(\langle\theta_1^{2,(r)}\rangle + D_{\theta_1^{(r)}}\big)\boldsymbol{\mathcal{M}}_{\langle\theta_2^{(r)}\rangle}\boldsymbol{\mathcal{M}}_{\langle\theta_2^{(r)}\rangle}\boldsymbol{S}_{00}^{(r)}\right\}\langle\tau^{2,(r)}\rangle\zeta^{2,(r)},$$

$$f_{\zeta_0^{2,(r)}} = \mathtt{Tr}\left\{\big(\widehat{\boldsymbol{v}}_0^{(r)}\widehat{\boldsymbol{v}}_0^{T,(r)} + \widehat{\boldsymbol{\Sigma}}_{0,0|N_t}^{(r)}\big)\right\}\langle\tau_0^{2,(r)}\rangle\zeta_0^{2,(r)}.$$

Based on the results above, these three q functions can be approximated with the Laplace method (Section 4.7 of Lindgren and Rue (2015)). Section S10 provides further details on the Laplace approximation related to the proposed approach.

S10. The Laplace approximation for the parameter without the conjugate prior. Assume the function $f(\xi)$ is the density of a random variable ξ. Based on the Laplace method, the integral of $f(\xi)$ can be approximated by a Gaussian integral. Specifically, the mean of the Gaussian distribution is a stationary point of $\log f(\xi)$ such that $\frac{\partial \log f(\xi)}{\partial \xi} = 0$, denoted by ξ_0, and the variance is $-1/\frac{\partial^2 \log f(\xi)}{\partial^2 \xi}\big|_{\xi=\xi_0}$, where the vertical bar indicates a specific value of a parameter. See Section 4.7 of Lindgren and Rue (2015) for more details on this topic.

Several parameters in the IDE (4) of the Manuscript do not have the conditionally conjugate prior, including θ_2, ζ_0, and ζ, leading to no closed-form expressions for their q functions. But because these functions are proportional to (S20)-(S22) of the manuscript, we can approximate their posterior distributions with the Gaussian by finding corresponding maximum posterior points by (S20), (S21), and (S22), respectively. Below are the relevant details:

1. Find the maximum posteriori estimate $\theta_2^{*,(r)}$ for $\theta_2^{(r)}$ by solving $\frac{\partial \log \pi\left(\theta_2^{(r)}\right)}{\partial \theta_2^{(r)}} - \frac{1}{2}\frac{\partial f_{\theta_2^{(r)}}}{\partial \theta_2^{(r)}} = 0$, where $\frac{\partial \log \pi\left(\theta_2^{(r)}\right)}{\partial \theta_2^{(r)}} = 0$ for an uniform distribution $\pi\left(\theta_2^{(r)}\right)$. The second term on the left-hand side of the equation above can be written as

$$\frac{\partial f_{\theta_2^{(r)}}}{\partial \theta_2^{(r)}} = \mathtt{Tr}\left\{-2\langle\theta_1^{(r)}\rangle\frac{\partial \boldsymbol{\mathcal{M}}_{\theta_2^{(r)}}}{\partial \theta_2^{(r)}}\langle\boldsymbol{Q}^{(r)}\rangle\boldsymbol{S}_{10}^{(r)} + \big(\langle\theta_1^{2,(r)}\rangle + D_{\theta_1^{(r)}}\big)\frac{\partial f^{\mathcal{M}}_{\theta_2^{(r)}}}{\partial \theta_2^{(r)}}\boldsymbol{S}_{00}^{(r)}\right\},$$

