The Annals of Applied Statistics
2014, Vol. 8, No. 3, 1538–1560
DOI: 10.1214/14-AOAS737
© Institute of Mathematical Statistics, 2014

NONLINEAR PREDICTIVE LATENT PROCESS MODELS FOR INTEGRATING SPATIO-TEMPORAL EXPOSURE DATA FROM MULTIPLE SOURCES

BY NIKOLAY BLIZNYUK[1,2,*], CHRISTOPHER J. PACIOREK[1,†],
JOEL SCHWARTZ[1,‡] AND BRENT COULL[1,‡]

University of Florida[*], *University of California*[†] *and*
Harvard School of Public Health[‡]

Spatio-temporal prediction of levels of an environmental exposure is an important problem in environmental epidemiology. Our work is motivated by multiple studies on the spatio-temporal distribution of mobile source, or traffic related, particles in the greater Boston area. When multiple sources of exposure information are available, a joint model that pools information across sources maximizes data coverage over both space and time, thereby reducing the prediction error.

We consider a Bayesian hierarchical framework in which a joint model consists of a set of submodels, one for each data source, and a model for the latent process that serves to relate the submodels to one another. If a submodel depends on the latent process nonlinearly, inference using standard MCMC techniques can be computationally prohibitive. The implications are particularly severe when the data for each submodel are aggregated at different temporal scales.

To make such problems tractable, we linearize the nonlinear components with respect to the latent process and induce sparsity in the covariance matrix of the latent process using compactly supported covariance functions. We propose an efficient MCMC scheme that takes advantage of these approximations. We use our model to address a temporal change of support problem whereby interest focuses on pooling daily and multiday black carbon readings in order to maximize the spatial coverage of the study region.

1. Introduction and background. An important scientific goal in environmental health research is the identification of sources of air pollution responsible for the well-documented health effects of air pollution. A pollution source of great interest is motor vehicle (i.e., traffic) emissions. Because traffic pollution is inherently higher near busy roads and major highways and falls off to background

Received February 2012; revised February 2014.

[1] Supported by Grants from the National Institute of Environmental Health Sciences (ES07142, ES000002, ES012044, ES016454, ES009825), the National Cancer Institute (CA134294), and the U.S. Environmental Protection Agency (EPA) Grant R-832416.

[2] Supported by the postdoctoral training Grant CA090301 from the National Cancer Institute.

Key words and phrases. Air pollution, approximate inference, covariance tapering, Gaussian processes, hierarchical model, likelihood approximation, particulate matter, semiparametric model, spatio-temporal model.

levels relatively quickly in space, concentrations of traffic-related pollutants exhibit large amounts of spatial heterogeneity within an urban area. Therefore, epidemiologic studies of the health effects of traffic pollution that use a centrally sited ambient monitor suffer from large amounts of exposure measurement error [Zeger et al. (2000)]. However, because it is not always feasible to obtain exposure recordings at each study subject's residence at a given time (a special case of spatio-temporal misalignment), it is now common practice in air pollution epidemiology for researchers to collect data from monitoring networks on the intra-urban spatio-temporal variability in traffic pollution levels. These data are used to make predictions of the exposure process, which are then used as a surrogate for true exposures in health effects models [Adar et al. (2010); Berhane et al. (2004); Wannemuehler et al. (2009)]. Note that this creates a measurement error problem [Gryparis et al. (2009)].

In this article we consider statistical models for prediction of spatio-temporal concentrations of black carbon (BC), thought to be a surrogate for traffic-related air particle levels [Janssen et al. (2011)], in the greater Boston-area. One complicating factor in the development of such models in our Boston-area analysis, however, is that the logistical and financial demands of maintaining a dedicated monitoring network are prohibitive. Accordingly, rather than set up a single network with a standardized monitoring protocol, our collaborators have augmented existing ambient monitoring data with targeted residence-specific indoor pollution monitoring aimed at increasing both the spatial and temporal coverage of the study region and period, respectively. Early work by our group [Gryparis et al. (2007)] focused on latent variable models for the integration of spatio-temporal data from multiple sources when all data were measured at the same temporal (in this case, daily) scale. The resulting number of monitors producing daily BC data was modest (under 90), limiting our ability to fully explore the spatio-temporal structure in the data. Specifically, such data sparsity motivated us to fit relatively simple spatial models separately for warm and cold seasons, as opposed to fitting more complex and likely more realistic spatio-temporal correlation structures across the entire study period.

Since this initial work, data at additional spatial locations have been collected. In this work, we consider data from 93 additional indoor monitors and explore how incorporation of these data improves our ability to explore the spatio-temporal patterns in the resulting monitoring data and ultimately the predictive performance of the resulting exposure models. One factor complicating the integration of these new data is the fact that this more recent monitoring campaign yielded concentration data at temporal scales different than the original BC data. Whereas the original data were collected on a daily time scale, the more recent monitoring campaign yielded multiday integrated readings. Therefore, our scientific interest focuses on the integration of data from these disparate sources into a unified exposure prediction framework, while rigorously accounting for changes in temporal support and the fact that different monitors operate irregularly in time. Given a modeling strategy that satisfies these goals, we assess the improvement in predictive performance of the models that incorporate all the data versus simpler models that only use the original daily data. While there has been a wide body of statistical work on spatio-temporal modeling of air pollution, most of these efforts have focused on data without substantial temporal misalignment and with a single type of pollution measurement. Although there is a considerable literature on the change of spatial and spatio-temporal support [Gelfand, Zhu and Carlin (2001)] and the use of aggregated data in spatial statistics [Gotway and Young (2002, 2007); Fuentes and Raftery (2005)], these proposed methods largely rely on the linear change-of-support and data assimilation. For example, Calder (2007, 2008) develops dynamic process convolution models—effectively, multivariate time series models—for multivariate spatio-temporal air quality data that allow one to solve the linear change-of-support problem. We are not aware of references that focus on the nonlinear change of temporal support in spatio-temporal statistics.

We now outline the structure of the available BC data in more detail. The three data types that we use in our model and describe below are (i) *daily* average outdoor BC concentrations (abbreviated as BCO), (ii) *daily* average indoor BC concentrations (BCI) and (iii) *multiday* aggregated indoor BC concentrations (BCA), in $\mu g/m^3$. Figure 1 and Figure 4 of the online supplements [Bliznyuk et al. (2014)]

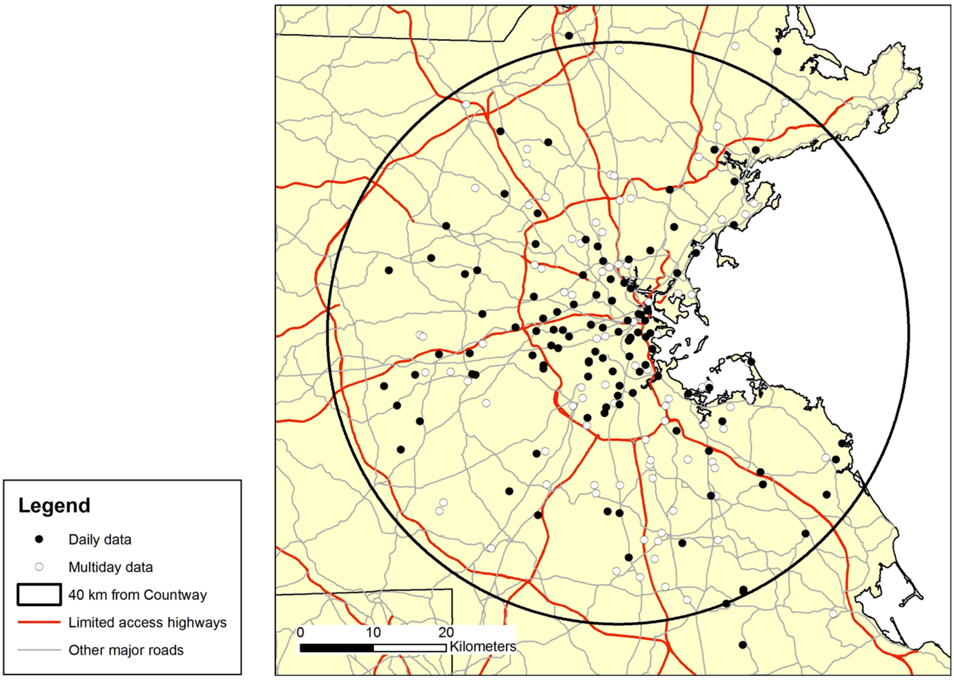

FIG. 1. *Spatial coverage of monitors. The HSPH* (*Countway*) *monitor is at the center of the circle in Downtown Boston.*

display the spatial and temporal coverage of the study region and periods, distinguishing the different data types.

