Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

*Biometrics*. Author manuscript; available in PMC 2015 May 25.

Published in final edited form as:

*Biometrics*. 2012 September ; 68(3): 837–848. doi:10.1111/j.1541-0420.2011.01725.x.

# Space-Time Data fusion Under Error in Computer Model Output: An Application to Modeling Air Quality

**Veronica J. Berrocal**[1,*], **Alan E. Gelfand**[2,**], and **David M. Holland**[3,***]

[1]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A

[2]Department of Statistical Science, Duke University, Durham, North Carolina 27708, U.S.A

[3]U.S. Environmental Protection Agency, National Exposure Research Laboratory, Research Triangle Park, North Carolina 27711, U.S.A

## Summary

We provide methods that can be used to obtain more accurate environmental exposure assessment. In particular, we propose two modeling approaches to combine monitoring data at point level with numerical model output at grid cell level, yielding improved prediction of ambient exposure at point level. Extending our earlier downscaler model (Berrocal, V. J., Gelfand, A. E., and Holland, D. M. (2010b). A spatio-temporal downscaler for outputs from numerical models. *Journal of Agricultural, Biological and Environmental Statistics* **15,** 176–197), these new models are intended to address two potential concerns with the model output. One recognizes that there may be useful information in the outputs for grid cells that are neighbors of the one in which the location lies. The second acknowledges potential spatial misalignment between a station and its putatively associated grid cell.

The first model is a Gaussian Markov random field smoothed downscaler that relates monitoring station data and computer model output via the introduction of a latent Gaussian Markov random field linked to both sources of data. The second model is a smoothed downscaler with spatially varying random weights defined through a latent Gaussian process and an exponential kernel function, that yields, at each site, a new variable on which the monitoring station data is regressed with a spatial linear model. We applied both methods to daily ozone concentration data for the Eastern US during the summer months of June, July and August 2001, obtaining, respectively, a 5% and a 15% predictive gain in overall predictive mean square error over our earlier downscaler model (Berrocal et al., 2010b). Perhaps more importantly, the predictive gain is greater at hold-out sites that are far from monitoring sites.

© 2011, The International Biometric Society

*berrocal@umich.edu
**alan@stat.duke.edu
***Holland.David@epa.gov

Supplementary Materials: Web Figures referenced in Sections 1 and 5 are available under the article information link at the *Biometrics* website http://www.biometrics.tibs.org.

**Disclaimer**: The U.S. Environmental Protection Agency through its Office of Research and Development partially collaborated in this research. Although it has been reviewed by the Agency and approved for publication, it does not necessarily reflect the Agency's policies or views.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

**Keywords**

Change of support; Data fusion; Gaussian Markov random field; Numerical model calibration; Smoothing; Spatially varying random weights

## 1. Introduction

The need for accurate assessment of exposure to air pollutants arises to effectively investigate the linkage between ambient exposure and health effects. It also arises with regard to compliance with legislated regulatory standards to control levels of environmental exposure. As a result, the US Environmental Protection Agency (EPA) monitors pollutant levels using information from monitoring networks as well as estimates generated by deterministic numerical models. The former measure pollutant concentrations using instruments at a sparse set of stations, while the latter yield estimates of the average concentration in grid cells of prespecified dimensions by numerically solving complex systems of differential equations capturing various diffusion, chemical, and atmospheric processes. The computer model output spans large spatial domains with no missingness.

Fusing these information sources can improve exposure assessment at high, in fact, point-level resolution. Combining data from multiple sources, so-called data assimilation, is well-known in the atmospheric sciences (Kalnay, 2003). There, the goal is to combine observational data on the current state of the atmosphere with a short-range forecast in order to obtain initial conditions for a numerical atmospheric model. Most methods proposed in atmospheric data assimilation are algorithmic, ad hoc and do not address the “change of support” problem (Cressie 1993; Gotway and Young 2002; Banerjee, Carlin, and Gelfand 2004, chapter 6).

