HHS Public Access

Author manuscript

Ann Appl Stat. Author manuscript; available in PMC 2021 June 09.

Published in final edited form as:

Ann Appl Stat. 2021 March ; 15(1): 323–342. doi:10.1214/20-aoas1399.

SPATIAL DISTRIBUTED LAG DATA FUSION FOR ESTIMATING AMBIENT AIR POLLUTION

JOSHUA L. WARREN[1], **MARIE LYNN MIRANDA**[2], **JOSHUA L. TOOTOO**[3,*], **CLAIRE E. OSGOOD**[3,†], **MICHELLE L. BELL**[4]

[1]Department of Biostatistics, Yale University

[2]Department of Applied and Computational Mathematics and Statistics, University of Notre Dame

[3]Children's Environmental Health Initiative, University of Notre Dame

[4]School of Forestry and Environmental Studies, Department of Environmental Health Sciences, Yale University

Abstract

We introduce spatial (DLfuse) and spatiotemporal (DLfuseST) distributed lag data fusion methods for predicting point-level ambient air pollution concentrations, using, as input, gridded average pollution estimates from a deterministic numerical air quality model. The methods incorporate predictive information from grid cells surrounding the prediction location of interest and are shown to collapse to existing downscaling approaches when this information adds no benefit. The spatial lagged parameters are allowed to vary spatially/spatiotemporally to accommodate the setting where surrounding geographic information is useful in one area/time but not in another. We apply the new methods to predict ambient concentrations of eight-hour maximum ozone and 24-hour average $PM_{2.5}$ at unobserved spatial locations and times, and compare the predictions with those from several state-of-the-art data fusion approaches. Results show that DLfuse and DLfuseST often provide improved model fit and predictive accuracy when the lagged information is shown to be beneficial. Code to apply the methods is available in the R package DLfuse.

Keywords

Air pollution; downscaling; spatial distributed lags; varying coefficients

1. Introduction.

Obtaining accurate estimates of ambient air pollutant concentrations at locations outside of those where data collection routinely occurs is vital to investigations of associations between

* jtootoo@nd.edu;. † cosgood@nd.edu.

SUPPLEMENTARY MATERIAL

Supplement A: Supplement to "Spatial distributed lag data fusion for estimating ambient air pollution" (DOI: 10.1214/20-AOAS1399SUPPA; .pdf). Supplemental tables and figures.

Supplement B: Supplement to "Spatial distributed lag data fusion for estimating ambient air pollution" (DOI: 10.1214/20-AOAS1399SUPPB; .zip). This package implements a hierarchical Bayesian spatially-varying (and spatiotemporally-varying) distributed lag regression analysis to predict ambient air pollution concentrations at new spatial locations and times.

exposure and adverse human health outcomes. This is especially true for rural areas which are consistently understudied and where reliable data are notably scarce. In addition, interpretations of associations between modeled air pollution data and health outcomes should, but often do not, take underlying uncertainty in the modeled exposure data into account. A number of methods for spatial/ spatiotemporal interpolation of air pollution concentrations have been previously used in this context, including inverse-distance weighting, land use regression and geostatistical approaches such as kriging (e.g., Berman et al. (2019), Gurung, Levy and Bell (2017), Hsu et al. (2019), Jin et al. (2019), Xu et al. (2019), Yu et al. (2019)).

More recently, advanced statistical techniques have been developed for combining multiple sources of air pollution information, potentially at different spatial scales, to obtain estimates of air pollution concentrations with improved accuracy and appropriate measures of uncertainty (e.g., Berrocal, Gelfand and Holland (2010a, 2010b, 2012), Fuentes and Raftery (2005), Guan et al. (2019), McMillan et al. (2010), Paciorek (2012), Reich, Chang and Foley (2014)). In order to model and to predict pollutant concentrations at locations and times without an active monitor, these methods typically use directly measured concentrations in combination with alternative pollutant estimates which have improved spatial/spatiotemporal coverage but which may be biased and/or produced at different spatial scales (e.g., grid averages). By estimating the associations between measured data and alternative sources of information, the methods leverage the full space-time coverage of the estimates to predict the measured concentrations at previously unobserved spatiotemporal locations.

The methods differ in how the auxiliary information is incorporated, with some specifying a joint model for measured and estimated concentrations (e.g., Fuentes and Raftery (2005)), some using the estimated concentrations directly as predictors along with spatially/ spatiotemporally-varying regression coefficients (e.g., Berrocal, Gelfand and Holland (2010a, 2010b, 2012)) and others using spectral methods to describe the associations (e.g., Guan et al. (2019), Reich, Chang and Foley (2014)). In a recent review, statistical data fusion techniques were shown to perform favorably to advanced machine learning algorithms with respect to predictive performance under the specified study settings (Berrocal et al. (2019)).

Regardless of how the auxiliary information is incorporated, the majority of these statistical models connect measured air pollutant concentrations with the closest unit of auxiliary information during model building (e.g., the grid cell that contains the air pollution monitor).

In this work we improve upon this limitation by introducing a spatial distributed lag data fusion model that incorporates potentially important auxiliary information from spatial locations surrounding the pollution monitor, not just the closest information. How much auxiliary information is used is determined by the data and allowed to vary spatially. This approach accommodates the notion that lagged predictors may be informative in some locations but uninformative in others. The model also includes spatially/ spatiotemporally-varying regression parameters to flexibly model complex biases in the auxiliary information and is shown to collapse to similar downscaling models (e.g., Berrocal, Gelfand and Holland (2010a)) in the setting where lagged information is not predictive of measured

concentrations. We also extend the spatially static version of our model to the spatiotemporal setting.

Berrocal, Gelfand and Holland (2012) introduced a model that incorporated auxiliary information from surrounding locations with the goal of improving predictions of measured concentrations. However, their model included each unit of auxiliary information (e.g., every grid cell estimate was used as a predictor) and spatially/spatiotemporally-varying regression parameters which allowed the set of important predictors to vary across space and time. With a large spatial domain and/or small grid cells, the number of predictors will be very large. As a result, a computationally efficient approximation was implemented during model fitting.

We avoid these computational issues by using spatially-lagged predictors, defined by taking the average of grid cell average concentrations in squares surrounding the grid cell containing the pollution monitor, similar to the spatial lags used in recent work from Baek et al. (2016). In Figure 1 an ambient air pollution monitor lies in the grid cell marked zero. Grid cells with the same number are used to calculate the corresponding lagged average of the auxiliary pollution information and are used as predictors in our newly developed model. We assume the larger the lag number (i.e., the further the distance from the index monitor grid cell), the lower the regression weight on the lagged average, reflecting the assumption that auxiliary information closer to the monitor are likely more reliably predictive than those further away. As a result, our method only introduces a new parameter for a grid cell that contains an active air pollution monitor, not for every grid cell as in Berrocal, Gelfand and Holland (2012). Due to the sparsity of pollution monitors with respect to the grid cells, this formulation leads to a drastic reduction in the number of unique predictors included in the model, avoiding the need for computational approximations and offering an intuitive and flexible framework for incorporating auxiliary information.

In Section 2 we describe the air pollution data sources used for modeling, predicting, and validating our new approaches. Section 3 introduces the spatial and spatiotemporal distributed lag data fusion methods with applications to eight-hour maximum ozone and 24-hour average particulate matter ≤ 2.5 micrometers ($PM_{2.5}$) given in Section 4. Also in Section 4, we include several state-of-the-art competing methods to compare predictive accuracy. We close in Section 5 with discussion and conclusions.

2. Data.

We analyze measured daily ambient concentrations of ozone (daily eight-hour maximum) and $PM_{2.5}$ (24-hour average) across the eastern United States (U.S.) during June 1–August 31, 2013. These data come from the air quality system (AQS) maintained by the U.S. Environmental Protection Agency (EPA) (https://www.epa.gov/aqs) and include the latitude and longitude of each monitoring location along with measured concentrations on each day where the monitor was active for a particular pollutant. For ozone we used the daily eight-hour maximum, and for $PM_{2.5}$ we used the 24-hour average to correspond to health-based regulations. The locations of the ozone and $PM_{2.5}$ AQS monitors that were active at any

point during the study period are shown in Figure 2 along with time series plots of the number of active monitors for each pollutant.

Additionally, comparable daily estimates of each pollutant produced by the Community Multiscale Air Quality (CMAQ) modeling system are obtained from the U.S. EPA (https://www.epa.gov/hesc/rsig-related-downloadable-data-files). The CMAQ model is a regional deterministic numerical air quality model that inputs information such as emissions and meteorology to produce estimates of a number of different pollutants across space and time. Whereas the AQS monitoring network can be sparse in space and time, CMAQ estimates have excellent space-time coverage and can be extremely useful when evaluating different air pollution scenarios (e.g., new emission regulations) (US EPA (2019)). Notably, the CMAQ concentrations are modeled and not measured; they can be biased, and this bias can differ by pollutant and across space and time. The daily CMAQ estimates we used are available on a 12-by-12 kilometer grid across the entire study region for both pollutants. The information at each grid cell includes the latitude, longitude of the grid centroid and the estimated average pollutant concentration within the grid cell. Unlike the AQS data, the CMAQ estimates are available on all days and at each grid cell during the study period.

3. Methods.

We develop spatial distributed lag data fusion methods for predicting pollutant concentrations at: (i) unobserved spatial locations based on data analyzed from a single day (spatial downscaler) and (ii) unobserved spatiotemporal locations/times based on data analyzed across multiple days (spatiotemporal downscaler). Improved predictions in space and time are made possible by incorporating CMAQ pollutant estimates from spatial lags surrounding each AQS monitor during modeling while computational efficiency is maintained through use of the distributed lag framework.