Daily outdoor data (*BCO*). A sizeable fraction of the BCO readings that we use come from Gryparis et al. (2007). These data, generated from two different exposure assessment studies, were collected by outdoor monitors at 48 spatial locations in Boston and its suburbs over the period from mid October 1999 to the end of September 2004. The length of time each monitor operated ranged from 2 weeks to hundreds of days. These monitoring efforts resulted in 4219 daily BC readings. Our analyses supplement these data with additional daily BC data collected as part of a recent NIH Program Project Grant (PPG), which added 2696 daily readings from 52 distinct sites taken between mid March 2006, and early November 2008. The observations from the two studies do not occur at the same spatial locations, thus, we have 100 distinct sites, with over 6900 daily BCO readings.

Daily indoor data (*BCI*). These data consist of 318 daily indoor concentrations of BC from 45 distinct households, recorded between mid November 1999 and early December 2003. Of these 45 sites, 30 overlap spatially and temporally with the BCO data. Further details are in Gryparis et al. (2007).

Multiday aggregated indoor data (*BCA*). Multiday measurements of indoor BC were collected as part of the Normative Aging Study (NAS). There are 93 observations, one per household, each of which is a measurement of concentration aggregated over multiple days; the corresponding daily concentrations are not available. The lengths of measurement, which range from 3 to 14 days, and starting dates of the monitoring periods are different across the households. The data correspond to the period from mid July 2006 through late March 2008. The spatial locations of the multiday data are distinct from those of the daily data.

To achieve the scientific goals outlined above, we develop a Bayesian hierarchical framework for inference and prediction where a joint model for all exposure measurements depends on a set of submodels, one for each data source, and a model for the latent process that relates the submodels to one another. In particular, we focus on the case in which the submodels depend on the latent process nonlinearly, which frequently occurs when the different data sources yield data on different surrogates of pollution or at varying temporal or spatial scales.

Inference for nonlinear hierarchical models with latent Gaussian structure is computationally challenging. When the likelihood is not Gaussian, likelihood-based and Bayesian inference involve high-dimensional integration with respect to the random effects that cannot be expressed in closed form. While MCMC is a standard approach for such models in a Bayesian framework, convergence and mixing are often troublesome [Christensen and Waagepetersen (2002), Christensen, Roberts and Sköld (2006)] because of the high-dimensionality of the random effects and the dependence between random effects (particularly in spatio-temporal specifications) and cross-level dependence between random effects and their hyperparameters [Rue and Held (2005), Rue, Martino and Chopin (2009)].

Our main methodological contribution is development of an efficient, yet straightforward, MCMC algorithm for Bayesian inference on model parameters and prediction of arbitrary functions of the latent process. Within our hierarchical model, it is based on the approximation of nonlinear regression functions by "linearizing" them with respect to the latent process values over the region of their high posterior probability.

The paper is organized as follows: in Section 2 we describe the overall hierarchical modeling strategy, proposing a nonlinear statistical model in Section 2.1 that is approximated through "linearization." In Section 3 we present a computational strategy to reduce the cost of Bayesian inference and prediction and discuss the relative merits of our approach and existing approximation methods. Section 4 is devoted to model selection and validation for pollutants in the greater Boston data, assessment of the adequacy of linearization and the results of Bayesian inference and prediction. Discussion and concluding remarks are in Section 5. Technical details and supplementary figures and tables are in the online supplements [Bliznyuk et al. (2014)].

2. Statistical model.

This section defines the joint model for the observed data. The individual models for the observations of each type are linked through the latent process. The nonlinear model for the multiday data is subsequently "linearized" for the sake of computational tractability.

2.1. *Nonlinear observation model.*

Following Gryparis et al. (2007), the latent spatio-temporal process, η, is a proxy for the logarithm of the true daily average concentration of outdoor black carbon (BC). The model for η will be specified in the next subsection. For notational simplicity, we will often abbreviate the space–time indices using subscripts, for example, $\eta(s_i, t_j)$ as η_{ij} for the value of the latent process at site s_i on date t_j. The logarithms of the observed outdoor and indoor daily average BC concentrations, Y_{ij}^O and Y_{ij}^I, are related to the latent process as

$$Y_{ij}^O = \eta_{ij} + \varepsilon_{ij}^O, \tag{1}$$

$$Y_{ij}^I = \alpha_{0i} + \alpha_{1I}\eta_{ij} + \varepsilon_{ij}^I, \tag{2}$$

where $\{\alpha_{0i}\}$ are household-specific fixed effects and ε_{ij}^O, ε_{ij}^I are instrument errors. The household-specific effects are introduced as in Gryparis et al. (2007) in order to account for that differences in penetration efficiencies of particles that depend on properties of the building. In the absence of instrument error, setting the slope $\alpha_{1I} = 1$ corresponds to the indoor BC being proportional to the outdoor BC on the original scale, with the proportionality constant $\exp(\alpha_{0i})$. The values of the slope

α_{1I} less than one—such as those observed with our data—allow one to account for the slower than linear increase in the indoor BC as the outdoor BC grows, relative to the proportional concentration model.

The model for the observed average *multiday* concentration of indoor black carbon at a site s_i is defined as

$$Y_i^A = \alpha_{0i} + g_i(\vec{\eta}_i^A) + \varepsilon_i^A, \tag{3}$$

where $\vec{\eta}_i^A$ is the vector of (daily) latent process values upon which the aggregate average reading at s_i depends and ε_i^A is the instrument error. We assume that the instrument error processes $\{\varepsilon_{ij}^O\}$, $\{\varepsilon_{ij}^I\}$ and $\{\varepsilon_i^A\}$ are mutually independent Gaussian white noise with zero mean and variances σ_O^2, σ_I^2 and σ_A^2, respectively.

Without loss of detail, let Y_i^A be the logarithm of the sum (as opposed to an average) of consecutive daily average concentrations of indoor black carbon at site s_i. The nonlinear regression model for the multiday data is

$$g_i(\vec{\eta}_i^A) = \log \sum_j \exp(\alpha_{1I} \cdot \eta_{ij}). \tag{4}$$

The nonlinearity arises because the multiday readings are aggregated on the original rather than on the logarithmic scale. For instance, without the instrument error, that is, if $\varepsilon_{ij}^I = \varepsilon_i^A = 0$, Y_i^A would be the logarithm of $\sum_j \exp(Y_{ij}^I)$, the sum of consecutive daily readings of (daily) average indoor black carbon concentrations at site s_i. Trivially, equation (2) is a special case of equation (3).

Note that there is only a single reading Y_i^A for each household, so the home-specific intercepts α_{0i} are not identifiable in the model of equation (3). We therefore absorbed them into ε_i^A, but introduced the parameter α_{00} to capture the population intercept. Exploratory analysis revealed that the slope parameter α_I can be significantly different for the models for Y^I and Y^A. Consequently, the model (3) was changed to

$$Y_i^A = \alpha_{00} + \log \sum_j \exp(\alpha_{1A} \cdot \eta_{ij}) + \varepsilon_i^A. \tag{5}$$

The coefficient α_{1A} is allowed to be different from α_{1I} in the model for Y^I, in order to account for (i) data aggregation and rounding errors, since the monitors from the NAS study do not run for an integer number of days, and (ii) demographic differences in households since the multiday data come from a study (targeting elderly people) different from the study providing the daily indoor data.

2.2. *Latent process model.* The latent process at site s_i on day t_j is modeled as

$$\eta(s_i, t_j) = x(s_i, t_j)^{\mathsf{T}} w_x + \zeta(s_i, t_j) + u(s_i, t_j), \tag{6}$$

where $x(s_i, t_j)$ is a vector of observable predictors and $\zeta(s_i, t_j) + u(s_i, t_j)$ accounts for unobservable spatio-temporal variability. In order to ensure identifiability, we let ζ capture the temporally long-range spatio-temporal variability and

u capture the temporally short-range variability. Equivalently, for a fixed value s_0 of s, $u(s_0, \cdot)$ is a stationary temporal process with rapidly decaying dependence, and $\zeta(s_0, \cdot)$ is a long-range temporal process, possibly with nondecaying dependence.