The statistics literature on “data fusion” can be grouped into two paths. One is Bayesian melding (Fuentes and Raftery 2005) where observational data is combined with computer model output by introducing a latent point-level process driving both sources of data. The numerical model output is then expressed as a *linearly* calibrated integral over a grid cell (scaled by the area of the cell) of the latent point-level process while the monitoring data is related to the latent process via a measurement error model. A spatio-temporal extension of Bayesian melding has been presented by McMillan et al. (2010). This approach offers a solution to the problem through upscaling to grid cells.

The second approach uses a two-stage regression, dating to Guillas et al. (2008) and subsequently Liu, Le, and Zidek (2008), with an ad hoc method to allow the coefficients of the linear regression to be spatially interpolated. Berrocal, Gelfand, and Holland (2010a, 2010b) propose univariate and bivariate hierarchical downscaler models that relate the monitoring station data and the computer model output using a spatial linear model with spatially varying coefficients in turn modeled as Gaussian processes (GPs). These models offer the advantage of local calibration of the numerical model output without incurring in problems due to the dimensionality of the computer model output, as for example in Bayesian melding, since they are only fitted at the numerical model grid cells where the monitoring stations reside.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The contribution of this article is to provide two useful neighbor-based extensions of our earlier downscaler modeling work. That is, there may be useful information in the output at neighboring grid cells to the one where the location lies and there may be misalignment between stations and putatively associated grid cells. These extensions do not seek directly to remedy other forms of error that may be built into the numerical model (errors reflecting uncertainty in model input, uncertainty in the partial differential equations for the dynamics of pollution transport, and uncertainty introduced by the numerical approximation methods used to solve the resulting system of partial differential equations). In particular, these new models provide adaptive smoothing to the computer model output which achieves stronger association with the observed station data. Improved spatial interpolation of the ambient exposures results; using hold out data, we achieve gains of 5% and 15%, respectively, in predictive mean square error over our original downscaler. One extension introduces a Gaussian Markov random field (GMRF) to smooth the computer model. The other introduces spatially varying weights driven by a latent GP to accomplish the smoothing. This last model falls in the realm of recent work to render conditionally autoregressive models (CAR; Besag 1974; Banerjee et al. 2004) more flexible by allowing adaptive adjacency structure (e.g. Lu and Carlin 2005; Kyung and Ghosh 2010).

We apply our approach to interpolate ozone levels in space and time for the summer of 2001 using station data and the Community Multiscale Air Quality model (CMAQ; Byun and Schere 2006) output. However, the strategy is applicable to other environmental contaminants and to other data fusion settings.

The format of the article is as follows. In Section 2, we present the motivating data. In Section 3, we review the elementary downscaler model (Berrocal et al., 2010b) and introduce two extensions. In Section 4, we discuss computation details relative to the fitting of these downscaler models, while in Section 5, we present results on the predictive performance of these models. We conclude with Section 6 where we briefly discuss future extensions of the smoothed downscaler with spatially varying random weights. Supplementary material including additional figures and results is available online at the Biometrics website (http://www.biometrics.tibs.org).

## 2. Data

Ground-level ozone is one of the six “criteria pollutants” that the US EPA is required to monitor by the Clean Air Act. To keep track of ozone concentration, the EPA utilizes monitoring devices sparsely located across the United States along with estimates of ground-level ozone concentration produced by the deterministic numerical air quality model, Models-3/ Community Mesoscale Air Quality model, CMAQ (Byun and Schere 2006; http://epa.gov/asmdnerl/CMAQ). We illustrate our fully model based fusion approach using these two sources. In both cases, the data refer to the daily 8-hour maxima ozone concentration (henceforth daily concentration) for the Eastern United States during the summer months of June, July, and August 2001, when the elevated temperatures and solar radiation exacerbate the production of ozone.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Figure 1 displays the locations of the 800 monitoring sites belonging to the National Air Monitoring Stations/State and Local Air Monitoring Stations (NAMS/SLAMS) network that we employed for our analysis. Of the 800 sites, we selected 700 at random as a fitting dataset (percentage of missing data during the 92 summer days of 2001 = 2.8%), while we used the remaining 100 (percentage of missing data = 2.2%) to assess the out-of-sample predictive performance of four approaches: (i) a kriging model using only the station data, (ii) the spatio-temporal downscaler model discussed above (Berrocal et al., 2010b), (iii) a downscaler model with GMRF smoothing, and (iv) a downscaler model with smoothing obtained using spatially varying random weights.