3.1. Distributed lag data fusion: Spatial downscaler.

We first introduce a data fusion model for air pollutant concentrations obtained from AQS monitors that were active during a single day. The model is given as

$$Y(\mathbf{s}_{ij}) = \tilde{\beta}_0(\mathbf{s}_{ij}) + \tilde{\beta}_1(\mathbf{s}_{ij}) \sum_{l=0}^{L} \bar{x}_{B_i,l} \left(\frac{\pi_{B_i,l}}{\sum_{k=0}^{L} \pi_{B_i,k}} \right) + \epsilon(\mathbf{s}_{ij}), \tag{3.1}$$

where $\epsilon(\mathbf{s}_{ij}) \mid \sigma_\epsilon^2 \overset{\text{iid}}{\sim} N(0, \sigma_\epsilon^2)$; $Y(\mathbf{s}_{ij})$ is the pollutant concentration (possibly transformed) measured at spatial location $\mathbf{s}_{ij}$, representing the location of AQS monitor j ($j = 1, \ldots, n_i$) that resides within CMAQ grid cell i (B_i; $i = 1, \ldots, m$); n_i is the number of active AQS monitors in B_i; m is the number of unique CMAQ grid cells that contain an active AQS monitor; $n = \sum_{i=1}^{m} n_i$ is the total number of observed AQS concentrations, and L is the number of included spatial lags for the CMAQ estimates (fixed at a large value). The average of the individual CMAQ estimates comprising spatial lag l surrounding CMAQ grid cell B_i (see Figure 1) is denoted as $\bar{x}_{B_i,l}$, where $\bar{x}_{B_i,0}$ represents the single CMAQ estimate

from the grid cell containing the AQS monitor. The spatially-varying weight and regression parameters are described in Sections 3.1.1 and 3.1.2, respectively.

3.1.1. Spatially-varying regression weights.—We use spatially-varying regression weights to describe the importance of average CMAQ estimates at different spatial lags in explaining patterns in the measured concentrations. The weight corresponding to the lag l average CMAQ estimate is given as

$$\frac{\pi_{B_i,l}}{\sum_{k=0}^{L}\pi_{B_i,k}}, \tag{3.2}$$

where the weights across all lags sum to one. We specify that $\pi_{B_i,l}$ is decreasing as one moves further away from the central CMAQ grid cell containing the AQS monitor (i.e., as l increases) to reflect the belief that CMAQ estimates closer to the AQS monitor are likely more predictive of $Y(\mathbf{s}_{ij})$ than those further away. In addition, we allow these weights to vary spatially to account for the possibility that the relative importance of the lagged average CMAQ estimates may differ due to spatially-varying bias in CMAQ. This specification results in a unique set of weights for each CMAQ location and the ability to spatially expand (i.e., large weights even at large lags) and constrict (i.e., large weights only at small lags) the amount of CMAQ information that is used in (3.1).

The corresponding model for $\pi_{B_i,l}$ is given as

$$\pi_{B_i,l} = \Phi\left(\mu + \alpha_{B_i}\right)^l,\ l = 0,\ldots,L,$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution, μ represents the global spatial lag structure common to all CMAQ locations and α_{B_i} is the deviation from the global lag structure specific to CMAQ grid cell B_i. The model is anchored at one for the CMAQ grid cell that contains the AQS monitor (i.e., $\pi_{B_i,0} = 1$) with values decreasing as the lag order increases. Additionally, we consider a second model for $\pi_{B_i,l}$ based on the spherical spatial isotropic correlation function such that

$$\pi_{B_i,l} = \left\{1.00 - 1.50\left(\frac{l}{\exp\left\{\mu + \alpha_{B_i}\right\}}\right) + 0.50\left(\frac{l}{\exp\left\{\mu + \alpha_{B_i}\right\}}\right)^3\right\},$$

when $l < \exp\left\{\mu + \alpha_{B_i}\right\}$ and $\pi_{B_i,l} = 0$ when $l \geq \exp\left\{\mu + \alpha_{B_i}\right\}$. Unlike the CDF-defined weights, this formulation allows for the weights to exactly equal zero after some estimated distance.

The α_{B_i} parameters are modeled using an intrinsic conditional autoregressive (ICAR) model (Besag (1974)) such that

$$\alpha_{B_i} \mid \boldsymbol{\alpha}_{-B_i}, \tau^2 \overset{\text{ind}}{\sim} \mathrm{N}\left(\frac{\sum_{j=1}^{m} z_{ij}\alpha_{B_j}}{\sum_{j=1}^{m} z_{ij}}, \frac{\tau^2}{\sum_{j=1}^{m} z_{ij}}\right),\ i = 1, \ldots, m, \tag{3.3}$$

where α_{-B_i} is the vector of all α_{B_j} parameters with α_{B_i} excluded and z_{ij} describes the spatial similarity between B_i and B_j. Because there is not an active AQS monitor in every CMAQ grid cell, we opt not to define z_{ij} based on shared common borders between CMAQ grid cells as is most common (i.e., many "islands" would be present across the spatial domain). Instead, we define z_{ij} as the inverse distance between B_i and B_j, with $z_{ii} = 0$ for all i representing the only zero entries in the full matrix (i.e., inverse distance computed for all pairs, not just those that are connected). We note that a Gaussian process could be used to model these parameters and would likely yield similar performance overall. However, we prefer the ICAR model for two reasons. First, even though we are using inverse distance to describe proximity, the α_{B_i} parameters arise on the CMAQ grid and not continuously over the spatial domain as the Gaussian process assumes. Therefore, the ICAR model more naturally accommodates this gridded setting from a conceptual perspective. Second, the conditional form of the ICAR model often provides computational benefits compared to working with a Gaussian process. Similarly, the ICAR model only introduces a single variance parameter where the GP includes a variance parameter and additional correlation parameter(s) that need to be estimated.

This model for the regression weights allows for different CMAQ areas to have unique lag structures that are spatially correlated, reflecting the idea that the bias in CMAQ estimation may be spatially smooth. Use of the final summation from (3.2) in (3.1) results in a weighted average of the lagged average CMAQ estimates. Critically, if $\pi_{B_i,k} = 0$ for all $k > 0$, then our model collapses to the original static downscaler method of Berrocal, Gelfand and Holland (2010a).

3.1.2. Spatially-varying regression parameters.—We allow for spatial variability in the regression parameters to account for the setting where multiple AQS monitors are located within a single CMAQ grid cell, similar to Berrocal, Gelfand and Holland (2010a). In that case the lag structure from (3.2) is not changing, leading to the exact same value of $\sum_{l=0}^{L} \bar{x}_{B_i,l}\left(\frac{\pi_{B_i,l}}{\sum_{k=0}^{L}\pi_{B_i,k}}\right)$ for those AQS concentrations. However, it may be important to consider that spatially-structured variability could remain with respect to the intercept and slope parameters for those concentrations.

Utilizing the linear model of coregionalization to allow for flexibility in each set of parameters as well as a general cross-covariance structure (Wackernagel (2013)), the joint model for the regression parameters is given as

$$\tilde{\beta}_k(\mathbf{s}_{ij}) = \beta_k + \beta_k(\mathbf{s}_{ij}),\ k = 0, 1,$$

where

$$\begin{pmatrix}\beta_0(\mathbf{s}_{ij})\\ \beta_1(\mathbf{s}_{ij})\end{pmatrix} = A\begin{pmatrix}w_0(\mathbf{s}_{ij})\\ w_1(\mathbf{s}_{ij})\end{pmatrix};\ A = \begin{pmatrix}A_{11} & 0\\ A_{21} & A_{22}\end{pmatrix}. \tag{3.4}$$

β_k represents the global intercept ($k = 0$) or slope ($k = 1$) parameter, and $\beta_k(\mathbf{s}_{ij})$ is the location-specific deviation from the global value. The spatially-correlated random effects are then modeled using a Gaussian process with spatially-structured correlation matrix such that

$$w_k = \left\{w_k(\mathbf{s}_{11}), \ldots, w_k(\mathbf{s}_{mn_m})\right\}^{\mathrm{T}} \mid \phi_k \overset{\text{ind}}{\sim} \mathrm{MVN}\{0, \Sigma_k(\phi_k)\},\ k = 0, 1,$$

where $\Sigma_k(\phi_k)$ describes the spatial correlation between entries of $\boldsymbol{w}_k$ such that $\mathrm{Corr}\{w_k(\mathbf{s}_{ij}), w_k(\mathbf{s}_{i'j'})\} = g_k(\|\mathbf{s}_{ij} - \mathbf{s}_{i'j'}\|; \phi_k)$; $g_k(\cdot\,; \phi_k)$ is an isotropic spatial correlation function, $\|\cdot\|$ represents the Euclidean distance function and $\boldsymbol{\phi}_k > 0$ describes the level of spatial correlation between parameters. The specification in (3.4) induces flexible correlation between the intercept and slope parameters (e.g., Warren et al. (2020)).

3.1.3. Prior specification.—We finalize the model by assigning prior distributions to the unknown parameters. The variance parameters are given weakly informative inverse gamma prior distributions such that $\sigma^2_\epsilon \sim \text{Inverse Gamma}(\alpha_{\sigma^2_\epsilon}, \beta_{\sigma^2_\epsilon})$ and $\tau^2 \sim \text{Inverse Gamma}(\alpha_{\tau^2}, \beta_{\tau^2})$ where $\alpha_{\sigma^2_\epsilon}$, $\beta_{\sigma^2_\epsilon}$, α_{τ^2} and β_{τ^2} are fixed at small values. The nonspatial components of the regression parameters are assigned weakly informative Gaussian priors such that $\beta_k \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2_\beta)$, $k = 0, 1$ with σ^2_β fixed at a large value. Similarly, the entries of A are assigned log-normal and Gaussian priors such that $\ln(A_{11})$, $\ln(A_{22})$, $A_{21} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2_A)$, with σ^2_A fixed. The nonspatial parameter of the regression weight definition is assigned an informative prior such that $\mu \sim \mathrm{N}(0, 1)$ due to identifiability concerns caused by use of $\Phi(\cdot)$ and the summation in (3.2). To explore the impact of this prior choice, we display 100 realizations of the weights across different lags generated by $\mu \sim \mathrm{N}(0, 1)$ (assuming all $\alpha_{B_i} = 0$) for both weight definitions in Figure S1 of the Supplementary Material (Warren et al. (2021)). The results suggest that the informative prior for μ does not have a large impact on the range of behavior of the weights at different lags for either weight definition. Both definitions allow realizations at the extremes, with weights close to (or equal to) zero almost immediately (i.e., at small lags) and with weights almost constant across all considered lags. Lastly, the spatial correlation hyperparameters are given weakly informative gamma priors such that $\phi_k \overset{\text{iid}}{\sim} \mathrm{Gamma}(\alpha_{\phi_k}, \beta_{\phi_k})$, $k = 0, 1$ with α_{ϕ_k} and β_{ϕ_k} fixed at small values.