For our case study, components of the vector of observable covariates $x(s, t)$ in equation (6) are provided in Table 1. They include (i) spatially-varying variables—population density, traffic density and land use; (ii) temporally-varying variables—readings from the Harvard School of Public Health (HSPH) central site monitor, meteorological variables (wind speed and planetary boundary layer); and (iii) interaction terms. We use the logarithm of readings from the HSPH central site monitor as a predictor rather than as a response in order to enable comparisons with earlier work of Gryparis et al. (2007) that set $u = 0$. The implication is that much of the temporal variability common to all sites is captured by observations from the central site and that the temporal components of the model capture variability above and beyond that measured at the central site.

Following Opsomer, Wang and Yang (2001), we let ζ capture the long-range spatio-temporal variation, often referred to as the unknown smooth spatio-temporal trend. In the spirit of Wang (1998), we use penalized splines, so that the trend can be represented as

$$\zeta(s, t) = z(s, t)^{\mathsf{T}} w_z, \tag{7}$$

where $z(s, t)$ is a column vector of known basis functions evaluated at (s, t) and w_z is a column vector of the corresponding coefficients. We define the actual form of $z(s, t)$ and constraints on w_z below. Because the spatio-temporal coverage by the monitors is sparse (about 7300 observations from over 2700 distinct days and at most 200 sites), unconstrained spatio-temporal smoothing would be unreliable in parts of the domain without observations. Instead we put constraints on the spatio-temporal smoother by requiring the smoother to be periodic, thereby borrowing strength across years when estimating the trend. This also allows one to make predictions outside the temporal range of the observations. The local deviations of the latent process from the periodic term will be accounted for by the $u(s, t)$ process.

We decompose the long-range spatio-temporal trend as

$$\zeta(s, t) = g_S(s) + g_T(t) + g_{ST}(s, t),$$

where g_S and g_T are smooth functions of spatial coordinates and of time, respectively, and g_{ST} is a function representing the long-range (in time) spatio-temporal interaction. Here, g_T is the annual (cyclic) temporal trend, so that $g_T(t) = g_T(d_t)$, where $d_t = \mathrm{mod}(t, 365)$ is the day of the year if leap years are ignored. We use a thin-plate spline with 60 knots to model g_S, a cubic spline with seven equally spaced knots to model g_T, and the tensor product of spatial and temporal basis functions to model the interaction, g_{ST} [Wood (2006)]. To ensure that the temporal trend is periodic, continuous and differentiable at $t = 0$, linear constraints were

TABLE 1

Posterior summaries of the coefficients of the observed predictors under model $M(U = 1, GST = 0, A = 1)$

$\boldsymbol{w_i}$	**Predictor**	**Mean**	**2.5%**	**50%**	**97.5%**
w_1	`1`, the intercept	4.580	−2.190	4.176	13.516
w_2	`log_pop_sqkm`, log of population per square km	0.259	0.021	0.262	0.494
w_3	`log_adtxlth100m`, log of traffic density	−0.177	−0.306	−0.176	−0.049
w_4	`nlcd`, land use index	$2.65 \cdot 10^{-4}$	$1.19 \cdot 10^{-4}$	$2.63 \cdot 10^{-4}$	$4.14 \cdot 10^{-4}$
w_5	`loghsph`, log of HSPH monitor readings	0.767	0.742	0.767	0.793
w_6	`wind_sp`, wind speed	0.129	0.014	0.130	0.244
w_7	`log_pbl`, log of planetary boundary layer	−0.073	−0.244	−0.074	0.095
w_8	`log_pop_sqkm * wind_sp`	−0.028	−0.053	−0.028	−0.003
w_9	`log_adtxlth100m * wind_sp`	0.002	−0.011	0.002	0.015
w_{10}	`log_pbl * wind_sp`	−0.023	−0.040	−0.023	−0.006
w_{13}	`log_pop_sqkm * log_pbl * wind_sp`	0.004	0.001	0.005	0.008
w_{14}	`log_adtxlth100m * log_pbl * wind_sp`	0.000	−0.002	0.000	0.002

placed on the coefficients of g_T and of g_{ST}; see online supplements [Bliznyuk et al. (2014)], Section A.5.3. Thus, the model of equation (6) can be written as a linear model

$$\eta(s,t) = c(s,t)^{\mathsf{T}} w + u(s,t), \tag{8}$$

where

$$c(s,t)^{\mathsf{T}} = \{x(s,t)^{\mathsf{T}}, [s; \boldsymbol{\phi}(s)]^{\mathsf{T}}, [t; \boldsymbol{\psi}(t)]^{\mathsf{T}}, [s; \boldsymbol{\phi}(s)]^{\mathsf{T}} \otimes [t; \boldsymbol{\psi}(t)]^{\mathsf{T}}\}$$

is a row vector of "predictors" and

$$w = [w_x^{\mathsf{T}}, w_S^{\mathsf{T}}, w_T^{\mathsf{T}}, w_{ST}^{\mathsf{T}}]^{\mathsf{T}}$$

is a column vector of coefficients. Here, the ith component of $\boldsymbol{\phi}(\cdot)$ is $\phi_i(\cdot) = \phi(\cdot, s^{(i)})$, the spatial basis function centered at the knot $s^{(i)}$; similarly, $\psi_j(\cdot) = \psi(\cdot, t^{(j)})$ is the jth temporal basis function centered at the knot $t^{(j)}$, for example, $\psi_j(d_t) = |d_t - t^{(j)}|^3$. Following Wood (2006), we penalize the square of the second derivative of the nonparametric smooth terms to prevent overfitting. This approach is attractive because the penalty matrices for g_S and g_T can be written as symmetric positive semidefinite quadratic forms in w_S and w_T. For example, the penalty for g_T is

$$\mathcal{P}_T = \int \{g_T''(t)\}^2 \, dt = w_T^{\mathsf{T}} \cdot M_T \cdot w_T \tag{9}$$

for some symmetric positive semidefinite matrix M_T. The spatial and temporal marginal penalty matrices $M_{ST,S}$ and $M_{ST,T}$ for the smooth interaction term g_{ST} are derived in the online supplements [Bliznyuk et al. (2014)], Section A.5.2. These penalty matrices are subsequently used to define a precision matrix for the multivariate normal prior on w as the Bayesian analogue of the penalized log-likelihood criterion with penalty matrices M_S, M_T, $M_{ST,S}$ and $M_{ST,T}$ [Ruppert, Wand and Carroll (2003)]. This prior has a zero mean and precision matrix

$$Q_w = \text{blkdiag}\left\{\Delta \cdot I_{\dim(w_x)}, \frac{M_S}{\tau_S^2}, \frac{M_T}{\tau_T^2}, \frac{M_{ST,S}}{\tau_{ST,S}^2} + \frac{M_{ST,T}}{\tau_{ST,T}^2}\right\}, \tag{10}$$

where a small multiple Δ of the identity matrix is used to ensure that the prior on the linear coefficients w_x is proper and where blkdiag is a block-diagonal matrix with blocks listed as arguments.

We use a Gaussian process model for u in order to account for the short-range temporal variability and spatio-temporal interaction. Given the data sparsity, we model the covariance function for u in a separable fashion for simplicity as

$$\text{Cov}\{u(s,t), u(s',t')\} = \sigma_u^2 \cdot C_S(s, s' | \theta_S) \cdot C_T(t, t' | \theta_T), \tag{11}$$

where C_S and C_T are spatial and temporal correlation functions.

To model spatial dependence, we use the Matérn family of correlation functions

$$C_S(s, s+h) = \bigl(2\sqrt{\nu}\theta_S\|h\|_2\bigr)^\nu \cdot K_\nu\bigl(2\sqrt{\nu}\theta_S\|h\|_2\bigr)/\bigl\{2^{\nu-1}\Gamma(\nu)\bigr\}, \tag{12}$$

where $\nu, \theta_S > 0$, $\Gamma(\cdot)$ is the gamma function and $K_\nu(\cdot)$ is the modified Bessel function of order ν [Banerjee, Carlin and Gelfand (2004)]. The smoothness parameter ν is difficult to estimate accurately unless the spatial resolution of the data is very fine [Gneiting, Ševčíková and Percival (2012)]. Due to the spatial sparsity of the set of monitors, we hold ν fixed at 2, thereby representing smooth short-range (based on the tapering described next) variation, in $u(s,t)$. Nonsmooth variability is accounted for by the errors ε.