Following Berrocal et al. (2010b), Sahu, Gelfand, and Holland (2007) and references therein, we have modeled ozone concentration on the square root scale to achieve approximate normality and stabilize the variance. The observed ozone concentration data display a fair degree of variability during the summer months of 2001: the daily mean of the daily ozone concentration over the 800 monitoring sites, ranges from 5.8 to 8.5 √ppb (parts per billion), while the daily standard deviation varies from 0.7 to 1.6 √ppb (Figure 2a).

CMAQ produces estimates of ozone concentration over the United States at predetermined spatial and temporal scales. We used CMAQ for the eastern United States at 12-km resolution yielding 40,044 daily grid cell values. Although at areal scale rather than at point level, the CMAQ output has the advantage of complete spatial coverage and no missingness. In addition, it is moderately to fairly strongly correlated with the observed ozone concentration data, suggesting its potential for improved spatial interpolation. In particular, Figure 2(b) displays the daily correlation between the square root of observed ozone concentration at site **s** and the square root of the CMAQ output at the grid cell *B* that contains **s**.

## 3. Modeling

We briefly review the downscaler model of Berrocal et al. (2010b) and then we present the two promising specifications that extend the downscaler model to improve the fusion in light of concerns mentioned in the Introduction. For each model, we first present its static version and then its spatio-temporal formulation.

### 3.1 Static Setting

**3.1.1 The univariate downscaler—**Let $Y(\mathbf{s})$ denote the square root of the daily ozone concentration observed at site **s**, and let $x(B)$ indicate the square root of the daily ozone concentration predicted by CMAQ over grid cell *B*.

In the downscaler model, the *change of support* problem is addressed by relating $Y(\mathbf{s})$ to the CMAQ output, $x(B)$, at the grid cell *B* that contains **s**, via the model

$$Y(\mathbf{s})=\tilde{\beta}_0(\mathbf{s})+\tilde{\beta}_1(\mathbf{s})x(B)+\varepsilon(\mathbf{s}),\ \varepsilon(\mathbf{s})\overset{ind}{\sim}\mathrm{N}(0,\tau^2),\quad (1)$$

where $\varepsilon(\mathbf{s})$ is a white noise process with nugget variance $\tau^2$, and $\tilde{\beta_0}(\mathbf{s})$ and $\tilde{\beta_1}(\mathbf{s})$ are spatially varying coefficients that can be decomposed as

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\begin{aligned}\tilde{\beta}_0(\mathbf{s}) &= \beta_0 + \beta_0(\mathbf{s}) \\ \tilde{\beta}_1(\mathbf{s}) &= \beta_1 + \beta_1(\mathbf{s}),\end{aligned} \quad (2)$$

with $\beta_0$ and $\beta_1$, respectively, the overall intercept and slope in calibrating the CMAQ model output and $\beta_0(\mathbf{s})$ and $\beta_1(\mathbf{s})$, respectively, the local adjustments to these terms. Anticipating association between $\beta_0(\mathbf{s})$ and $\beta_1(\mathbf{s})$, the two spatially varying coefficients are in turn modeled as correlated mean-zero Gaussian spatial processes using the method of coregionalization (Wackernagel 2003; Gelfand et al. 2004). Thus, we assume two mean-zero unit-variance independent GPs $v_0(\mathbf{s})$ and $v_1(\mathbf{s})$ each, for convenience, equipped with an exponential covariance structure having decay parameters, respectively, $\varphi_0$ and $\varphi_1$, such that