3.1.4. Spatial prediction.—The main purpose of the proposed model is for predicting pollutant concentrations at unobserved spatial locations. Due to the spatially-varying lag structure, this prediction requires additional considerations outside of the typical geostatistical prediction setting. We are interested in obtaining samples from the posterior

predictive distribution (ppd) of the measured concentration at a new spatial location, $Y(\mathbf{s}_0)$, where this ppd is defined as

$$f\{Y(\mathbf{s}_0) \mid y\} = \int \dots \int f\{Y(\mathbf{s}_0) \mid \Theta, y\} f(\Theta \mid y) d\Theta$$

where Θ represents the collection of all previously described model parameters, $\boldsymbol{y}$ is the complete vector of measured concentrations and $f(\Theta|\boldsymbol{y})$ is the joint posterior distribution of all model parameters.

There are two different scenarios for predicting at a new spatial location and each results in a different predictive density. First, the desired prediction location may be located within a CMAQ grid cell where a measured concentration is already located, say B_i. In this case, $f\{Y(\mathbf{s}_0)|\Theta, \boldsymbol{y}\}$ can be written as

$$\iint f\left\{Y(\mathbf{s}_0) \mid \sigma_\epsilon^2, \tilde{\beta}_0(\mathbf{s}_0), \tilde{\beta}_1(\mathbf{s}_0), \mu, \alpha_{B_i}\right\} \times \prod_{k=0}^{1} f\{w_k(\mathbf{s}_0) \mid \mathbf{w}_k, \phi_k\} dw_k(\mathbf{s}_0)$$

due to conditional independence, where each component represents a Gaussian distribution based on the likelihood in (3.1) and conditional properties of a multivariate Gaussian distribution (Banerjee, Carlin and Gelfand (2015)).

Second, the prediction location may fall within a previously unobserved CMAQ grid cell, say B_0. In this case, $f\{Y(\mathbf{s}_0)|\Theta, \boldsymbol{y}\}$ can be written as

$$\iiint f\left\{Y(\mathbf{s}_0) \mid \sigma_\epsilon^2, \tilde{\beta}_0(\mathbf{s}_0), \tilde{\beta}_1(\mathbf{s}_0), \mu, \alpha_{B_0}\right\} \times f\left(\alpha_{B_0} \mid \boldsymbol{\alpha}, \tau^2\right) \times \left[\prod_{k=0}^{1} f\{w_k(\mathbf{s}_0) \mid w_k, \phi_k\} dw_k(\mathbf{s}_0)\right] d\alpha_{B_0},$$

where, again, each component represents a Gaussian distribution but now prediction of α_{B_0} is also required. In either case we can collect samples from the ppd of interest using composition sampling based on the samples collected from the joint posterior distribution (Tanner (1996)).

3.2. Distributed lag data fusion: Spatiotemporal downscaler.

Next, we extend the spatial version of the data fusion model to the spatiotemporal setting by allowing for a spatiotemporally-varying lag structure and regression parameters. This model allows for a unified framework of prediction across space and time and the potential to not only predict at new spatial locations but at future time points as well. Similar to (3.1), the spatiotemporal model is given as

$$Y_t(\mathbf{s}_{ij}) = \tilde{\beta}_{0t}(\mathbf{s}_{ij}) + \tilde{\beta}_{1t}(\mathbf{s}_{ij}) \sum_{l=0}^{L} \overline{X}_{B_i,t,l}\left(\frac{\pi_{B_i,t,l}}{\sum_{k=0}^{L} \pi_{B_i,t,k}}\right) + \epsilon_t(\mathbf{s}_{ij}) \tag{3.5}$$

for $t = 1, \ldots, d$ days of measured concentrations where $\epsilon_t(\mathbf{s}_{ij}) \mid \sigma_\epsilon^2 \overset{\text{iid}}{\sim} \text{N}(0, \sigma_\epsilon^2)$. The average of the individual CMAQ estimates comprising spatial lag l surrounding CMAQ grid cell B_i on day t is denoted as $\overline{X}_{B_i,t,l}$. Every AQS monitor is not active on each day of data collection and therefore, every CMAQ grid cell is not represented on each day. The full set of unique CMAQ grid cells that contain an active AQS monitor during any day of the study is denoted as $D = \{B_1, \ldots, B_m\}$. The subset of these grid cells that contain an active AQS monitor during day t is defined as $C_t \subseteq D$ where $m_t^* = |C_t|$ is the number of CMAQ grid cells that contain an active AQS monitor during day t and $\cup_{t=1}^{d} C_t = D$.

The full set of unique active AQS monitoring locations within CMAQ grid cell B_i across all time periods is given as $F_i = \{\mathbf{s}_{i1}, \ldots, \mathbf{s}_{in_i}\}$ where the total number of these locations is given as $n = \sum_{i=1}^{m} n_i$. The subset of unique active AQS monitoring locations within B_i during time period t is given as $E_{ti} \subseteq F_i$ where $n_{ti}^* = |E_{ti}|$ is the number of these locations in B_i during time period t and $\cup_{t=1}^{d} E_{ti} = F_i$. Therefore, the total number of measured concentrations on day t is given as $n_t^* = \sum_{i \in C_t} n_{ti}^*$ and the total number of measured concentrations in the entire dataset is described by $n^* = \sum_{t=1}^{d} n_t^*$. The spatiotemporally-varying weight and regression parameters are described in Sections 3.2.1 and 3.2.2, respectively.

3.2.1. Spatiotemporally-varying regression weights.—The regression weights are now extended to accommodate additive changes across space and time such that

$$\pi_{B_i,t,l} = \Phi\left(\mu + \alpha_{B_i} + \mu_t\right)^l,\ 0 = 1, \ldots, L$$

and

$$\pi_{B_i,t,l} = \left\{1.00 - 1.50\left(\frac{l}{\exp\left\{\mu + \alpha_{B_i} + \mu_t\right\}}\right) + 0.50\left(\frac{l}{\exp\left\{\mu + \alpha_{B_i} + \mu_t\right\}}\right)^3\right\}$$

when $l < \exp\left\{\mu + \alpha_{B_i} + \mu_t\right\}$ and $\pi_{B_i,t,l} = 0$ when $l \geq \exp\left\{\mu + \alpha_{B_i} + \mu_t\right\}$, where μ is the global spatial lag structure parameter and α_{B_i} is described in (3.3). The neighborhood adjacency matrix for the α_{B_i} parameters is defined based on the complete set of unique CMAQ grid cells that contain an active AQS monitor during any day of the study (i.e., D). The μ_t parameters are modeled using an autoregressive structure to allow for the lagged weights to change across time such that

$$\mu_t = \kappa\mu_{t-1} + \delta_t,\ t = 1, \ldots, d,$$

where $\kappa \in (0, 1)$, $\delta_t \mid \sigma_\delta^2 \overset{\text{iid}}{\sim} \text{N}(0, \sigma_\delta^2)$, and $\mu_0 \equiv 0$.

3.2.2. Spatiotemporally-varying regression parameters.—Similarly, the regression parameters are allowed to evolve across space and time such that

$$\tilde{\beta}_{kt}(\mathbf{s}_{ij}) = \beta_k + \beta_k(\mathbf{s}_{ij}) + \beta_{kt},\ k = 0, 1,$$

where β_k are the global intercept/slope parameters, $\beta_k(\mathbf{s}_{ij})$ are the spatial deviations from these global parameters, previously described by (3.4) and β_{kt} represent temporal deviations from the global parameters. They are modeled using a multivariate autoregressive framework such that

$$\begin{pmatrix}\beta_{0t}\\ \beta_{1t}\end{pmatrix} = \Omega\begin{pmatrix}\beta_{0,t-1}\\ \beta_{1,t-1}\end{pmatrix} + \eta_t,\ t = 1, \ldots, d. \tag{3.6}$$

where Ω is a two-by-two diagonal matrix with $\Omega_{ii} = \rho_i$ and $\rho_i \in (0, 1)$; $\eta_t \mid V \overset{\text{iid}}{\sim} \text{MVN}(\mathbf{0}_2, V)$, $\mathbf{0}_2$ is a column vector of length two with each entry equal to zero and V a two-by-two covariance matrix describing potential correlation between the intercept and slope deviations, and $(\beta_{00}, \beta_{10})^{\text{T}} = \mathbf{0}_2$.

3.2.3. Prior distributions.—For the parameters shared across the spatial and spatiotemporal models, the prior specifications remain the same (see Section 3.1.3 for full details). For the parameters specific to the spatiotemporal model, we opt for weakly informative prior distributions when possible such that $\kappa, \rho_1, \rho_2 \overset{\text{iid}}{\sim} \text{Uniform}(0, 1)$, $\sigma_\delta^2 \sim \text{Inverse Gamma}(3, 2)$ and $V^{-1} \sim \text{Wishart}(I_2, 3)$ (Gelman et al. (2014)).

3.2.4. Spatiotemporal prediction.—Prediction within the spatiotemporal modeling framework can also fall under a few different categories: prediction at a previously unobserved spatial location on a previously observed day or prediction at a previously unobserved spatial location during a day in the future. We assume here that the spatial location for prediction falls within one of the previously observed CMAQ grid cells, say $B_0 \in D$, but this can easily be extended; see Section 3.1.4 for further discussion on this point.