Examination of the plots of autocorrelation and partial autocorrelation functions for residuals from a monitoring station with a long series of daily measurements suggested that temporal dependence can be explained well by an order-one autoregressive process with moderate lag-one correlation (of less than 0.5). Under the plausible assumption that the rates of decay of the temporal autocorrelation are similar across all monitoring stations, it can be seen that the components of u that are 7 days or more apart are practically uncorrelated since the correlation is less than 10^{-2}. Consequently, we introduce sparsity structure into the covariance matrix explicitly via covariance tapering [Furrer, Genton and Nychka (2006)]. As a temporal correlation function, we use the product of the exponential and the (compactly supported) spherical correlation functions

$$C_T(t, t+h) = \exp(-\theta_T \cdot h) \cdot \max\bigl\{(1 - h/r), 0\bigr\}^2\bigl\{1 + h/(2r)\bigr\} \tag{13}$$

for $r = 7$, which behaves similarly to the exponential correlation function when h is small, and is exactly zero when $h \geq 7$. The benefits of tapering for the computational aspects of Bayesian inference will be discussed Section 3.

2.3. *Linearized observation model.*

We will refer to the set of equations (1)–(6) as the *nonlinear model*. MCMC for such models can be very inefficient, if tractable at all. For example, if one puts a Gaussian spatio-temporal process prior on u, one needs to sample from a nonstandard density for the vector of latent process values (here, of dimension 712) that enters the nonlinear model for Y^A. The values cannot be analytically integrated over in the joint model. In this subsection we develop the idea of "linearization" of the nonlinear regression function of equation (4) about some "central" value η^{A*} of the latent process and briefly discuss the practical choices for η^{A*}.

2.3.1. *Linearization.*

The linearized model is obtained by doing a Taylor series expansion of the nonlinear regression function g_i in equation (4) about some "central" value $\vec{\eta}_i^{A*}$ of vector $\vec{\eta}_i^A$:

$$Y_i^A = G_i(\alpha_{1A}) + \alpha_{00} + \sum_{j=1}^{J_i} b_{ij}(\alpha_{1A})\eta_{ij} + \varepsilon_i^A, \tag{14}$$

where J_i is the number of days in the aggregated reading at site s_i, $J_i \in \{3, \dots, 14\}$. Here, b_{ij} and G_i are known deterministic functions of α_{1A}:

$$G_i(\alpha_{1A}) = g_i(\vec{\eta}_i^{A*}) - \{b_i(\alpha_{1A})\}^{\mathsf{T}} \vec{\eta}_i^{A*}, \tag{15}$$

$$b_i(\alpha_{1A}) = [b_{i1}, \dots, b_{iJ_i}]^{\mathsf{T}} = \frac{\partial g_i(x)}{\partial x}\bigg|_{x=\vec{\eta}_i^{A*}} \quad \text{and} \tag{16}$$

$$\vec{\eta}_i^{A} = [\eta_{i1}, \dots, \eta_{iJ_i}]^{\mathsf{T}}, \tag{17}$$

$$\frac{\partial g_i(\vec{\eta}_i^{A})}{\partial \eta_{ij}} = \frac{\alpha_{1A} \exp(\alpha_{1A}\eta_{ij})}{\sum_j \exp(\alpha_{1A}\eta_{ij})}. \tag{18}$$

Notice that the model obtained by replacement of equation (3) by equation (14) is a conditionally linear model given α_{1A}.

Define $v = (w; \{\alpha_{0i}\})$ and let γ be the vector of all remaining parameters, which includes $\alpha_{1I}, \alpha_{1A}, \sigma_O^2, \sigma_I^2, \sigma_A^2$, variance components controlling the smoothness of ζ and parameters of the covariance function of u. We can then write the linearized joint model for the observed data of all types in matrix form as

$$Y = H(\alpha_1) \cdot (1; v) + \xi, \tag{19}$$

where $\xi = X(\alpha_1)u + \varepsilon$ and $\alpha_1 = (\alpha_{1I}, \alpha_{1A})$. Here, H and X are matrices that do not depend on v, as follows from equations (2)–(6) and (14). The scalar 1 is necessary to capture the offset due to $G_i(\alpha_{1A})$ in the linearized model for Y^A, equation (14). Notice that, conditional on α_1, this is a linear model with dependent Gaussian errors, which allows a computationally efficient implementation of an MCMC sampler, discussed in Section 3.

2.3.2. *Choice of the central value of the latent process.* The scheme outlined above assumes the availability of the point η^{A*} about which the linearization is performed. In this section we detail how this value can be obtained and justified. We use a standard bracket notation for marginal, $[\cdot]$, and conditional, $[\cdot|\cdot]$, densities [Ruppert, Wand and Carroll (2003)].

Upon defining $Y^{OI} = (Y^O, Y^I)$ and changing the order of conditioning as

$$\begin{aligned}&[Y^{OI}, Y^A|\eta^A, w, \gamma][\eta^A, w, \gamma]\\ &\quad = [Y^A|\eta^A, w, \gamma, Y^{OI}][\eta^A|w, \gamma, Y^{OI}][w, \gamma|Y^{OI}][Y^{OI}],\end{aligned}$$

it is seen that the posterior $[\eta^A, w, \gamma|Y^{OI}]$ for the daily data, Y^O and Y^I, implicitly acts as an informative prior for the parameters and the latent process in the multiday model likelihood $[Y^A|\eta^A, w, \gamma, Y^{OI}]$. (Since $\{\alpha_{0i}\}$ can be integrated out analytically, v is replaced by w here.) Because $[\eta^A, w, \gamma|Y^{OI}] = [\eta^A|w, \gamma, Y^{OI}][w, \gamma|Y^{OI}]$, the mass of the density of the latent process vector η^A is concentrated around the best linear unbiased predictor (BLUP) $E(\eta^A|Y^{OI}, w =$

$\widehat{w}, \gamma = \widehat{\gamma})$, where $\widehat{w}$ and $\widehat{\gamma}$ are some "central" values of w and γ. This suggests the use of $\eta^{A*} = E(\eta^A | Y^{OI}, w = \widehat{w}, \gamma = \widehat{\gamma})$ in the linearization. In fact, as the daily data become dense in space, infill asymptotics suggest that the BLUP $E(\eta^A | Y^{OI}, w = \widehat{w}, \gamma = \widehat{\gamma})$ converges to the true unobserved value of η^A. Consequently, (14) provides a likelihood for Y^A that results in Bayesian inferences and predictions that are asymptotically equivalent to those under the true nonlinear model. Of course, the validity of this large-sample argument may be questionable in some applications. For our case study, we justify use of the linearized model for BCA empirically using a cross-validation study in Section 4.2. In Section 4.3 we assess the accuracy of inferences under the linearized model against those under the nonlinear model in the simplest case when neither long- nor short-range dependence is included in the model.

Taylor expansion about $\eta^{A*} = E(\eta^A | Y^{OI}, w = \widehat{w}, \gamma = \widehat{\gamma})$ is computationally tractable because the marginal posterior $[\gamma | Y^{OI}]$ or the profile posterior $\sup_w [\gamma, w | Y^{OI}]$ can be obtained analytically (up to a constant of proportionality) and hence maximized efficiently to get $\widehat{\gamma}$; the corresponding value of w is available analytically. In contrast, a possible alternative of expanding about the mode of $[\eta^A, w, \gamma | Y^{OIA}]$ would require a costly numerical optimization run.

Notice that the naïve solution to the temporal change of support problem, that is,

$$Y_i^A = \beta_0 + \beta_1 \sum_{j=1}^{J_i} \eta_{ij} + \varepsilon_i^A, \tag{20}$$

arises as a special case of our linearized model when $\vec{\eta}_i^{A*}$ is set to zero. In this case, $G_i(\alpha_{1A}) = \log J_i$ and $b_{ij}(\alpha_{1A}) = \alpha_{1A}/J_i$, where J_i is the observation period length at the site s_i. The linearized model becomes

$$Y_i^A = \log J_i + \alpha_{00} + \frac{\alpha_{1A}}{J_i} \sum_{j=1}^{J_i} \eta_{ij} + \varepsilon_i^A,$$

which is equivalent to the above "naive" model when the observation period lengths J_i are all equal. However, this is hardly appropriate in our case study since the observation periods lengths vary from 3 to 14 days, which implies that α_{00} and α_{1A} cannot be identified from β_0 and β_1. More importantly, the naïve linearization about 0 is inferior from the methodological standpoint since, unlike the linearization about $E(\eta^A | Y^{OI}, w = \widehat{w}, \gamma = \widehat{\gamma})$, the approximation error in the Taylor expansion does not go to zero as the spatial design becomes dense.

3. Computational considerations for Bayesian inference by MCMC. In this section we develop three strategies that lower the computational burden of model fitting and prediction: (i) covariance tapering, (ii) strategies for sampling from the posterior density of the model parameters, and (iii) sampling strategies for latent process prediction.