$$\begin{pmatrix} \beta_0(\mathbf{s}) \\ \beta_1(\mathbf{s}) \end{pmatrix} = \mathbf{A} \begin{pmatrix} v_0(\mathbf{s}) \\ v_1(\mathbf{s}) \end{pmatrix}, \quad (3)$$

where the unknown **A** matrix in (3) can be assumed, without loss of generality, to be lower-triangular. To complete the hierarchical specification, we need to provide prior distributions for the overall bias terms, $\beta_0$ and $\beta_1$, the nugget variance, $\tau^2$, the three nonnull elements of the coregionalization matrix **A**, and the decay parameters $\varphi_0$ and $\varphi_1$.

The downscaler model "fuses" the two sources of data while avoiding problems due to the large number of grid cells (>40,000) associated with the CMAQ model output since only numerical model grid cells with monitoring station observations are used in fitting. As a result, we can address the change of support problem while avoiding introduction of stochastic integrals, needed in the scaling up associated with Bayesian melding. Such integrals render the latter models computationally challenging to fit for a large number of grid cells and hopeless when we introduce measurements over time.

Comparison with ordinary kriging has shown that the downscaler model yields better *in* and *out-of-sample* predictive performance. Lower predictive mean square and absolute value errors along with predictive intervals having coverage close to the nominal value result.

**3.1.2 A GMRF smoothed downscaler—**In (1) the CMAQ model output enters as a covariate. Here, we introduce a smoothed version of the $\{x(B)\}$ surface arising from a latent Gaussian Markov random field (GMRF), i.e., let

$$x(B) = \mu + V(B) + \eta(B) \;\; \eta(B) \overset{ind}{\sim} N(0, \sigma^2) \quad (4)$$

with $\mu$ an overall mean and $V(B)$ a mean-zero Gaussian Markov random field equipped with a conditionally autoregressive structure (CAR; Besag 1974; Banerjee et al. 2004). In other words, if $g$ is the number of numerical model grid cells, then we assume that

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$V(B_i)|\{V(B_j):j \neq i\} \sim N\left(\sum_{j\in\delta B_i}\frac{V(B_j)}{m_i},\frac{\xi^2}{m_i}\right)\ i=1,\ldots,g \quad (5)$$

where $\delta B_i$ denotes the set of grid cells that are neighbors to $B_i$. Although the joint distribution of the $\{V(B_i)\}$ is improper, the joint distribution for the $\{x(B_i)\}$ given the $V$'s is proper and so we have a valid model for the *data* $\{x(B_i)\}$. Also, the GMRF specification makes it clear that $\{\tilde{V}(B)\colon \tilde{V}(B) = \mu + V(B)\}$ is a smoothed version of $\{x(B)\}$. Hence, for $\mathbf{s} \in B$, we revise (1) to

$$Y(\mathbf{s})=\tilde{\beta}_0(\mathbf{s})+\beta_1\tilde{V}(B)+\varepsilon(\mathbf{s})\ \varepsilon(\mathbf{s}) \sim N(0,\tau^2) \quad (6)$$

where, again, $\tilde{\beta_0}(\mathbf{s}) = \beta_0 + \beta_0(\mathbf{s})$ with $\beta_0(\mathbf{s})$ modeled as a mean-zero GP with exponential covariance structure, decay parameter $\varphi_0$ and marginal variance $\sigma^2_{\beta_0}$.[1]

Differently from the downscaler model, with the GMRF smoothed downscaler model we sacrifice dimension reduction. Though (6) is still fitted only at those grid cells $B$ with observations, (4) with (5) requires the entire latent field $V(B)$. Fortunately, the computation associated with a GMRF is *local* so we can still fit this model efficiently. Also, (6) through (4) and (5) clarifies that we are implicitly relating the $Y(\mathbf{s})$ with the CMAQ model output at all the grid cells that are neighbors of the grid cell $B$ containing $\mathbf{s}$.