The posterior predictive distribution of interest is given as

$$f\{Y_{t_0}(\mathbf{s}_0) \mid Y\} = \int \cdots \int f\{Y_{t_0}(\mathbf{s}_0) \mid \Theta, Y\} f(\Theta \mid Y) d\Theta,$$

where $Y_{t_0}(\mathbf{S}_0)$ is the pollutant concentration of interest at unobserved spatial location $\mathbf{s}_0$ and potentially unobserved day t_0. The remaining terms have been previously described in Section 3.1.4. When predicting at a previously observed day such that $t_0 \in \{1, \ldots, d\}$, $f\{Y_{t_0}(\mathbf{s}_0) \mid \Theta, Y\}$ can be written as

$$\iint f\left\{Y_{t_0}(\mathbf{s}_0) \mid \sigma_\epsilon^2, \tilde{\beta}_{0t_0}(\mathbf{s}_0), \tilde{\beta}_{1t_0}(\mathbf{s}_0), \mu, \alpha_{B_0}, \mu_{t_0}\right\} \times \prod_{k=0}^{1} f\{w_k(\mathbf{s}_0) \mid w_k, \phi_k\} dw_k(\mathbf{s}_0). \tag{3.7}$$

Similarly, when predicting for a new day, say $t_0 = d+1$, $f\{Y_{d+1}(\mathbf{s}_0)|\Theta,\boldsymbol{Y}\}$ maintains the same form from (3.7) with the following additional terms included in the multiplication:

$$\iiint f(\boldsymbol{\beta}_{d+1} \mid \rho_1, \rho_2, \boldsymbol{\beta}_d, V) \times f\left(\mu_{d+1} \mid \kappa, \mu_d, \sigma_\delta^2\right) d\boldsymbol{\beta}_{d+1} d\mu_{d+1}.$$

Similar to the spatial prediction solution, we can generate samples from the ppd using composition sampling, given that these densities are all Gaussian (or multivariate Gaussian) with known mean and variance/covariance.

4. Daily ozone and $PM_{2.5}$ data fusion.

4.1. Creating lagged predictors.

To implement the newly developed methods, we begin by linking each AQS monitor that was active during the study period (June 1–August 31, 2013, see Figure 2) with the corresponding CMAQ grid cell in which it was located (B_i). This CMAQ estimate represents the lag-zero predictor from (3.1) $\left(\bar{x}_{B_i,0}\right)$. Next, we average the CMAQ estimates from the grid cells that immediately surround B_i and define this to be the lag-one predictor $\left(\bar{x}_{B_i,1}\right)$.

We repeat this process, each time averaging the CMAQ grid cell estimates contained in the square surrounding the previous lagged average grid cell estimates, until we the obtain the lag-10 predictor ($\bar{x}_{B_i,10}$; i.e., $L = 10$ in (3.1)); see Figure 1 for an example.

4.2. *Spatial downscaler*: *June–August*, 2013.

We first apply the distributed lag data fusion model in (3.1), hereafter referred to as DLfuse, to three different days of data: June 3, July 3 and August 2, 2013. These days were the first day in each month where the number of active $PM_{2.5}$ monitors was high in comparison to surrounding days (see Figure 2), allowing for a large number of concentrations with which to validate the model. We randomly select 64, 128 and 192 of the monitors for three separate analyses on each day. The random selection of monitors is done separately for each day, pollutant and sample size. The remaining nonselected AQS locations are used to validate the model predictions. We select these sample sizes for analysis because 192 represents the median number of active $PM_{2.5}$ monitors across this spatial domain on the "low monitoring" activity days across the entire month of June 2013. Additionally, we select one- and two-thirds of 192 (64, 128) and apply the model to see how predictive performance changes when significantly less data are used. In Table S1 of the Supplementary Material, we summarize the number of uniquely observed CMAQ grid cells (i.e., those that contain an active AQS monitor) with respect to the total number of AQS observations in each dataset (Warren et al. (2021)).

4.2.1. Competing methods.—We consider several competing methods in the study. These methods vary greatly in complexity and include: (i) raw CMAQ model estimates, (ii) ordinary kriging (OK), (iii) the original static downscaler (DS) from Berrocal, Gelfand and Holland (2010a), (iv) the spatially-varying random weights smoothed downscaler (DS-Smooth) from Berrocal, Gelfand and Holland (2012) and (v) the spectral spatial downscaling method (DS-Spectral) of Reich, Chang and Foley (2014).

The raw CMAQ approach simply uses the lag zero CMAQ estimate $\left(\bar{x}_{B_i,0}\right)$ as the prediction of the AQS concentration at a new spatial location residing within CMAQ grid cell B_i. This method will perform well if CMAQ is unbiased for the measured concentrations and if there is no variability in the measured concentrations within a CMAQ grid cell. Next, we consider the geostatistical prediction approach of OK which is a special case of (3.1). OK does not include any CMAQ estimates in the regression model but does incorporate a spatially structured intercept term such that

$$Y(\mathbf{s}_{ij}) = \beta_0 + A_{11} w_0(\mathbf{s}_{ij}) + \epsilon(\mathrm{s}_{ij}), \tag{4.1}$$

where each of the terms have been previously described. OK will predict well if the measured concentrations are spatially correlated across the domain and the correlation can be accurately described by the specified isotropic correlation model.

DS is also a special case of (3.1) when the lag structure is ignored (i.e., $L = 0$) and is given as

$$Y(\mathbf{s}_{ij}) = \tilde{\boldsymbol{\beta}}_0(\mathbf{s}_{ij}) + \tilde{\boldsymbol{\beta}}_1(\mathbf{s}_{ij})\bar{X}_{B_i,0} + \epsilon(\mathbf{s}_{ij}), \tag{4.2}$$

where each of the terms have been previously described. DS allows for spatially varying bias adjustment through the intercepts and slopes but does not consider lagged CMAQ predictors.

For full details on DS-Smooth, see Section 3.1.3 of Berrocal, Gelfand and Holland (2012). Unlike the previously described competing methods, DS-Smooth is not a special case of DLfuse, but its goals are similar. DS-Smooth attempts to utilize information from more than just a single CMAQ grid cell when modeling/predicting the pollutant concentrations. Instead of considering lagged CMAQ estimates, however, DS-Smooth includes a weighted average of all CMAQ grid cell estimates within the regression model where the weights are AQS location specific, defined by an exponential kernel function and a spatially-structured Gaussian process and sum to one. DS-Smooth also includes spatially-correlated intercept terms, while the slope is constant due to identifiability issues. The weighted CMAQ average, defined by (3.1), can be written as

$$\tilde{X}(\mathrm{s}_{ij}) = \tilde{X}(B_i) = \sum_{l=0}^{L} \bar{X}_{B_i,l} \left(\frac{\pi_{B_i,l}}{\sum_{k=0}^{L} \pi_{B_i,k}} \right)$$

and only depends on two parameters, μ and α_{B_i}. Therefore, a new α_{B_i} parameter is only introduced when it appears in the data (i.e., when an AQS monitor is located in CMAQ grid cell B_i). The weighted CMAQ average of DS-Smooth is defined as

$$\widetilde{X}(\mathbf{s}_{ij}) = \sum_{k=1}^{g} \mathrm{x}(B_k)\eta_k(\mathbf{s}_{ij}),$$

where $\mathrm{x}(B_k)$ is the CMAQ estimate from grid cell B_k and g represents the total number of CMAQ grid cells across the entire spatial domain. Each weight parameter, $\eta_k(\mathbf{s}_{ij})$, is defined by one shared parameter and a unique spatial parameter that varies by CMAQ grid cell. Therefore, for a single AQS observation the full set of g spatial parameters is needed to define this weighted average. For reference, in the spatial domain considered here we have g = 31,722 total CMAQ grid cells while the number of grid cells that contain an active AQS monitor is significantly smaller for both pollutants (see Figure 2). The introduction of such a large number of individual CMAQ predictors and accompanying regression weights results in a model which is computationally difficult to fit in this setting. As a result, Berrocal, Gelfand and Holland (2012) implemented the predictive process modeling idea of Banerjee et al. (2008) to reduce the computational burden, and we do the same in this application. While the weights defined by Berrocal, Gelfand and Holland (2012) are more flexible than those defined by DLfuse, it is important to note that use of the spatially-varying intercepts and slopes in DLfuse helps to increase modeling and prediction flexibility overall.

Finally, we consider DS-Spectral which is also not a special case of DLfuse but, instead, uses a computationally efficient spectral method to model complex associations between measured concentrations and CMAQ estimates. Using code from Reich, Chang and Foley (2014), we calculate K = 20 spectral covariates at each CMAQ grid cell ($\widetilde{X}_i$), assign them to the measured observations that are located in each grid cell and enter them as predictors into a regression model with spatially-varying intercepts such that

$$Y(\mathbf{s}_{ij}) = \beta_0 + A_{11}w_0(\mathbf{s}_{ij}) + \sum_{j=1}^{20} \theta_j \widetilde{x}_i + \epsilon(\mathbf{s}_{ij}),$$

where $\theta_j \mid \mu_\theta, \sigma^2_\theta \overset{\text{iid}}{\sim} \mathrm{N}(\mu_\theta, \sigma^2_\theta)$, $j = 1, \ldots, 20$ and the other terms have been previously described.