Without tapering, the covariance matrix of the vector ξ in equation (19), Σ_Y, is numerically dense. It can take on the order of several seconds on a modern computer to form and factorize this matrix, making a long MCMC sample computationally expensive. Tapering reduces the proportion of nonzero entries (the fill) of Σ_Y to less than 2%. In addition, we reorder the observed data Y^O lexicographically with respect to the temporal index, which makes unnecessary the formal element reordering approaches [Furrer, Genton and Nychka (2006)]. This makes Σ_Y a banded (block) arrowhead matrix (see Figure 5 in the online supplements [Bliznyuk et al. (2014)] for a visualization), which yields a very efficient sparse Cholesky factorization [Golub and Van Loan (1996)]. As a result, the cost to evaluate the likelihood drops by at least an order of magnitude. For a general nonlinear model in which the joint posterior density of (v, γ) is computationally expensive to evaluate and tapering is not appealing, our linearization strategy can be supplemented by the dimension reduction scheme of Bliznyuk, Ruppert and Shoemaker (2011) for efficient approximation of high-dimensional densities.

We now discuss a strategy for sampling from the posterior density of model parameters. Recall from Section 2.3.1 that $\alpha = \{\alpha_{0i}\}$, $v = (w; \alpha_0)$ and γ is the vector of all other parameters. We analytically integrate v from the model as $[v|\gamma, Y]$ is multivariate normal. Consequently, we draw from $[v, \gamma|Y]$ using *composition sampling*, that is, by sampling $\gamma^{(i)}$ from $[\gamma|Y]$, and then by exactly sampling v from $[v|Y, \gamma = \gamma^{(i)}]$, which is in the spirit of the partially collapsed Gibbs samplers work of van Dyk and Park (2008). In order to sample from $[\gamma|Y]$, we use an adaptive random walk Metropolis–Hastings (RWMH) sampling scheme, in the spirit of Haario, Saksman and Tamminen (2001), that calibrates the covariance matrix of the proposal distribution based on the past trajectory of the Markov chain. The lag-1 autocorrelation in the components of γ in the actual sampling was below 0.95, while mixing for the components of v was considerably better; see Section 4.3 for details. The actual expressions for $[\gamma|Y]$ and $[v|\gamma, Y]$ are provided in the online supplements [Bliznyuk et al. (2014)], Section A.1.

For health effects studies and for fine visualization of the spatio-temporal variability of the latent process, one often needs to predict the values of the latent process, η^P, at a large set of spatio-temporal indices, say, at a regular grid with N_s spatial sites over the course of N_t days. In order to simulate from $[\eta^P|Y]$ under the linearized model, one needs to (i) sample from $[\gamma, v|Y]$ as in Section 3 and (ii) for each state in the (γ, v)-chain, sample exactly from $[\eta^P|\gamma, v, Y]$, which is a multivariate normal density. If (γ^*, v^*) is a given value of (γ, v) and $\operatorname{Cov}(u^P, Y|v^*, \gamma^*) = \Sigma_{u^P,Y}(\gamma^*)$ and $\operatorname{Var}(Y|v^*, \gamma^*) = \Sigma_{Y,Y}(\gamma^*)$, one generally needs to efficiently compute

$$\begin{aligned}
&E(\eta^P|Y, v = v^*, \gamma = \gamma^*) \\
&\quad = E(\eta^P|v = v^*, \gamma = \gamma^*) \\
&\qquad + \Sigma_{u^P,Y}(\gamma^*) \cdot \Sigma_{Y,Y}^{-1}(\gamma^*) \cdot \{Y - E(Y|v = v^*, \gamma = \gamma^*)\}.
\end{aligned}$$

For example, if one estimates $E(\eta^P|Y)$ by Monte Carlo via "Rao–Blackwellization" [e.g., Robert and Casella (1999)], then $E(\eta^P|Y) \approx M^{-1}\sum_{i=1}^{M} E(\eta^P|Y, v = v^{(i)}, \gamma = \gamma^{(i)})$. "Poor man's" approximations of the form $E(\eta^P|Y, v = v^*, \gamma = \gamma^*)$, where (v^*, γ^*) is the posterior mode or the posterior mean, are also possible. Section A.2 of the online supplements [Bliznyuk et al. (2014)] provides computational details of evaluation of $E(\eta^P|Y, v, \gamma)$ and of sampling from $[\eta^P|v, \gamma, Y]$.

4. Analysis and results for the greater Boston data.

4.1. *Candidate models.* Here we consider whether simpler models, such as the model of Gryparis et al. (2007), achieve comparable predictive accuracy to the full model presented in Section 2. We examine eight candidate models, each determined by a combination of following 3 indicator variables: U—whether the model includes a Gaussian process model for the short-range dependence term, u, or assumes that $u = 0$; GST—whether an extra term g_{ST} for the smooth long-range spatio-temporal interaction is included; and A—whether the aggregated multiday data, Y^A, are used (so as to assess their importance in improving predictions). We use this labeling scheme to abbreviate the models, for example, $M(U = 0, GST = 0, A = 1)$. We assess the models through cross-validation with spatially nonoverlapping subsets.

4.2. *Assessment of predictive performance on validation data.* We allocated a total of 1593 daily outdoor black carbon readings from 48 distinct sites into four disjoint groups of 12 sites, with each group having roughly 400 data values. To achieve this, we generated random partitions of the 48 sites into 4 groups many times and chose the partition that maximized the minimum pairwise distance between sites and achieved roughly the same number of observations in each group. We held out each of the four validation subsets in turn, training the model with the remaining observations and obtaining predictions to compare with the held-out subset. Although the training and validation subsets of data are spatially nonoverlapping, they are not temporally disjoint. To expedite model fitting, we used optimization to find the mode $\widehat{\gamma}$ of $[\gamma|\text{train}]$ and then analytically obtained the corresponding value $\widehat{v}(\widehat{\gamma})$ that maximizes $[v|\gamma = \widehat{\gamma}, \text{train}]$, after which we use the (empirical) BLUP $E(Y^V|\text{train}, v = \widehat{v}, \gamma = \widehat{\gamma})$ to obtain predictions. Here, train is the "training" data, which is $\{Y^{OI} \setminus Y^V\}$ or $\{Y^{OIA} \setminus Y^V\}$, depending on the model. This can be viewed as an analogue of the frequentist procedure that estimates the variance components and smoothing parameters by REML (restricted maximum likelihood) and then solves the quadratic minimization problem to fit the penalized spline. Of course, rather than estimating the mode, the more time-consuming alternative of estimating the posterior mean via MCMC could be used. Treating v and γ as known (set to their estimated values) allows us to derive the predictive distribution of the validation data and prediction errors used for the prediction interval and probability scores below [Gneiting and Raftery (2007)].

TABLE 2

Comparisons of cross-validation performance for the 8 *candidate models using averaged* (*over four subsets*) *criteria. Columns*: *B—MSPE*; *C—correlation*; *D—empirical coverage of the prediction interval*; *E—average width of the prediction interval*; *F—negatively oriented interval score*, *equation* (43) *of Gneiting and Raftery* (2007); *G—negatively oriented CRPS*, *equations* (20) *and* (21) *of Gneiting and Raftery* (2007); *H—plug-in maximum likelihood prequential score*, *equation* (54) *of Gneiting and Raftery* (2007)

U: is	**GST: is**	**A: is**							
u used?	**g_{ST} used?**	**Y^A used?**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
0	0	0	0.264	0.674	0.878	1.359	3.336	0.955	−0.083
0	0	1	0.161	0.777	0.907	1.360	2.165	0.527	−0.034
0	1	0	0.580	0.610	0.840	1.338	6.184	2.376	−0.175
0	1	1	0.167	0.768	0.896	1.336	2.265	0.562	−0.045
1	0	0	0.143	0.802	0.937	1.402	1.975	0.435	−0.004
1	0	1	0.132	0.816	0.938	1.394	1.918	0.396	0.007
1	1	0	0.172	0.759	0.906	1.383	2.267	0.543	−0.035
1	1	1	0.141	0.803	0.931	1.381	1.984	0.430	−0.005