**3.1.3 A smoothed downscaler using spatially varying random weights—**Here, we introduce smoothing using weights that are random and spatially varying. Now, we regress the observation at site $\mathbf{s}$, $Y(\mathbf{s})$, on a point-level regressor, $x(\tilde{\mathbf{s}})$, obtained by creating, at each site $\mathbf{s}$, a weighted average of all the numerical model output with weights that are site-specific.

We replace (6) with

$$Y(\mathbf{s})=\tilde{\beta}_0(\mathbf{s})+\beta_1\tilde{x}(\mathbf{s})+\varepsilon(\mathbf{s})\ \varepsilon(\mathbf{s}) \sim N(0,\tau^2) \quad (7)$$

where $\tilde{\beta_0}(\mathbf{s})$ is as in (6) and

$$\tilde{x}(\mathbf{s})=\sum_{k=1}^{g}w_k(\mathbf{s})x(B_k). \quad (8)$$

The weights $w_k(\mathbf{s})$ are in turn defined as follows: let $\mathbf{r}_k$, $k = 1,\ldots, g$ be the centroids of the $g$ numerical model grid cells, and let $Q(\mathbf{r})$ be a mean-zero GP having exponential covariance function with decay parameter $\varphi_Q$ and marginal variance $\sigma^2_Q$. Then, given $\{Q(\mathbf{r}_k)\}$, the $g$-dimensional random vector of weights $\{w_k(\mathbf{s})\}_{k=1,\ldots,g}$ at $\mathbf{s}$ is given by

[1]With $\tilde{V}(B)$ unobserved, a spatially varying $\tilde{\beta_1}(\mathbf{s})$ will not be identifiable.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$w_k(\mathbf{s}) := \frac{\mathcal{K}(\mathbf{s}-\mathbf{r}_k;\psi)\cdot\exp(Q[\mathbf{r}_k])}{\sum_{l=1}^{g}\mathcal{K}(\mathbf{s}-\mathbf{r}_l;\psi)\cdot\exp(Q[\mathbf{r}_l])} \quad (9)$$

where $\mathcal{K}(\mathbf{s}-\mathbf{r}_k;\ \psi)$ is an exponential kernel with decay parameter $\psi$.

Expression (9) is reminiscent of the discretized version of process convolution introduced by Higdon (1998). However, that work developed a stochastic process model with covariance function induced by the kernel $\mathcal{K}$. Here, we are only interested in creating a spatially varying set of weights that are spatially dependent, positive and sum to 1. If we define the weights $w_k(\mathbf{s})$ without introducing the latent GP$Q(\mathbf{r})$ they would not be directional and would have the same circular contour when moving from site to site. From (9), the $Q(\cdot)$ process is not identified; its center is arbitrary. So, we impose a "sum to 0" constraint, implementing it *on the fly* in the model fitting. Further discussion regarding the effect on the weights due to $\psi$, $\varphi_Q$, and $\sigma_Q^2$ is provided in Section 4.

Evidently, we allow calibration in the association between the observational data at $\mathbf{s}$, $Y(\mathbf{s})$, and the *revised* numerical model output at the grid cell $B$ that contains $\mathbf{s}$. Also, we clearly relate $Y(\mathbf{s})$ to CMAQ levels at neighbors of the grid cell $\mathbf{s}$ belongs to. Moreover, as in the GMRF smoothed downscaler model, the collection $\{x(\tilde{\mathbf{r}}_k)\}_{k=1,\ldots,g}$ can be interpreted as a smoothed version of the CMAQ model output, analogous to the collection $\{\tilde{V}(B_k)\}_{k=1,\ldots,g}$.