For methods OK and DS, we use the prior distributions, as specified in Section 3.1.3, given that these methods are subsets of DLfuse. Specifically, $\alpha_{\sigma^2_\epsilon} = \beta_{\sigma^2_\epsilon} = 0.01$, $\alpha_{\tau^2} = 3$, $\beta_{\tau^2} = 2$, $\sigma^2_{\beta} = 100^2$, $\sigma^2_A = 1$ and $\alpha_{\phi_k} = 1$, $\beta_{\phi_k} = 5$ for $k = 0, 1$. For DS-Smooth, we use the prior specifications, as described in Section 4.1 of Berrocal, Gelfand and Holland (2012), with minor adjustments to the choice of prior distribution hyperparameters. For the predictive process approximation we randomly selected 400 knot locations from the full set of CMAQ centroids that contained an AQS monitor on the selected day of analysis. This random selection was carried out separately for each pollutant and day. Similar to Section A.1 of Berrocal, Gelfand and Holland (2012), we also use a Metropolis block-updating algorithm

for the spatially-structured Gaussian process parameters with a chosen block size of 50. For DS-Spectral, $\mu_\theta \sim N(0, 100^2)$, σ^2_θ~Inverse Gamma $(0.01,\ 0.01)$ and A^2_{11}~ Inverse Gamma$(3, 2)$.

For all models with spatially-correlated intercepts and/or slope parameters, we implement a sum-to-zero constraint *on the fly* (Berrocal, Gelfand and Holland (2012), Besag et al. (1995)), separately for each set of parameters, to improve identifiability (other than for DS-Spectral where the spatial random effects were marginalized out prior to model fitting). For similar reasons and for the spatial parameters corresponding to the lag weights (see (3.3)), we impose the constraint that the parameters sum to zero and have unit variance and enforce it *on the fly*.

4.2.2. Model comparison metrics.—We apply each competing method to each of the subsampled datasets for both pollutants across all days and predict at the validation locations based on results in Section 3.1.4. We log-transform the AQS concentrations prior to model fitting to achieve approximate normality. In order to compare the predictive results between the competing methods, we calculate the predictive mean absolute error (PMAE), average empirical coverage of the 95% equal tailed, quantile-based credible intervals (CIs) and average length of those CIs. Posterior medians are used as point estimates for the predictions. Runtimes for each method are also recorded.

In addition to predictive accuracy, we also calculate Watanabe–Akaike information criteria (WAIC) (Watanabe (2010)) for each fitted model. In practice, a user may not have access to a hold-out sample of validation data that can be used to compare predictive performances. However, if a model comparison metric like WAIC, which balances model fit and complexity (p_{WAIC}), can be used to identify the "best" model among the competitors, and this choice aligns with the method that also has the optimal predictive performance, then it may be useful in practice for determining which method to use for prediction. Given the lack of statistical modeling, we note that WAIC cannot be calculated for the raw CMAQ method. To make the comparisons more fair across the various methods which have different regression forms, we calculate WAIC based on a marginalized version of the likelihood of the data for each method. We integrate over all of the spatial "random effect" parameters (in the intercept and slopes, not the lagged predictors) so that each method's likelihood has a fixed-effect regression trend and spatially correlated error terms, if present.

4.2.3. Results.—From each model we collect 450,000 posterior samples after removing the first 50,000 as a burn-in period. We thin the remaining samples by a factor of 45 to reduce posterior autocorrelation and the dimension of our results, leading to 10,000 posterior samples with which to make posterior inference. Convergence was assessed using individual parameter trace plots and by calculating the Geweke diagnostic (Geweke (1992)) for each relevant parameter across the different models, with no obvious signs of nonconvergence being observed. The runtimes from the June 3rd analyses are shown in Table S2 of the Supplementary Material with OK and DS-Smooth consistently reporting the shortest and longest runtimes, respectively (Warren et al. (2021)).

We first describe the $PM_{2.5}$ results shown in Table 1. The WAIC findings suggest that DLfuse and DS-Smooth provide improved fits across each dataset, even when accounting for

the additional complexity of the methods. Their WAIC values are often similar to each other while typically being smaller than the other methods. These findings are encouraging but do not necessarily suggest that these models are providing improved predictions at the validation sites. However, the $PM_{2.5}$ prediction results often align with the WAIC findings, suggesting that WAIC may be a useful tool for determining which model to use for prediction in the absence of validation data.

While DLfuse and DS-Smooth have similar WAIC values, DLfuse typically produces improved predictions. Out of the nine total $PM_{2.5}$ analyses, some version of DLfuse produces the smallest PMAE value six times, with no other method producing the smallest value more than once. In head-to-head comparisons with DS-Smooth and DS-Spectral, DLfuse produces smaller PMAE values seven out of nine times each. All methods generally provide CI coverage near 95% with DS-Smooth and DS-Spectral consistently providing shorter intervals. The DLfuse results based on the weights defined by the CDF of the standard normal distribution are generally superior to those based on the spherical spatial correlation function.

In the first column of Figure 3, we present the posterior predictive means at 1000 randomly selected CMAQ grid cells (with inverse distance weighting used to interpolate between those 1000 grid cells for presentation purposes) of the spatially varying lag structure parameters $\left(\mu + \alpha_{B_i}\right)$, based on both versions of DLfuse being fitted to the 192 sample size $PM_{2.5}$ dataset from June 3. The plots corresponding to July 3 and August 2 are shown in Figures S2 and S3 of the Supplementary Material, respectively (Warren et al. (2021)). The positively estimated lag component suggests that the lag structure may be important to consider when predicting $PM_{2.5}$ concentrations—something clearly supported by Table 1. Recall from (3.1) that if $\mu + \alpha_{B_i}$ is large and positive, both versions of the lag weights for $l > 0$ will be nonzero, indicating that surrounding CMAQ information may be useful for improved prediction. Figure 3 and Figures S2–S3 of the Supplementary Material also show that the spatial pattern of the lag structure is similar across both weight definitions (Warren et al. (2021)).

The ozone results in Table 2 are more variable across analyses and methods, with DS performing better than it did in the $PM_{2.5}$ analyses. We see that DL-Smooth and DLfuse most often produce smaller WAIC values, as in the $PM_{2.5}$ analyses. The PMAE results are mixed, with DLfuse producing the smallest value in four out of the nine total analyses and DS, DS-Spectral producing the smallest values twice each (only once for DS-Smooth). In head-to-head comparisons with DS-Smooth and DS-Spectral, DLfuse once again produces smaller PMAE values seven out of nine times each. Each method performs well with respect to coverage of the CIs with no consistent trend for the average CI lengths. The results suggest that, in this case, the included lag structure is less helpful in predicting measured concentrations and, as a result, the DLfuse results closely resemble those from DS in many analyses, with other methods also performing well. Posterior predictive means of the spatially-varying lag parameters displayed in the second columns of Figure 3 and Figures S2–S3 of the Supplementary Material provide evidence to further support these findings (Warren et al. (2021)). The negative estimates for the lag parameters indicate that the lag structure is less important for ozone. Given the chemistry of ozone formation, which tends to

occur over larger geographic areas, compared to that of $PM_{2.5}$ where levels tend to fall off with distance from the emitting source, this result is not surprising.

The third columns of Figure 3 and Figures S2–S3 of the Supplementary Material allows us to better understand the estimated lag structure (Warren et al. (2021)). For both versions of DLfuse fit to the 192 sample size $PM_{2.5}$ and ozone datasets, we display the posterior means of (3.2) across all lags, where each plotted line represents a uniquely observed CMAQ grid cell. Regardless of which weight definition is used, the $PM_{2.5}$ results suggest a slow decrease in the weights as the lag degree increases, further indicating that the average lagged CMAQ estimates are useful in predicting pollutant concentrations. The ozone results suggest that the lagged CMAQ estimates are less useful when predicting, as the regression weights decrease much more quickly. The figure also suggests that there is not a lot of spatial variability in the lag structure for either pollutant. These features of the lags are generally consistent across all days of analysis. Overall, the results for both pollutants remain fairly consistent across all considered sample sizes and days.

4.3. *Spatiotemporal downscaler*: *June–August* 2013.

Next, we apply the spatiotemporal version of the newly developed model, hereafter referred to DLfuseST, to three datasets of different lengths: June 1–June 30, June 1–July 31 and June 1–August 31, 2013. For each analysis we remove the measured concentrations from the final day and use it as validation data. For computational purposes we first select 128 AQS monitors that were active at some point during the study period (separately for each pollutant and dataset) and use these as the measured concentrations. Therefore, on each day of the study we use measured concentrations from the subset of those 128 selected monitors which were active to fit the model. Using results from Section 3.2.4, we then predict at all active monitor locations on the final day of each dataset.

4.3.1. Competing methods.—Similar to the spatial data application, we consider a number of competing spatiotemporal methods. These methods represent special cases of (3.5) and include spatiotemporal versions of the competing models previously described in Section 4.2.1, though we do not include the spatiotemporal versions of DS-Smooth and DS-Spectral in the comparisons. The raw CMAQ procedure has already been described. For the spatiotemporal version of DS (DS-ST), we now include temporally-varying regression parameters such that $\tilde{\beta}_k(\mathbf{s}_{ij})$, $k = 0, 1$, in (4.2) is replaced by $\beta_k+\beta_k(\mathbf{s}_{ij})+\beta_{kt}$ for DS-ST. The temporally-varying intercepts and slopes are described in (3.6). Similarly, for the spatiotemporal version of OK (OK-ST), we include a temporally-varying intercept parameter such that β_0 in (4.1) is replaced by $\beta_0 + \beta_{0t}$ and β_{0t} follows a univariate version of the autoregressive model detailed in (3.6). Each of the prior distributions have been previously described in Sections 3.1.3, 3.2.3 and 4.2.1. We implement a sum-to-zero constraint *on the fly* for the β_{0t}, β_{1t} and μ_t parameters separately and further impose that the μ_t parameters have unit variance.

4.3.2. Results.—All methods were applied to the same datasets (log-transformed measured concentrations) where we collected 100,000 posterior samples after removing the first 10,000 as a burn-in period. We thinned the samples by a factor of 10, resulting in

10,000 posterior samples used for posterior inference. There were no obvious signs of nonconvergence based on trace plots and the Geweke diagnostic analysis.