where the (l,l')th element of $\frac{\partial \boldsymbol{\mathcal{M}}_{\theta_2^{(r)}}}{\partial \theta_2^{(r)}}$ is $\frac{\partial W\left(d_{l,l'};\theta_2^{(r)}\right)}{\partial \theta_2^{(r)}} = \frac{d_{l,l'}}{4\theta_2^{2,(r)}}\big\{\big(1 - d_{l,l'}/\theta_2^{(r)}\big)^2\big(1 + 3d_{l,l'}/\theta_2^{(r)}\big) - \big(1 - d_{l,l'}/\theta_2^{(r)}\big)^3\big\}$, $f^{\mathcal{M}}_{\theta_2^{(r)}} = \boldsymbol{\mathcal{M}}_{\theta_2^{(r)}}\langle\boldsymbol{Q}^{(r)}\rangle\boldsymbol{\mathcal{M}}_{\theta_2^{(r)}}$, and the derivative of $f^{\mathcal{M}}_{\theta_2^{(r)}}$ is given by $\frac{\partial f^{\mathcal{M}}_{\theta_2^{(r)}}}{\partial \theta_2^{(r)}} = \frac{\partial \boldsymbol{\mathcal{M}}_{\theta_2^{(r)}}}{\partial \theta_2^{(r)}}\langle\boldsymbol{Q}^{(r)}\rangle\boldsymbol{\mathcal{M}}_{\theta_2^{(r)}} + \boldsymbol{\mathcal{M}}_{\theta_2^{(r)}}\langle\boldsymbol{Q}^{(r)}\rangle\frac{\partial \boldsymbol{\mathcal{M}}_{\theta_2^{(r)}}}{\partial \theta_2^{(r)}}$. Moreover, $\frac{\partial^2 f_{\theta_2^{(r)}}}{\partial \theta_2^{(r)}} = \mathtt{Tr}\Big\{-2\langle\theta_1\rangle\frac{\partial^2 \boldsymbol{\mathcal{M}}_{\theta_2^{(r)}}}{\partial \theta_2^{(r)}}\langle\boldsymbol{Q}^{(r)}\rangle\boldsymbol{S}_{10}^{(r)} + \big(\langle\theta_1^{2,(r)}\rangle + D_{\theta_1^{(r)}}\big)\frac{\partial^2 f^{\mathcal{M}}_{\theta_2^{(r)}}}{\partial \theta_2^{(r)}}\boldsymbol{S}_{00}^{(r)}\Big\}$, where $\frac{\partial^2 f^{\mathcal{M}}_{\theta_2^{(r)}}}{\partial \theta_2^{(r)}} = \frac{\partial^2 \boldsymbol{\mathcal{M}}_{\theta_2^{(r)}}}{\partial \theta_2^{(r)}}\langle\boldsymbol{Q}^{(r)}\rangle\boldsymbol{\mathcal{M}}_{\theta_2^{(r)}} + 2\frac{\partial \boldsymbol{\mathcal{M}}_{\theta_2^{(r)}}}{\partial \theta_2^{(r)}}\langle\boldsymbol{Q}^{(r)}\rangle\frac{\partial \boldsymbol{\mathcal{M}}_{\theta_2^{(r)}}}{\partial \theta_2^{(r)}} + \boldsymbol{\mathcal{M}}_{\theta_2^{(r)}}\langle\boldsymbol{Q}^{(r)}\rangle\frac{\partial^2 \boldsymbol{\mathcal{M}}_{\theta_2^{(r)}}}{\partial \theta_2^{(r)}}$, here the (l,l')th element of $\frac{\partial^2 \boldsymbol{\mathcal{M}}_{\theta_2^{(r)}}}{\partial \theta_2^{(r)}}$ is $\frac{\partial^2 W\left(d_{l,l'};\theta_2^{(r)}\right)}{\partial \theta_2^{(r)}} = -\frac{d_{l,l'}}{2\theta_2^{3,(r)}}\Big\{\big(1-d_{l,l'}/\theta_2^{(r)}\big)^2\big(1+3d_{l,l'}/\theta_2^{(r)}\big) - \big(1-d_{l,l'}/\theta_2^{(r)}\big)^3\Big\} +$

$\frac{d_{l,l'}^2}{2\theta_2^{4,(r)}}\Big\{\big(1-d_{l,l'}/\theta_2^{(r)}\big)^2\big(1+3d_{l,l'}/\theta_2^{(r)}\big)-3\big(1-d_{l,l'}/\theta_2^{(r)}\big)^2\Big\}$. We then have $\theta_2^{(r)}\sim \mathcal{N}\Big(\theta_2^{*,(r)},-1\Big/\frac{\partial^2 f^{\mathcal{M}}_{\theta_2^{(r)}}}{\partial\theta_2^{(r)}}\Big|_{\theta_2^{(r)}=\theta_2^{*,(r)}}\Big)$.