### 3.2 Spatio-Temporal Modeling

We now extend these downscaler models to handle data collected over space and time.

**3.2.1 The univariate downscaler—**Let $t$ denote time with $t = 1,\ldots, T$, and let $Y(\mathbf{s}, t)$ denote the square root of the daily ozone concentration observed at site $\mathbf{s}$ and time $t$. Following Section 3.1.1, $x(B, t)$ is the square root of the CMAQ predicted daily average ozone concentration over grid cell $B$ at time $t$. As in the static setting, we associate to each point $\mathbf{s}$ the CMAQ grid cell $B$ in which it lies, and extend (1) to

$$Y(\mathbf{s},t) = \tilde{\beta}_0(\mathbf{s},t) + \tilde{\beta}_1(\mathbf{s},t)x(B,t) + \varepsilon(\mathbf{s},t), \quad (10)$$

where $\varepsilon(\mathbf{s},t) \overset{ind}{\sim} N(0,\tau^2)$. For each $t = 1,\ldots, T$, we decompose $\tilde{\beta}_i(\mathbf{s}, t)$, $i = 0, 1$ as the sum of an overall coefficient and a local adjustment to it, that is: $\tilde{\beta}_i(\mathbf{s}, t) = \beta_{i,t} + \beta_i(\mathbf{s}, t)$, $i = 0, 1$.

We consider two ways to introduce temporal dependence in the time varying parameters $\beta_{0,t}$, $\beta_{1,t}$, $\beta_0(\mathbf{s}, t)$, and $\beta_1(\mathbf{s}, t)$. The first is to assume that they are nested, i.e. they are independent across time; the second is to assume that they evolve dynamically in time (West and Harrison 1999). For the former, we would adopt $\beta_{i,t} \overset{ind}{\sim} N(\mu_i, \sigma_i^2), i=0,1$ while if they are dynamic, we would assume

$$\beta_{i,t} = \rho_i\beta_{i,t-1} + \varsigma_{i,t}\ \varsigma_{i,t} \overset{ind}{\sim} N(0,\varsigma_i^2)\ i=0,1. \quad (11)$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

If the $\beta_0(\mathbf{s}, t)$ and $\beta_1(\mathbf{s}, t)$ are assumed nested within time, then for each $t = 1,\ldots, T$, they are expressed as a linear combination of uncorrelated latent mean-zero unit-variance GPs $v_0(\mathbf{s}, t)$ and $v_1(\mathbf{s}, t)$ having exponential covariance functions with decay parameters, respectively $\varphi_{0,t}$ and $\varphi_{1,t}$, i.e. similar to (3),

$$\begin{pmatrix} \beta_0(\mathbf{s},t) \\ \beta_1(\mathbf{s},t) \end{pmatrix} = \mathbf{A} \begin{pmatrix} v_0(\mathbf{s},t) \\ v_1(\mathbf{s},t) \end{pmatrix} \quad (12)$$

with $\mathbf{A}$ coregionalization matrix and $v_0(\mathbf{s}, t)$ and $v_1(\mathbf{s}, t)$ independent replicates of two GPs. Conversely, if $\beta_0(\mathbf{s}, t)$ and $\beta_1(\mathbf{s}, t)$ evolve in time, then, following Gelfand, Banerjee, and Gamerman (2005), for each $t = 1,\ldots, T$, we could assume

$$\beta_i(\mathbf{s},t) = \gamma_i \beta_i(\mathbf{s},t-1) + \nu_i(\mathbf{s},t) \; i=0,1, \quad (13)$$

where the innovations $\nu_i(\mathbf{s}, t)$ are correlated mean-zero GPs defined as:

$$\begin{pmatrix} \nu_0(\mathbf{s},t) \\ \nu_1(\mathbf{s},t) \end{pmatrix} = \mathbf{A} \begin{pmatrix} v_0(\mathbf{s},t) \\ v_1(\mathbf{s},t) \end{pmatrix}$$

with the $v_i(\mathbf{s}, t)$ as above. In addition, for this model, we set $\beta_i(\mathbf{s}, 0) = 0$, for $i = 0, 1$. In both (12) and (13), we might envision $\mathbf{A} = \mathbf{A}_t$ with $\mathbf{A}_t$ independent across time.