The results shown in Table 3 suggest that one of the versions of DLfuseST always yields improved predictions and shorter CIs across both pollutants and all temporal periods. Similar to the spatial results, larger differences between competing methods are often seen for the $PM_{2.5}$ applications, indicating that the lagged CMAQ predictors may play a larger role. To better understand the changes in the lag parameters across time, Figure 4 and Figures S4–S5 of the Supplementary Material display the posterior means of the temporal components of the lag structures (μ_l) from both versions of DLfuseST, plotted across the days of the analyses, with more variability (and less smoothness across time) generally seen in the ozone results and similar behavior observed across both weight definitions (Warren et al. (2021)). The higher variability in the temporal component of the lag structure for ozone makes sense given the higher dependence of ozone formation on temperature.

5. Discussion.

We introduced distributed lag data fusion methods for the analysis of spatial (DLfuse) and spatiotemporal (DLfuseST) ambient air pollution data that resulted in improved predictive performances over existing state-of-the-art approaches in the majority of analyses. Importantly, other than DS-Smooth (Berrocal, Gelfand and Holland (2012)) and DS-Spectral (Reich, Chang and Foley (2014)), these existing approaches were shown to be special cases of the more general method developed here, including the original downscaler from Berrocal, Gelfand and Holland (2010a). This indicates that this new methodology should be used in most applications, given that it can collapse to resemble the competing methods when needed. In fact, in the spatial ozone data application we saw evidence of this behavior, while for $PM_{2.5}$, use of the lagged CMAQ predictors resulted in improved predictive performance. Additionally, we provided evidence to suggest that WAIC may be a useful tool in the spatial prediction setting to determine which method will produce more accurate predictions. In addition to WAIC, future users of DLfuse and DLfuseST can analyze the spatially-varying estimated weights (see Figure 3 and Figures S2–S3 of the Supplementary Material) to determine if there are locations that are benefiting from the inclusion of lagged CMAQ information (Warren et al. (2021)). If there are not, however, DLfuse will likely perform similarly to DS, given the connection between both methods.

In comparison to DS-Smooth, DLfuse represents a more parsimonious and less computationally intensive model. While both aim to incorporate CMAQ information from surrounding spatial locations, DS-Smooth introduces a unique parameter for every CMAQ grid cell in the spatial domain while DLfuse only introduces a parameter for those grid cells that contain an active AQS monitor. In Figures S6–S7 of the Supplementary Material, we present trace plots from randomly selected spatial locations of the weighted CMAQ covariate $\left(\sum_{l=0}^{L} \overline{x}_{B_i,l}\left(\frac{\pi_{B_i,l}}{\sum_{k=0}^{L}\pi_{B_i,k}}\right)\right)$ and the location-specific slope ($\tilde{\beta}_k(\mathbf{s})$) from DLfuse (weights defined by the CDF of the standard normal distribution) applied to the June 3 datasets for both pollutants. Similar plots corresponding to DLfuseST are shown in Figures

S8–S9 of the Supplementary Material (Warren et al. (2021)). The results suggest that there are no major identifiability concerns caused by modeling spatial variation in the weights and slopes. This is likely due to a few factors. First, when there are multiple active AQS monitors located in the same CMAQ grid cell, this results in replications of the α_{B_i} parameter across multiple observations. This is because two locations in the same CMAQ grid cell have the exact same set of weights $\left(\pi_{B_i,l}/\sum_{k=0}^{L}\pi_{B_i,k}\right)$ (see Table S1 of the Supplementary Material for how often this happens in our case studies) (Warren et al. (2021)). Second, a single α_{B_i} parameter is used to define the weighted CMAQ covariate for a selected spatial location. Also, because each of the weights are bounded between zero and one and sum to one, the slope and weight parameters are better identified in this setting. DS-Spectral offers great computational improvements over DLfuse and DS-Smooth, since it can be fit in the mixed model framework and it predicts the validation data well with respect to many of the methods. However, it is consistently outperformed by DLfuse with respect to predictive accuracy in our specific applications, particularly for the $PM_{2.5}$ analyses where surrounding auxiliary information may be more beneficial.

Environmental health disparities are typically conceived as problems of differences in exposure or differences in effect. Another important dimension, however, is differences in access to information on those exposures and, therefore, on the expressed toxic effects. Our method provides improved estimation of pollutant levels across both space and time, including for those areas where pollutant levels are sparsely measured, if at all. We focused on the eastern U.S. to be comparable with past work in this area but note that extensions to include the western U.S. will require additional considerations, given the relatively poor spatial coverage of the air pollution monitors as well as differing atmospheric conditions and transport phenomena.

In addition to being run for retrospective studies, the CMAQ model can also be used for future time periods under different projections of what future meteorological conditions may look like. For example, recent climate change studies have used CMAQ output to describe air pollution levels under different future scenarios (e.g., Nolte et al. (2018)). Therefore, it is possible that downscaling methods that can predict future values (potentailly on aggregated time scales) could be useful in this context. Based on our experiences, future studies that plan to utilize DLfuseST will likely benefit from focusing on a smaller spatial domain and including more time periods, a larger spatial domain with fewer time periods or a larger spatial domain with a fixed set of AQS monitors that does not change over time (as we have done in our study) and more time periods. Computationally, DLfuseST becomes slower as the number of unique AQS locations across all time periods of data increases, and we recommend its use when prediction of future pollution concentrations is of interest. When spatial prediction on a previously observed day is the primary goal of an analysis, we recommend DLfuse which can be implemented at a reduced computational cost. Users of DLfuseST should also think carefully about the assumptions being made for the temporal components of the model. In particular, we opt for a rather simple additive space-time specification for many of the parameters in order to ease the computational burden associated with modeling potentially large datasets. While this form does appear to perform

well in our specific case studies, particularly with respect to the raw CMAQ baseline approach, it may be restrictive for applications that anticipate an interaction between space and time or some other complex relationship between the two.

The introduced framework can also be extended to include alternative estimates of pollutant concentrations, such as satellite data. In the case of multiple, possibly correlated auxiliary information, new distributed lag methodology may be needed to account for joint modeling and interactions. Another future extension of the framework would be to allow the CMAQ estimates corresponding to a specific lag to be differentially weighted before being averaged and used as a predictor in the model. Currently for DLfuse, all CMAQ estimates from the same lag receive the same weight, whereas Berrocal, Gelfand and Holland (2012) overcome this by allowing for a unique parameter at every CMAQ grid cell. This feature could be important to consider for pollutants where wind speed and direction could impact measured concentrations. New methodology may be required to incorporate this into the distributed lag setting. However, with increased flexibility comes the inclusion of potentially many additional parameters; so future work should balance this effort with practical computing considerations. While the focus of this work is on spatial distributed lag models, extensions to include temporal lagged pollution information may also be useful in producing accurate predictions of pollution concentrations at new spatial locations and time periods. Including both spatial and temporal lags (and their possible interaction) would require a careful consideration of correlation patterns across parameters with an emphasis on computational aspects of the modeling.

In practice, a potential drawback of DLfuse is that it does require the creation of the lagged covariates not only at the location of the AQS monitors but also at locations where predictions are to be made. This covariate construction process requires more effort than simply using the CMAQ estimate from the grid cell containing the AQS monitor, but our results suggest that this additional effort may be warranted in terms of prediction accuracy. Code to implement DLfuse and DLfuseST is available at https://github.com/warrenjl/DLfuse and also in Warren et al. (2021).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments.

Supported by 1R01MD012769-01A1 from NIMHD and Assistance Agreement No. RD835871 awarded by the U.S. EPA to Yale University. It has not been formally reviewed by EPA. The views expressed in this document are solely those of the authors and do not necessarily reflect those of the Agency.

REFERENCES

BAEK J, SÁNCHEZ BN, BERROCAL VJ and SANCHEZ-VAZNAUGH EV (2016). Distributed lag models: Examining associations between the built environment and health. Epidemiology 27 116. [PubMed: 26414942]

BANERJEE S, CARLIN BP and GELFAND AE (2015). Hierarchical Modeling and Analysis for Spatial Data, 2nd ed. Monographs on Statistics and Applied Probability 135. CRC Press, Boca Raton, FL. MR3362184

BANERJEE S, GELFAND AE, FINLEY AO 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 10.1111/j.1467-9868.2008.00663.x

BERMAN J, JIN L, BELL M and CURRIERO FC (2019). Developing a geostatistical simulation method to inform the quantity and placement of new monitors for a follow-up air sampling campaign. J. Expo. Sci. Environ. Epidemiol. 29 248. [PubMed: 30237550]

BERROCAL VJ, GELFAND AE and HOLLAND DM (2010a). A spatio-temporal downscaler for output from numerical models. J. Agric. Biol. Environ. Stat. 15 176–197. MR2787270 10.1007/s13253-009-0004-z [PubMed: 21113385]

BERROCAL VJ, GELFAND AE and HOLLAND DM (2010b). A bivariate space-time downscaler under space and time misalignment. Ann. Appl. Stat. 4 1942–1975. MR2829942 10.1214/10-AOAS351 [PubMed: 21853015]

BERROCAL VJ, GELFAND AE and HOLLAND DM (2012). Space-time data fusion under error in computer model output: An application to modeling air quality. Biometrics 68 837–848. MR3055188 10.1111/j.1541-0420.2011.01725.x [PubMed: 22211949]

BERROCAL VJ, GUAN Y, MUYSKENS A, WANG H, REICH BJ, MULHOLLAND JA and CHANG HH (2019). A comparison of statistical and machine learning methods for creating national daily maps of ambient $PM_{2.5}$ concentration. Atmos. Environ.. 10.1016/j.atmosenv.2019.117130

BESAG J (1974). Spatial interaction and the statistical analysis of lattice systems. J. Roy. Statist. Soc. Ser. B 36 192–236. MR0373208

BESAG J, GREEN P, HIGDON D and MENGERSEN K (1995). Bayesian computation and stochastic systems. Statist. Sci. 10 3–66. With comments and a reply by the authors. MR1349818

FUENTES M and RAFTERY AE (2005). Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. Biometrics 61 36–45. MR2129199 10.1111/j.0006-341X.2005.030821.x [PubMed: 15737076]

GELMAN A, CARLIN JB, STERN HS, DUNSON DB, VEHTARI A and RUBIN DB (2014). Bayesian Data Analysis, 3rd ed. Texts in Statistical Science Series. CRC Press, Boca Raton, FL. MR3235677

GEWEKE J (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In Bayesian Statistics, 4 (Peñíscola, 1991) 169–193. Oxford Univ. Press, New York. MR1380276

GUAN Y, REICH BJ, MULHOLLAND JA and CHANG HH (2019). Multivariate spectral downscaling for $PM_{2.5}$ species. arXiv preprint arXiv:1909.03816.