Once the predictions are available, we measure the predictive accuracy using the mean squared prediction error (MSPE) and correlation between the predicted and observed validation values (columns B and C in Table 2). We also considered criteria based on 95% prediction intervals, particularly the observed proportion of coverage (column D), average width (column E) and the negatively oriented interval score of Gneiting and Raftery (2007) defined as

$$S_\alpha^{\mathrm{int}}(l,u;x) = (u-l) + \frac{2}{\alpha}(l-x)\mathbb{I}(l-x) + \frac{2}{\alpha}(x-u)\mathbb{I}(x>u),$$

where $\alpha = 0.05$, l and u are the lower and upper bounds of the size $(1-\alpha)$ central prediction interval, and $\mathbb{I}(\cdot)$ is the indicator function (column F). Given comparable empirical coverages, lower values in columns E and F correspond to better fitting models. Column G summarizes the negatively oriented continuously ranked probability scores defined as

$$\mathrm{CRPS}(F,x) = \int_{-\infty}^{\infty} \{F(y) - \mathbb{I}(y \geq x)\}^2\, dy,$$

where $F(y)$ is the predictive distribution of interest, which has recently drawn the attention of the atmospheric sciences community [see Gneiting and Raftery (2007) and the references therein]. In column H we include summaries based on the equivalent of the plug-in maximum likelihood prequential score $\sum_{j\in V_i}\log[Y^{(j)}|\,\text{train}, v=\widehat{v}, \gamma=\widehat{\gamma}]$, where $Y^{(j)}$ is the jth observation in V_i. Higher values in columns G and H correspond to better fitting models. The criteria reported in the table are averaged over the four subsets of data, for example, the average MSPE is $\sum_{i=1}^4 \|Y^{V_i} - \widehat{Y}^{V_i}\|_2^2/n_i$, where Y^{V_i} is the ith subset of validation data, $\widehat{Y}^{V_i}$ is the corresponding vector of predictions, and n_i is the size of the validation subset.

The cross-validation results in Table 2 suggest that, for every model (and, actually, for every validation subset; see representative results in Tables 4 and 5 in Section A.3 of the online supplements [Bliznyuk et al. (2014)]) inclusion of the multiday data through the linearized model—in order to increase spatial coverage—always improves the predictive performance relative to the corresponding model without multiday data. In particular, it can be seen from Figures 8–11 in the online supplements [Bliznyuk et al. (2014)] that models with long-range interaction term g_{ST} ($GST = 1$) do not perform well near the boundaries of the study region if the model for Y^A is excluded ($A = 0$).

The two best models are $M(U = 1, GST = 0, A = 1)$ and $M(U = 1, GST = 1, A = 1)$. With an exception of one station where the model $M(U = 1, GST = 1, A = 1)$ overpredicts, predictions from the two models are very similar, suggesting that inclusion of the long-range spatio-temporal interaction is not helpful for prediction given the observations available. It is notable that, for the better models, the empirical prediction interval coverage is close to the nominal 95%. The small difference of 1–2% from the nominal coverage could be due to holding the values of the parameters fixed at the estimated values. Failure to include the short-range dependence term u appears to result in underestimation of the prediction error variance and, consequently, narrower intervals with below-nominal coverage.

We also compared the predictive performance of our linearized models and the corresponding "simple models" based on equation (20) proposed by a reviewer using the MSPE and the correlation between held-out data and predictions. For each validation subset, our linearized models outperformed the models of equation (20). Surprisingly, the naïve linearization of the "simple models" occasionally caused the predictive performance to deteriorate, relative to the corresponding models without the multiday data. Our findings are fully described in Section A.3.2 in the online supplements [Bliznyuk et al. (2014)].

Validating the model on a spatially and temporally disjoint subset of data (online supplements [Bliznyuk et al. (2014)], Section A.3), which is indicative of the models' out-of-sample prediction performance, yielded the same choice of best model and the same conclusion that incorporation of the aggregated data via linearization uniformly improves the quality of predictions.

4.3. *Assessment of the adequacy of linearization.* Here we assess the impact of linearization on Bayesian inference using models that include the multiday data based on the results of Section 4.2. We compare nonlinear and linearized versions of model $M(U = 0, GST = 0, A = 1)$ of Table 2 because the models with ($U = 1$) and/or ($GST = 1$) are computationally less tractable.

Sampling from the linearized model was discussed in Section 3. To sample from $[\gamma|Y^{OIA}]$ under the linearized model, we initialized two Markov chains in a neighborhood of the mode of $[\gamma|Y^{OIA}]$, sampling as discussed in Section 3 for 75,000 iterations. The chains mixed well, with lag-one correlations in the component-wise chains $\{\gamma_j^{(i)}\}_i$ and $\{\log([\gamma^{(i)}|Y^{OIA}])\}_i$ around 0.95; lag-one correlations between the corresponding components of v are of much smaller magnitude, typically between 0.2 and 0.3. A burn-in sample of 2500 states was discarded from each chain.

To draw samples under the nonlinear model, we first reduced the dimension of the posterior by analytically integrating out the vector α_0. We sampled from $[\gamma, w|Y^{OIA}]$ using the adaptive RWMH sampler discussed in Section 3. Here, we drew $\{w, \gamma\}$ in a single step when sampling from $[w, \gamma|Y^{OIA}]$. This Markov chain mixes very slowly, with typical lag-one correlations in $\{\log([\gamma^{(i)}, w^{(i)}|Y^{OIA}])\}_i$ on the order of 0.995. We used 6 Markov chains, each of length 200,000, initialized in the high probability region of $[\gamma, w|Y^{OIA}]$. A burn-in sample of 25,000 states was discarded from each chain. Based on the effective sample size calculations, the Markov chain based on the nonlinear model is about 10 times less efficient than the one based on the linearized model.

Estimates of the marginal posterior densities of γ and w are shown in Figure 2 and Figure 6 in the online supplements [Bliznyuk et al. (2014)]. The marginal densities of elements of γ and w are remarkably similar between the nonlinear and linearized models, with the exception of the densities of $\gamma_2 = \alpha_{1A}$; these are still close to one another. Plots of spatial predictions—obtained as means of the posterior predictive distribution—for the two models (Figure 7 in the online supplements [Bliznyuk et al. (2014)]) are also visually indistinguishable, which provides further support for the use of linearization. The correlation between the spatial predictions under the two models is 0.9999 on both the logarithmic and original scale. We also examined the distribution and spatial variability of the pointwise

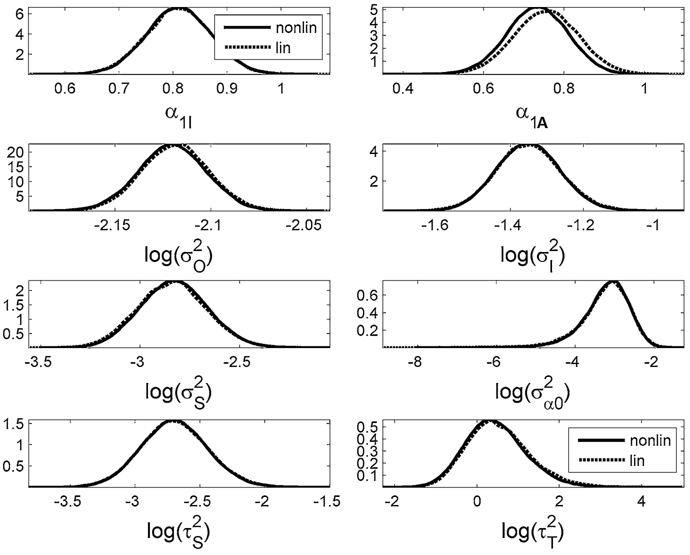

FIG. 2. *Estimates of marginal densities of nonlinear parameters under nonlinear* (*solid*) *and linearized* (*dashed*) *versions of the model* $M(U = 0, GST = 0, A = 1)$.

prediction differences. The variability of the differences tends to increase with the distance from the central site monitor (as expected due to the curse of dimensionality), but the predictions are still very close to each other. The relative accuracy of predictions on the original scale, computed pointwise as the absolute value of the differences of the predictions divided by the predicted value under the nonlinear model, is very high. For example, the 90th, 95th, 99th and 99.5th percentiles for the empirical distribution of the relative errors were 0.011, 0.014, 0.033 and 0.040, respectively.

4.4. *Bayesian inference and prediction.* In this section we report results under the model chosen in Section 4.2, which includes the short-range process, u, and the multiday data but excludes the long-range g_{ST} process.

To sample from $[\gamma|Y^{OIA}]$ using the computational strategy of Section 3, we launched four Markov chains, initialized in the region of high posterior probability of γ. Each chain had a length of 12,500, and a burn-in sample of size 2500 was discarded from each. We examined trace plots of MCMC states and the corresponding posterior density estimates to determine that the chains mixed rapidly and converged to the same posterior.