With two different ways in which time dependence can be modeled for each of the time varying parameters, $\beta_{0,t}$, $\beta_{1,t}$, $\beta_0(\mathbf{s}, t)$, and $\beta_1(\mathbf{s}, t)$, we can formulate four different versions of the spatio-temporal downscaler model. In experiments carried out with ozone concentration data for 2001 (Berrocal et al. 2010b), the spatio-temporal downscaler model with all time varying parameters nested within time yielded the best predictive performance. The flexibility to choose daily decay parameters is *better* than the introduction of autoregressive structure in the $\beta$s. In addition, fitting conditionally independent daily models is computationally much faster. Hence, in what follows we only consider this specification.

**3.2.2 The GMRF smoothed downscaler—**To extend the GMRF smoothed downscaler model to the space-time setting, we assume that

$$Y(\mathbf{s},t) = \tilde{\beta}_0(\mathbf{s},t) + \beta_{1,t}\tilde{V}(B,t) + \varepsilon(\mathbf{s},t) \; \varepsilon(\mathbf{s},t) \overset{ind}{\sim} N(0,\tau^2), \quad (14)$$

where we decompose $\tilde{\beta_0}(\mathbf{s}, t)$ as $\tilde{\beta_0}(\mathbf{s}, t) = \beta_{0,t} + \beta_{0,t}(\mathbf{s})$. Potential temporal dependence models for $\beta_{0,t}$, $\beta_{1,t}$ and the single GP $\beta_0(\mathbf{s}, t)$ can take the forms described in Section 3.2.1. Extending the measurement error model (4), we have:

$$x(B,t) = \mu_t + V(B,t) + \eta(B,t), \quad (15)$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

where $\eta(B,t) \overset{ind}{\sim} N(0,\sigma^2)$ and, then $\tilde{V}(B, t) = \mu_t + V(B, t)$

To model the temporal dependence in the latent field $\tilde{V}(B, t)$, as before, we consider two cases. In the first, we assume that for each $t$, the $\mathbf{V}_t = \{V(B, t)\}$ are independent replicates over time of a Gaussian Markov random field provided with a conditional autoregressive prior, that is, for each $t$:

$$V(B_i,t)|\{V(B_j,t):j \neq i\} \sim N\left(\sum_{j\in\delta B_i}\frac{V[B_j,t]}{m_i},\frac{\xi^2}{m_i}\right). \quad (16)$$

In a second case, we assume that the $g$-dimensional random vectors $\mathbf{V}_t$, $t = 1,\ldots, T$, have an AR(1) structure in time. Therefore, for each $t$

$$\mathbf{V}_t=\rho\mathbf{V}_{t-1}+\boldsymbol{\kappa}_t, \quad (17)$$

where $\boldsymbol{\kappa}_t = \{\kappa(B, t)\}$ is a Gaussian Markov random field with a conditional autoregressive structure and $\mathbf{V}_0 = \{V(B, 0)\} \equiv \mathbf{0}$. As noted above, we only consider the model where the $\beta_{0,t}$, $\beta_{1,t}$, $\beta_0(\mathbf{s}, t)$, and the $\mathbf{V}_t = \{V(B, t)\}$ are independent in time.