GURUNG A, LEVY JI and BELL ML (2017). Modeling the intraurban variation in nitrogen dioxide in urban areas in Kathmandu Valley, Nepal. Environ. Res. 155 42–48. [PubMed: 28189072]

HSU C-Y, WU J-Y, CHEN Y-C, CHEN N-T, CHEN M-J, PAN W-C, LUNG S-CC, GUO YL and WU C-D (2019). Asian culturally specific predictors in a large-scale land use regression model to predict spatial-temporal variability of ozone concentration. Int. J. Environ. Res. Public Health 16 1300.

JIN L, BERMAN JD, WARREN JL, LEVY JI, THURSTON G, ZHANG Y, XU X, WANG S, ZHANG Y et al. (2019). A land use regression model of nitrogen dioxide and fine particulate matter in a complex urban core in Lanzhou, China. Environ. Res. 177 108597. [PubMed: 31401375]

MCMILLAN NJ, HOLLAND DM, MORARA M and FENG J (2010). Combining numerical model output and particulate data using Bayesian space-time modeling. Environmetrics 21 48–65. MR2842223 10.1002/env.984

NOLTE CG, SPERO TL, BOWDEN JH, MALLARD MS and DOLWICK PD (2018). The potential effects of climate change on air quality across the conterminous US at 2030 under three representative concentration pathways. Atmos. Chem. Phys. 18 15471. [PubMed: 30972111]

PACIOREK CJ (2012). Combining spatial information sources while accounting for systematic errors in proxies. J. R. Stat. Soc. Ser. C. Appl. Stat. 61 429–451. MR2914520 10.1111/j.1467-9876.2011.01035.x

REICH BJ, CHANG HH and FOLEY KM (2014). A spectral method for spatial downscaling. Biometrics 70 932–942. MR3295754 10.1111/biom.12196 [PubMed: 24965037]

TANNER MA (1996). Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions, 3rd ed. Springer Series in Statistics. Springer, New York. MR1396311 10.1007/978-1-4612-4024-2

US EPA (2019). CMAQ models. https://www.epa.gov/cmaq/cmaq-models-0.

WACKERNAGEL H (2013). Multivariate Geostatistics: An Introduction with Applications. Springer Science & Business Media.

WARREN JL, KONG W, LUBEN TJ and CHANG HH (2020). Critical window variable selection: Estimating the impact of air pollution on very preterm birth. Biostatistics 21 790–806. 10.1093/biostatistics/kxz006 [PubMed: 30958877]

WARREN JL, MIRANDA ML, TOOTOO JL, OSGOOD CE and BELL ML (2021). Supplement to "Spatial distributed lag data fusion for estimating ambient air pollution." 10.1214/20-AOAS1399SUPPA, 10.1214/20-AOAS1399SUPPB

WATANABE S (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. J. Mach. Learn. Res. 11 3571–3594. MR2756194

XU H, BECHLE MJ, WANG M, SZPIRO AA, VEDAL S, BAI Y and MARSHALL JD (2019). National $PM_{2.5}$ and NO_2 exposure models for China based on land use regression, satellite measurements, and universal kriging. Sci. Total Environ. 655 423–433. [PubMed: 30472644]

YU Y, YAO S, DONG H, WANG L, WANG C, JI X, JI M, YAO X and ZHANG Z (2019). Association between short-term exposure to particulate matter air pollution and cause-specific mortality in Changzhou, China. Environ. Res. 170 7–15. [PubMed: 30554054]

FIG. 1.

Example of community multiscale air quality grid cells where the air quality system monitoring location is assumed to lie within the grid cell marked by 0. *Grid cells with a shared number are included in the corresponding lagged average of pollution estimates and used as predictors in* (3.1).

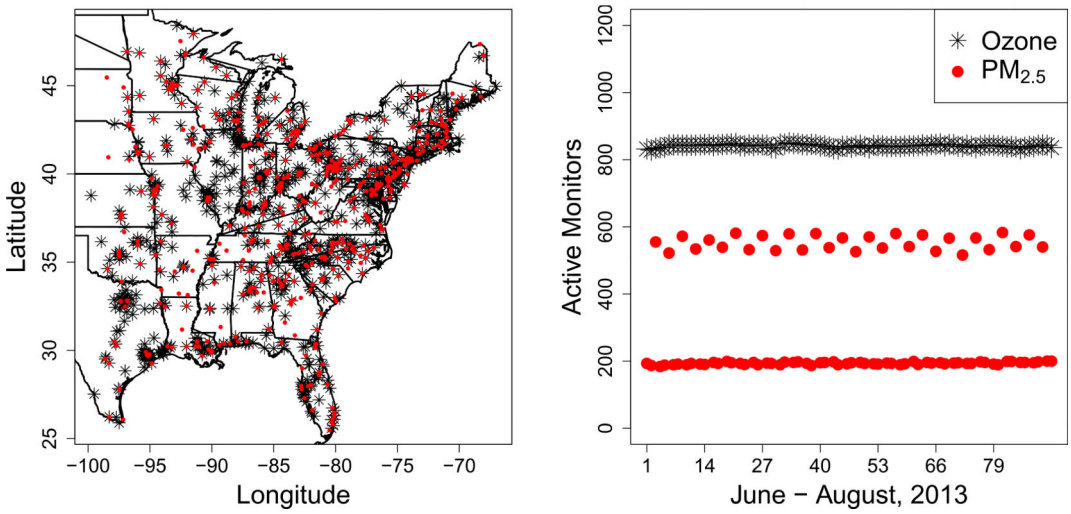

FIG. 2.

Air quality system active monitor locations (*left panel*) *and time series plots of the number of daily active monitors from June* 1–*August* 31, 2013 (*right panel*) *for ozone and* $PM_{2.5}$.

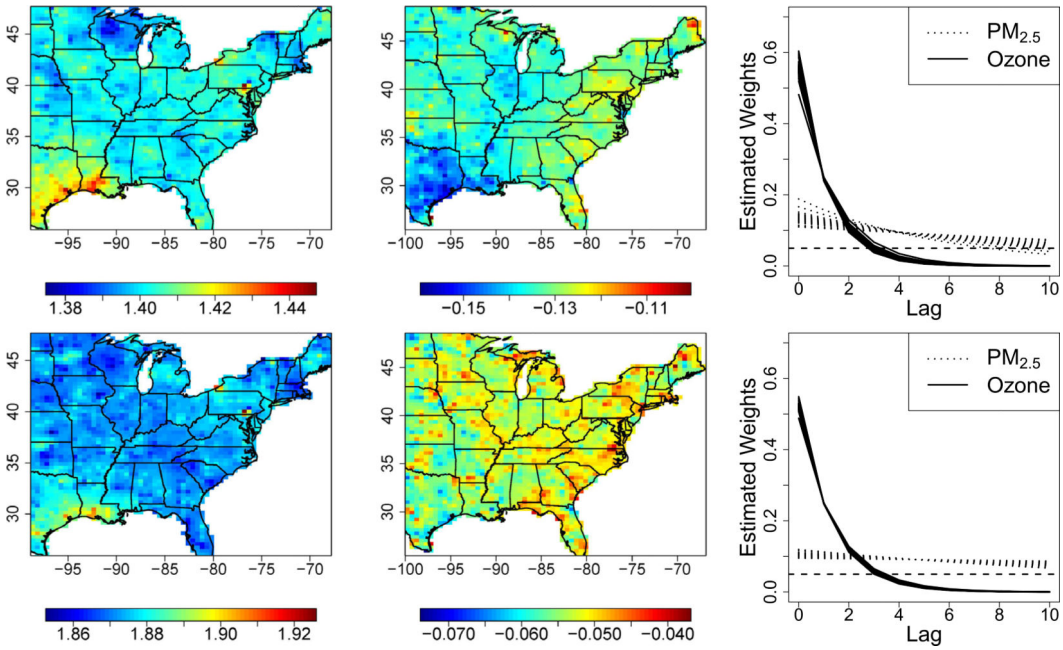

FIG. 3.

Posterior means from the distributed lag data fusion model fit to the 192 *sample size*, *June* 3, 2013 *dataset for the lag parameters* $(\mu + \alpha_{B_i})$ *for* 24*-hour average* $PM_{2.5}$ (*first column*), *daily eight-hour maximum ozone* (*second column*) *and of the weights in* (3.2) (*third column*). *The first row uses weights defined by the CDF of the standard normal distribution while the second row uses weights defined by the spherical spatial correlation function*.

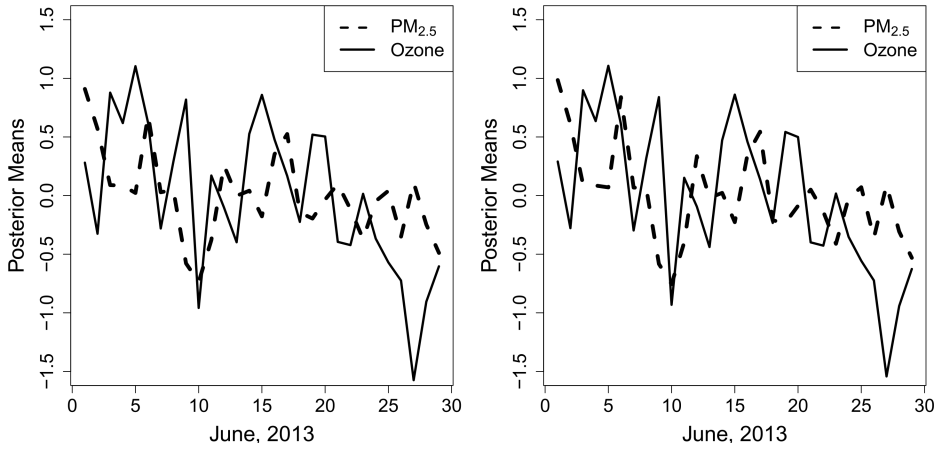

FIG. 4.