Posterior means and quantiles for γ are given in Table 3. Even though parameters α_{1I} and α_{1A} have similar interpretations, α_{1A} is smaller in magnitude than α_{1I}. This suggests that multiday indoor data are less informative for daily predictions of the outdoor exposure process than daily indoor data. This is plausible because readings from 30 out of 45 BCI sites overlap spatially and temporally with those from BCO sites, whereas all BCA sites are spatially and often temporally disjoint

TABLE 3
Posterior summaries of nonlinear parameters under model $M(U = 1, GST = 0, A = 1)$

Parameter	**Mean**	**2.5%**	**50%**	**97.5%**
α_{1I}	0.956	0.870	0.957	1.040
α_{1A}	0.698	0.576	0.702	0.817
σ_O^2	0.045	0.041	0.045	0.049
σ_I^2	0.129	0.101	0.125	0.161
σ_S^2	0.037	0.021	0.035	0.056
$\sigma_{\alpha 0}^2$	0.030	0.011	0.024	0.057
τ_S^2	0.030	0.016	0.027	0.047
τ_T^2	2.774	0.182	0.974	7.168
σ_u^2	0.098	0.090	0.098	0.106
θ_S	0.054	0.043	0.053	0.066
θ_T	0.120	0.073	0.120	0.169

from the BCO sites. Consequently, the spatio-temporal mismatch causes measurement error in the regressor (the latent process here), and the coefficients are shrunk more toward zero whenever there is more error in the covariate. The temporal decay parameter $\theta_T \approx 0.12$, interpreted in light of the tapering structure, corresponds to a model with temporal correlation function that is about 0.7 at lag one but decays faster than that of the $AR(1)$ process with lag-one correlation of 0.7. The spatial decay parameter $\theta_S \approx 0.054$ (on the 1 km distance scale) corresponds to spatial correlation that decays to 0.05 by about 35 km. Consequently, when predicting within the temporal range of measurements, the short-range process, u, "pulls" the predictions toward the observed data, thereby capturing the nonperiodic features of the exposure process not accounted for by ζ.

Posterior means and quantiles for the coefficients of the observable covariates are reported in Table 1. Based on preliminary exploratory analysis using only the outdoor data, the logarithm of readings from the central site monitor (`logHSPH`) was the most important covariate for spatio-temporal prediction. From the Bayesian model fit using data from all sources, this conjecture was further supported by the relative widths and quantiles of the credible intervals. The effect of other temporally-varying covariates such as wind speed and the planetary boundary layer is not easily interpretable in the presence of interactions of spatial and temporal covariates. However, certain two- and three-way interactions have been shown to add to the predictive ability of other prediction models in the Boston area [Zanobetti et al. (2014)]. The spatially-varying population and land use covariates are positively associated with the response. The traffic density covariate is of most interest because of the relationship between black carbon and traffic that motivates this work. Its marginal effect—once the interactions with temporally varying covariates have been accounted for—is positive, which can be clearly seen in Figure 3, in which predictions follow the road network. In the early phase of this project we considered models with fewer predictors and without interactions, which yielded a similar relative ranking of the models and slightly less accurate predictions.

The primary goal of our work is to predict a vector of latent process values, η^P, in the region for any temporal period of interest for health effects analysis, which is done using $E(\eta^P|Y^{OIA})$. Because of the temporal covariance tapering, if the minimum distance between the temporal indices in η^P and in Y^{OIA} exceeds the range of the taper function, then $E(\eta^P|Y^{OIA}) = E(C^P w|Y^{OIA})$, where C^P is the "design matrix" for η^P. Figure 3 shows predictions for an example day (July 31, 2006) based on the MCMC estimate of $E(\eta^P|Y^{OIA})$.

5. Discussion. In this paper we developed a unified exposure prediction framework that aggregates air pollutant concentration data from multiple disparate sources that are available at different levels of temporal resolution, which is of great importance for health effects models arising in environmental science. We

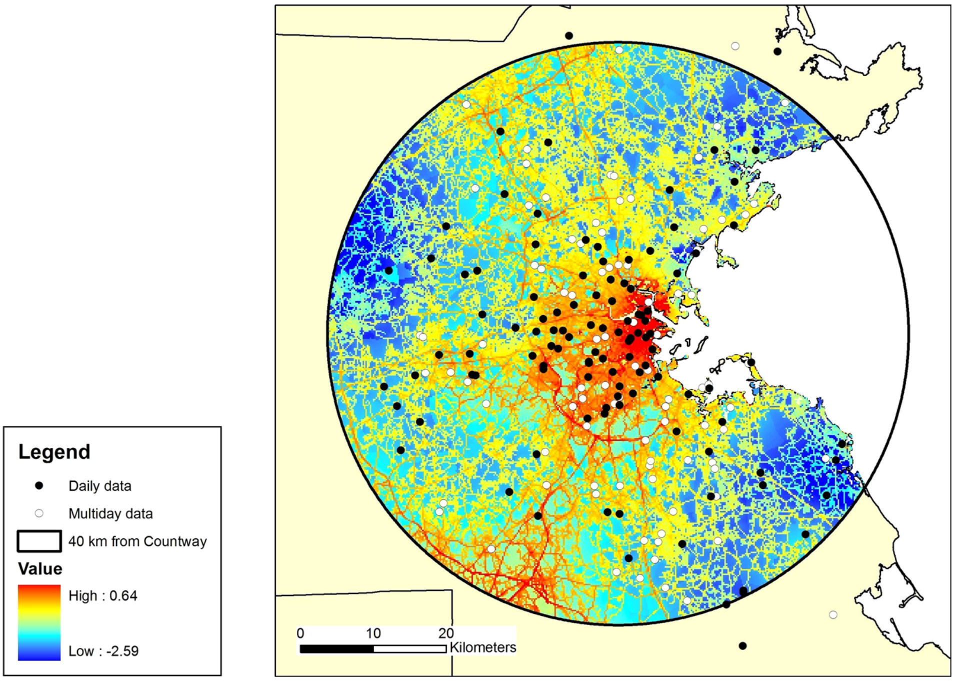

FIG. 3. *Log black carbon predictions for July* 31, 2006 *based on the mean of the predictive distribution*, $E(\eta^P|Y^{OIA})$, *under the final model* $M(U = 1, GST = 0, A = 1)$. *The unit is the natural logarithm of* $\mu g/m^3$.

found that incorporation of even a modest number of observations (93 or under 1.5% of the overall observation count) of the multiday data from a relatively large spatial network (roughly doubling the number of unique spatial sites) uniformly improves the prediction quality in a number of models that may or may not include long-term and short-term spatio-temporal signal. To our surprise, incorporation of a periodic long-range spatio-temporal trend did not produce considerable improvements over the models without the long-range interaction. We attribute this to the fact that the air monitors in the networks corresponding to each study are scattered in space and operate irregularly in time, which implies that the observed data correspond to under 5% of the dates from all the monitors over the whole study period. In our models, the departures from the periodic trend are being captured by the short-range process. If the temporal coverage were richer, we would be able to identify the nonperiodic component of long-range variability better and to rigorously test its presence.

Our linearization approach provides a computationally efficient means to build two quadratic approximations: (i) the logarithm of $[\eta^A|\upsilon, \gamma, Y^{OIA}]$ and (ii) the logarithm of $[\eta^A, \eta^P|\upsilon, \gamma, Y^{OIA}]$, where η^P is a vector of latent process values we want to predict. This produces a linearized model with Gaussian approximations for the marginal likelihood and required conditional posterior densities. An alternative that also results in Gaussian approximations is to approximate (i) and (ii) using a two-term Taylor expansion about the appropriate modes, which needs to be located by a costly optimization run. Using these approximations to integrate out η^A is equivalent to the Laplace approximation [Tierney and Kadane (1986)]. The downside of this scheme is that the approximation needs to be built for every value of (υ, γ) of interest, which is infeasible in practice.

While we adopt an MCMC-based approach to Bayesian inference and prediction, a promising direction for future work is to consider an approximation scheme in the spirit of Rue, Martino and Chopin (2009), the integrated nested Laplace approximation (INLA). Methodologically, one will need to address the following two issues that are critical to the computational performance of INLA for *inference* in latent process models that combine multiple data sets, both in our case study and in general. First, one needs to be able to enforce the Markov property of the spatio-temporal latent process. Second, increasing the number of data sets in the joint model (that are linked by the latent process) and the complexity of the model for the latent process adds to the dimension of the hyperparameter vector γ [$\boldsymbol{\theta}$ in Rue, Martino and Chopin (2009)], which can make accurate numerical integration computationally demanding, if feasible.