2. Find the maximum posteriori estimate $\zeta^{2*,(r)}$ for $\zeta^{2,(r)}$ by solving $N_t\frac{\partial\log\left|\boldsymbol{G}^{(r)}+\zeta^{2,(r)}\boldsymbol{I}_{m_r}\right|}{\partial\zeta^{2,(r)}}-\frac{\partial f_{\zeta^{2,(r)}}}{\partial\zeta^{2,(r)}}+\frac{\partial\log\pi\left(\zeta^{2,(r)}\right)}{\partial\zeta^{2,(r)}}=0$, where $\frac{\partial\log\left|\boldsymbol{G}^{(r)}+\zeta^{2,(r)}\boldsymbol{I}_{m_r}\right|}{\partial\zeta^{2,(r)}}=\mathtt{Tr}\Big\{\big(\boldsymbol{G}^{(r)}+\zeta^{2,(r)}\boldsymbol{I}_{m_r}\big)^{-1}\Big\}$, $\frac{\partial f_{\zeta^{2,(r)}}}{\partial\zeta^{2,(r)}}=\mathtt{Tr}\left\{\boldsymbol{S}_{11}^{(r)}-2\langle\theta_1^{(r)}\rangle\boldsymbol{\mathcal{M}}_{\langle\theta_2^{(r)}\rangle}\boldsymbol{S}_{10}^{(r)}+\Big(\langle\theta_1^{2,(r)}\rangle+D_{\theta_1^{(r)}}\Big)\boldsymbol{\mathcal{M}}_{\langle\theta_2^{(r)}\rangle}\boldsymbol{\mathcal{M}}_{\langle\theta_2^{(r)}\rangle}\boldsymbol{S}_{00}^{(r)}\right\}\langle\tau^{2,(r)}\rangle$, and $\frac{\partial\log\pi\left(\zeta^{2,(r)}\right)}{\partial\zeta^{2,(r)}}=0$. On the other hand, we find $\sigma^2_{\zeta^{2*,(r)}}=-1\Big/\frac{\partial^2\log\left|\boldsymbol{G}^{(r)}+\zeta^{2,(r)}\boldsymbol{I}_{m_r}\right|}{\partial\zeta^{4,(r)}}\Big|_{\sigma^2_{\zeta^{2,(r)}}=\zeta^{2*,(r)}}$, where $\frac{\partial^2\log\left|\boldsymbol{G}^{(r)}+\zeta^{2,(r)}\boldsymbol{I}_{m_r}\right|}{\partial\zeta^{4,(r)}}=-\big(\boldsymbol{G}^{(r)}+\zeta^{2,(r)}\boldsymbol{I}_{m_r}\big)^{-1}\big(\boldsymbol{G}^{(r)}+\zeta^{2,(r)}\boldsymbol{I}_{m_r}\big)^{-1}$ by using $\frac{\partial\mathtt{Tr}\widetilde{\boldsymbol{G}}^{(r)}}{\partial\widetilde{\boldsymbol{G}}^{(r)}}=\boldsymbol{I}_{m_r}$ and $\frac{\partial\widetilde{\boldsymbol{G}}^{-1,(r)}}{\partial\zeta^{2,(r)}}=-\widetilde{\boldsymbol{G}}^{-1,(r)}\frac{\partial\widetilde{\boldsymbol{G}}^{(r)}}{\partial\zeta^{2,(r)}}\widetilde{\boldsymbol{G}}^{-1,(r)}$, here $\widetilde{\boldsymbol{G}}^{(r)}=\boldsymbol{G}^{(r)}+\zeta^{2,(r)}\boldsymbol{I}_{m_r}$.
3. Find the maximum posteriori estimate $\zeta_0^{2*,(r)}$ for $\zeta_0^{2,(r)}$ by solving $\frac{\partial\log\left|\boldsymbol{G}^{(r)}+\zeta_0^{2,(r)}\boldsymbol{I}_{m_r}\right|}{\partial\zeta_0^{2,(r)}}-\frac{\partial f_{\zeta_0^{2,(r)}}}{\partial\zeta_0^{2,(r)}}+\frac{\partial\log\pi\left(\zeta_0^{2,(r)}\right)}{\partial\zeta_0^{2,(r)}}=0$, where $\frac{\partial\log\left|\boldsymbol{G}^{(r)}+\zeta_0^{2,(r)}\boldsymbol{I}_{m_r}\right|}{\partial\zeta_0^{2,(r)}}=\mathtt{Tr}\Big\{\big(\boldsymbol{G}^{(r)}+\zeta_0^{2,(r)}\boldsymbol{I}_{m_r}\big)^{-1}\Big\}$, $\frac{\partial f_{\zeta_0^{2,(r)}}}{\partial\zeta_0^{2,(r)}}=\mathtt{Tr}\left\{\big(\widehat{\boldsymbol{v}}_0\widehat{\boldsymbol{v}}_0^T+\widehat{\boldsymbol{\Sigma}}_{0,0|N_t}\big)\right\}\langle\tau_0^{2,(r)}\rangle$, and $\frac{\partial\log\pi\left(\zeta_0^{2,(r)}\right)}{\partial\zeta_0^{2,(r)}}=0$. On the other hand, we find $\sigma^2_{\zeta_0^{2*,(r)}}=-1\Big/\frac{\partial^2\log\left|\boldsymbol{G}^{(r)}+\zeta_0^{2,(r)}\boldsymbol{I}_{m_r}\right|}{\partial\zeta_0^{4,(r)}}\Big|_{\sigma^2_{\zeta_0^{2,(r)}}=\zeta_0^{2*,(r)}}$.