Posterior means of the time series components of the lags (μ_l) *from the DLfuse model fit to the June* 1–29, 2013, *dataset. The first panel uses weights defined by the CDF of the standard normal distribution while the second panel uses weights defined by the spherical spatial correlation function.*

TABLE 1

Results from the 24*-hour average* $PM_{2.5}$ *data analyses in* 2013. *EC: Empirical coverage; Length: Average credible interval length; VSS: Sample size of the validation data; VSD: Standard deviation of the validation data. PMAE and VSD are multiplied by* 100 *for presentation purposes*

		June 3			July 3			August 2		
Metric	**Method**	**64**	**128**	**192**	**64**	**128**	**192**	**64**	**128**	**192**
WAIC	OK	11.85	67.14	118.78	38.62	84.85	109.96	17.70	−28.05	15.79
	DS	3.73	62.31	106.47	38.54	70.03	93.66	17.47	−36.51	−38.50
	DS-Smooth	**−4.92**	48.23	94.94	**30.35**	**64.88**	**77.98**	**10.61**	−39.10	−19.18
	DS-Spectral	10.24	66.00	108.67	36.26	70.83	95.09	16.89	−33.12	−6.81
	DLfuse[a]	−1.66	**47.75**	**94.03**	36.09	67.46	78.59	13.31	**−48.46**	**−61.63**
	DLfuse[b]	3.67	57.20	102.27	37.53	67.71	91.57	19.83	−36.15	−60.60
PMAE	CMAQ	39.75	39.60	37.75	53.12	58.59	59.44	27.90	29.58	30.35
	OK	23.15	21.53	19.39	23.50	20.52	18.35	16.40	17.11	**15.75**
	DS	21.92	19.63	18.73	21.69	20.15	18.92	15.51	16.01	16.56
	DS-Smooth	20.58	19.48	18.04	**20.49**	20.18	18.67	15.82	16.07	16.52
	DS-Spectral	21.26	19.81	18.47	21.73	20.20	**18.15**	15.66	16.25	16.03
	DLfuse[a]	**20.51**	**18.72**	**18.01**	20.89	**19.98**	19.30	**15.35**	15.83	16.54
	DLfuse[b]	21.03	18.82	18.11	21.36	20.06	19.08	15.37	**15.77**	16.73
EC	OK	0.89	0.95	0.97	0.93	0.94	0.97	0.95	0.92	0.96
	DS	0.92	0.96	0.97	0.96	0.95	0.97	0.96	0.92	0.95
	DS-Smooth	0.91	0.96	0.97	0.95	0.96	0.97	0.96	0.92	0.94
	DS-Spectral	0.91	0.96	0.97	0.95	0.95	0.97	0.96	0.92	0.94
	DLfuse[a]	0.90	0.96	0.97	0.96	0.96	0.97	0.97	0.93	0.94
	DLfuse[b]	0.91	0.96	0.97	0.96	0.95	0.97	0.96	0.93	0.93
Length	OK	0.96	1.12	1.16	1.10	1.13	1.13	0.90	0.78	0.87
	DS	0.98	1.13	1.18	1.17	1.13	1.12	0.96	0.79	0.84
	DS-Smooth	**0.88**	**1.07**	**1.14**	**1.06**	1.11	**1.07**	0.91	0.77	**0.81**
	DS-Spectral	0.94	1.11	1.15	1.07	**1.11**	1.10	**0.89**	**0.76**	0.82
	DLfuse[a]	0.92	1.11	1.16	1.18	1.14	1.09	0.92	0.81	0.84
	DLfuse[b]	0.95	1.11	1.17	1.17	1.13	1.10	0.95	0.82	0.83
VSS		491	427	363	515	451	387	512	448	384
VSD		39.41	38.24	37.96	53.54	53.65	51.22	39.63	40.80	38.99

[a] Weights defined by the CDF of the standard normal distribution.

[b] Weights defined by the spherical spatial correlation function.

TABLE 2

Results from the eight-hour maximum ozone data analyses in 2013. *EC: Empirical coverage; Length: Average credible interval length; VSS: Sample size of the validation data; VSD: Standard deviation of the validation data. PMAE and VSD are multiplied by* 100 *for presentation purposes*

		June 3			July 3			August 2		
Metric	**Method**	**64**	**128**	**192**	**64**	**128**	**192**	**64**	**128**	**192**
WAIC	OK	−37.72	0.54	−19.65	33.49	−12.86	−14.71	−35.18	56.86	19.18
	DS	−48.59	−5.53	−31.64	24.80	−39.94	**−48.05**	−52.64	−31.53	−82.62
	DS-Smooth	**−50.39**	**−7.30**	**−32.45**	**20.19**	−39.43	−45.91	**−54.30**	23.90	−22.04
	DS-Spectral	−37.37	4.96	−18.06	34.86	−24.68	−32.21	−40.49	48.06	2.67
	DLfuse[a]	−48.85	−5.18	−32.16	25.06	−40.66	−45.45	−54.19	**−32.52**	**−82.96**
	DLfuse[b]	−49.01	−5.55	−32.12	24.88	**−40.72**	−47.51	−53.24	−30.50	−77.48
PMAE	CMAQ	22.63	22.17	21.82	27.86	28.22	27.97	20.08	20.13	19.27
	OK	14.86	13.29	13.04	16.39	13.44	13.11	15.18	14.50	13.68
	DS	11.82	12.35	**11.76**	**14.79**	13.16	13.20	11.67	11.96	11.22
	DS-Smooth	12.03	**12.07**	11.76	17.35	13.63	13.39	11.87	12.36	11.18
	DS-Spectral	12.27	12.48	11.88	15.11	**13.07**	**13.06**	12.72	12.52	11.47
	DLfuse[a]	**11.74**	12.34	11.86	14.82	13.12	13.13	**11.32**	**11.89**	**11.08**
	DLfuse[b]	11.81	12.29	11.80	14.80	13.16	13.15	11.59	11.92	11.08
EC	OK	0.91	0.95	0.98	0.98	0.96	0.97	0.94	0.99	0.98
	DS	0.95	0.96	0.98	0.98	0.96	0.96	0.95	0.96	0.96
	DS-Smooth	0.93	0.98	0.98	0.98	0.96	0.97	0.94	0.98	0.99
	DS-Spectral	0.93	0.97	0.98	0.98	0.96	0.97	0.93	0.99	0.98
	DLfuse[a]	0.94	0.97	0.98	0.98	0.96	0.97	0.95	0.96	0.96
	DLfuse[b]	0.95	0.96	0.98	0.98	0.96	0.97	0.95	0.96	0.96
Length	OK	0.64	0.83	0.82	1.10	0.80	0.83	0.69	1.14	0.96
	DS	0.61	**0.82**	**0.80**	1.08	0.75	0.78	0.59	**0.85**	0.78
	DS-Smooth	0.60	0.83	0.81	1.12	0.77	0.79	0.60	1.10	0.88
	DS-Spectral	**0.59**	0.84	0.81	**1.05**	**0.75**	**0.78**	0.59	1.03	0.88
	DLfuse[a]	0.61	0.83	0.81	1.08	0.76	0.79	**0.57**	0.85	**0.77**
	DLfuse[b]	0.60	0.83	0.80	1.08	0.75	0.78	0.58	0.87	0.79
VSS		770	706	642	786	850	914	780	844	908
VSD		28.30	26.82	27.59	42.67	42.53	41.77	24.79	23.13	23.49

[a]Weights defined by the CDF of the standard normal distribution.

[b]Weights defined by the spherical spatial correlation function.

TABLE 3

Results from the spatiotemporal data analyses in 2013. *EC: Empirical coverage; Length: Average credible interval length; VSS: Sample size of the validation data; VSD: Standard deviation of the validation data. PMAE and VSD are multiplied by* 100 *for presentation purposes*

		June 30		July 31		August 31	
Metric	**Method**	**$PM_{2.5}$**	**Ozone**	**$PM_{2.5}$**	**Ozone**	**$PM_{2.5}$**	**Ozone**
PMAE	CMAQ	67.24	19.48	41.06	29.20	33.84	25.38
	OK-ST	49.02	30.65	51.21	26.62	56.26	18.56
	DS-ST	48.66	18.19	28.57	18.58	34.26	13.55
	DLfuseST[a]	**44.99**	**16.81**	**25.52**	**17.98**	**25.18**	12.82
	DLfuseST[b]	45.04	16.89	25.59	17.99	25.41	**12.82**
EC	OK-ST	0.90	1.00	0.97	0.99	0.93	1.00
	DS-ST	0.76	0.89	0.94	0.88	0.94	0.96
	DLfuseST[a]	0.79	0.91	0.92	0.88	0.94	0.96
	DLfuseST[b]	0.79	0.91	0.92	0.89	0.94	0.96
Length	OK-ST	1.87	2.07	2.13	1.83	2.05	1.62
	DS-ST	1.33	0.77	1.41	0.77	1.30	0.77
	DLfuseST[a]	1.28	**0.76**	**1.32**	0.75	**1.24**	0.75
	DLfuseST[b]	**1.28**	0.76	1.32	**0.75**	1.24	**0.75**
VSS		529	835	199	841	200	836
VSD		37.75	22.96	48.02	27.95	47.36	22.80

[a] Weights defined by the CDF of the standard normal distribution.

[b] Weights defined by the spherical spatial correlation function.