Acknowledgment. We thank Steven Melly for help with figures. The contents of this publication are solely the responsibility of the grantee and do not necessarily represent the official views of the U.S. EPA. Further, the U.S. EPA does not endorse the purchase of any commercial products or services mentioned in this publication.

SUPPLEMENTARY MATERIAL

Supplement to "Nonlinear predictive latent process models for integrating spatio-temporal exposure data from multiple sources" (DOI: 10.1214/14-AOAS737SUPP; .pdf). Online supplements contain technical details and supplementary figures and tables.

REFERENCES

ADAR, S. D., KLEIN, R., KLEIN, B. E. K., SZPIRO, A. A., COTCH, M. F., WONG, T. Y., O'NEILL, M. S., SHRAGER, S., BARR, R. G., SISCOVICK, D. S., DAVIGLUS, M. L., SAMPSON, P. D. and KAUFMAN, J. D. (2010). Air pollution and the microvasculature: A cross-sectional assessment of in vivo retinal images in the population-based multi-ethnic study of atherosclerosis (MESA). *PLOS Medicine* **7** e1000372.

BANERJEE, S., CARLIN, B. P. and GELFAND, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall, Boca Raton, FL.

BERHANE, K., GAUDERMAN, W. J., STRAM, D. O. and THOMAS, D. C. (2004). Statistical issues in studies of the long-term effects of air pollution: The Southern California children's health study. *Statist. Sci.* **19** 414–449. MR2185625

BLIZNYUK, N., RUPPERT, D. and SHOEMAKER, C. A. (2011). Efficient interpolation of computationally expensive posterior densities with variable parameter costs. *J. Comput. Graph. Statist.* **20** 636–655. MR2878994

BLIZNYUK, N., PACIOREK, C. J., SCHWARTZ, J. and COULL, B. (2014). Supplement to "Nonlinear predictive latent process models for integrating spatio-temporal exposure data from multiple sources." DOI:10.1214/14-AOAS737SUPP.

CALDER, C. A. (2007). Dynamic factor process convolution models for multivariate space–time data with application to air quality assessment. *Environ. Ecol. Stat.* **14** 229–247. MR2405328

CALDER, C. A. (2008). A dynamic process convolution approach to modeling ambient particulate matter concentrations. *Environmetrics* **19** 39–48. MR2416543

CHRISTENSEN, O. F., ROBERTS, G. O. and SKÖLD, M. (2006). Robust Markov chain Monte Carlo methods for spatial generalized linear mixed models. *J. Comput. Graph. Statist.* **15** 1–17. MR2269360

CHRISTENSEN, O. F. and WAAGEPETERSEN, R. (2002). Bayesian prediction of spatial count data using generalized linear mixed models. *Biometrics* **58** 280–286. MR1908167

FUENTES, M. and RAFTERY, A. E. (2005). Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. *Biometrics* **61** 36–45. MR2129199

FURRER, R., GENTON, M. G. and NYCHKA, D. (2006). Covariance tapering for interpolation of large spatial datasets. *J. Comput. Graph. Statist.* **15** 502–523. MR2291261

GELFAND, A., ZHU, L. and CARLIN, B. (2001). On the change of support problem for spatio-temporal data. *Biostatistics* **2** 31–45.

GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. MR2345548

GNEITING, T., ŠEVČÍKOVÁ, H. and PERCIVAL, D. B. (2012). Estimators of fractal dimension: Assessing the roughness of time series and spatial data. *Statist. Sci.* **27** 247–277. MR2963995

GOLUB, G. H. and VAN LOAN, C. F. (1996). *Matrix Computations*, 3rd ed. Johns Hopkins Univ. Press, Baltimore, MD. MR1417720

GOTWAY, C. A. and YOUNG, L. J. (2002). Combining incompatible spatial data. *J. Amer. Statist. Assoc.* **97** 632–648. MR1951636

GOTWAY, C. A. and YOUNG, L. J. (2007). A geostatistical approach to linking geographically aggregated data from different sources. *J. Comput. Graph. Statist.* **16** 115–135. MR2345750

GRYPARIS, A., COULL, B. A., SCHWARTZ, J. and SUH, H. H. (2007). Semiparametric latent variable regression models for spatiotemporal modelling of mobile source particles in the greater Boston area. *J. Roy. Statist. Soc. Ser. C* **56** 183–209. MR2359241

GRYPARIS, A., PACIOREK, C. J., ZEKA, A., SCHWARTZ, J. and COULL, B. A. (2009). Measurement error caused by spatial misalignment in environmental epidemiology. *Biostatistics* **10** 258–274.

HAARIO, H., SAKSMAN, E. and TAMMINEN, J. (2001). An adaptive Metropolis algorithm. *Bernoulli* **7** 223–242. MR1828504

JANSSEN, N. A. H., HOEK, G., SIMIC-LAWSON, S., FISCHER, P., VAN BREE, L., TEN BRINK, H., KEUKEN, M., ATKINSON, R. W., ANDERSON, H. R., BRUNEKREEF, B. and CASEE, F. R. (2011). Black carbon as an additional indicator of the adverse health effects of airborne particles compared with PM10 and PM2.5. *Environ. Health Perspect.* **119** 1691–1699.

OPSOMER, J., WANG, Y. and YANG, Y. (2001). Nonparametric regression with correlated errors. *Statist. Sci.* **16** 134–153. MR1861070

ROBERT, C. P. and CASELLA, G. (1999). *Monte Carlo Statistical Methods*. Springer, New York. MR1707311

RUE, H. and HELD, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. *Monographs on Statistics and Applied Probability* **104**. Chapman & Hall, Boca Raton, FL. MR2130347

RUE, H., MARTINO, S. and CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71** 319–392. MR2649602

RUPPERT, D., WAND, M. P. and CARROLL, R. J. (2003). *Semiparametric Regression*. Cambridge Univ. Press, Cambridge. MR1998720

TIERNEY, L. and KADANE, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.* **81** 82–86. MR0830567

VAN DYK, D. A. and PARK, T. (2008). Partially collapsed Gibbs samplers: Theory and methods. *J. Amer. Statist. Assoc.* **103** 790–796. MR2524010

WANG, Y. (1998). Smoothing spline models with correlated random errors. *J. Amer. Statist. Assoc.* **93** 341–348.

WANNEMUEHLER, K. A., LYLES, R. H., WALLER, L. A., HOEKSTRA, R. M., KLEIN, M. and TOLBERT, P. (2009). A conditional expectation approach for associating ambient air pollutant exposures with health outcomes. *Environmetrics* **20** 877–894. MR2838493

WOOD, S. N. (2006). *Generalized Additive Models*: *An Introduction with R*. Chapman & Hall, Boca Raton, FL. MR2206355

ZANOBETTI, A., COULL, B. A., GRYPARIS, A., SPARROW, D., VOKONAS, P. S., WRIGHT, R. O., GOLD, D. R. and SCHWARTZ, J. (2014). Associations between arrhythmia episodes and temporally and spatially resolved black carbon and particulate matter in elderly patients. *Occup. Environ. Med.* **71** 201–207.

ZEGER, S. L., THOMAS, D., DOMINICI, F., SAMET, J. M., SCHWARTZ, J., DOCKERY, D. and COHEN, A. (2000). Exposure measurement error in time-series studies of air pollution: Concepts and consequences. *Occup. Environ. Med.* **108** 419–426.

N. BLIZNYUK
DEPARTMENT OF AGRICULTURAL
AND BIOLOGICAL ENGINEERING
UNIVERSITY OF FLORIDA
406 MCCARTY HALL C
GAINESVILLE, FLORIDA 32611
USA
E-MAIL: nbliznyuk@ufl.edu

C. J. PACIOREK
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
367 EVANS HALL
BERKELEY, CALIFORNIA 94720
USA
E-MAIL: paciorek@stat.berkeley.edu

J. SCHWARTZ
DEPARTMENT OF ENVIRONMENTAL HEALTH
HARVARD SCHOOL OF PUBLIC HEALTH
665 HUNTINGTON AVENUE
LANDMARK CENTER ROOM 415
BOSTON, MASSACHUSETTS 02115
USA
E-MAIL: jschwrtz@hsph.harvard.edu

B. COULL
DEPARTMENT OF BIOSTATISTICS
HARVARD SCHOOL OF PUBLIC HEALTH
655 HUNTINGTON AVENUE
BOSTON, MASSACHUSETTS 02115
USA
E-MAIL: bcoull@hsph.harvard.edu