In practice, when optimization is used to find a stationary point, to make the algorithm more stable, the logarithmic transformation may be useful for these three parameters, i.e., $\log\theta_2^{(2)}=\vartheta_1,\log\zeta_0^{2,(r)}=\vartheta_2$, and $\log\zeta^{2,(r)}=\vartheta_3$, while other details remain the same as above.

REFERENCES

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.

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.

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*. John Wiley & Sons, New York.

BLEI, D. M., KUCUKELBIR, A. and MCAULIFFE, J. D. (2017). Variational inference: A review for statisticians. *J. Am. Stat. Assoc.* **112** 859–877.

BOLIN, D., WALLIN, J. and LINDGREN, F. (2019). Latent Gaussian random field mixture models. *Comput. Stat. Data Anal.* **130** 80–93.

DIEBOLD, F. X. and MARIANO, R. S. (1995). Comparing predictive accuracy. *Journal of Business & economic statistics* **13** 253–263.

GNEITING, T. (2002). Nonseparable, stationary covariance functions for space-time data. *J. Am. Stat. Assoc.* **97** 590–600.

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

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

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.

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., 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 Methodol.* **73** 423–498.

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

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.

RODU, J. and KAFADAR, K. (2022). The q-q Boxplot. *J. Comput. Graph. Stat.* **31** 26–39.

RUE, H. and HELD, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall/CRC, Boca Raton, FL.

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 Methodol.* **71** 319–392.

RUE, H., RIEBLER, A., SØRBYE, S. H., ILLIAN, J. B., SIMPSON, D. P. and LINDGREN, F. K. (2017). Bayesian computing with INLA: a review. *Annu. Rev. Stat. Appl.* **4** 395–421.

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

WOOD, S. N. (2017). *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC, New York.

WOOD, S. N. (2022). Package 'mgcv'. R package version 1.8-41. https://CRAN.R-project.org/package=mgcv.