### 3.2.3 The smoothed downscaler using spatially varying random weights—

Extending (7), we assume that:

$$Y(\mathbf{s},t)=\tilde{\beta}_0(\mathbf{s},t)+\beta_{1,t}\tilde{x}(\mathbf{s},t)+\varepsilon(\mathbf{s},t)\ \varepsilon(\mathbf{s},t)\sim N(0,\tau^2), \quad (18)$$

where $\tilde{\beta_0}(\mathbf{s}, t) = \beta_{0,t} + \beta_0(\mathbf{s}, t)$. Potential models for the time varying parameters $\beta_{0,t}$, $\beta_{1,t}$ and $\beta_0(\mathbf{s}, t)$ are as in Section 3.2.1 and 3.2.2. To define $x(\tilde{\mathbf{s}}, t)$, straightforward extension of (8) yields

$$\tilde{x}(\mathbf{s},t)=\sum_{k=1}^{g}w_k(\mathbf{s},t)x(B_k,t), \quad (19)$$

where the weights $w_k(\mathbf{s}, t)$ are random and varying both in space and time. Again, we let $\mathbf{r}_k$, $k = 1,\ldots, g$ denote the centroids of the $g$ numerical model grid cells. First, we introduce independent latent mean-zero GPs, $Q(\mathbf{r}, t)$, $t = 1,\ldots, T$, with exponential covariance structure, decay parameter $\varphi_{Q_t}$ and variance $\sigma^2_Q$. Then, the weights $w_k(\mathbf{s}, t)$ take the form:

$$w_k(\mathbf{s},t):=\frac{\mathscr{K}(\mathbf{s}-\mathbf{r}_k;\psi_t)\cdot\exp(Q[\mathbf{r}_k;t])}{\sum_{l=1}^{g}\mathscr{K}(\mathbf{s}-\mathbf{r}_l;\psi_t)\cdot\exp(Q[\mathbf{r}_l;t])}, \quad (20)$$

where $\mathcal{K}(\mathbf{s} - \mathbf{r}_k;\ \psi_t) = \exp(-\psi_t\,|\mathbf{s} - \mathbf{r}_k|)$, an exponential kernel with decay parameter $\psi_t$. This model allows the flexibility of spatially and temporally varying weights.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

In an alternative formulation, the weights can be specified dynamically by assuming that the latent GP $Q(\mathbf{r}, t)$ follows an AR(1) process in time, i.e.,

$$Q(\mathbf{r},t)=\gamma Q(\mathbf{r},t-1)+\lambda(\mathbf{r},t) \quad (21)$$

with $\lambda(\mathbf{r}, t)$ independent GPs, $Q(\mathbf{r}, 0) \equiv 0$ and the weights as in (20). Again, we only consider the $\gamma = 0$ case, i.e. independent $Q$ surfaces over time. As in the static setting, we impose a sum to 0 constraint on each of the GPs $Q(\mathbf{r}, t)$, $t = 1,\ldots, T$.

## 4. Model fitting

### 4.1 Priors

All the downscaler models introduced in Section 3 arise as a Bayesian hierarchical formulation and are completely specified given priors for all the parameters. In this section, we briefly discuss the priors adopted for the various model parameters.

First, we consider the static case. Global calibration of the numerical model output results from $\beta_0$ and $\beta_1$ for which we employ a bivariate normal prior with prior mean equal to $(0\ 1)'$ and a diagonal prior covariance matrix with very large diagonal entries corresponding to vague prior variances. For the coregionalization matrix $\mathbf{A}$ introduced in the downscaler model in (3) we specify a prior via its entries. More precisely, we place vague lognormal priors on the diagonal terms of $\mathbf{A}$, as they are related to the variances of the local adjustments $\beta_0(\mathbf{s})$ and $\beta_1(\mathbf{s})$ and a vague normal prior on the off-diagonal element $A_{21}$. For all models, we adopt standard conjugate inverse gamma priors for all the variance terms, that is, for $\tau^2$, $\sigma^2$, $\xi^2$, $\sigma^2_{\beta_0}$, $\sigma^2_Q$. In particular for $\sigma^2_Q$ we chose an inverse gamma prior with prior scale equal to 2 and with prior mean equal to 1.0. With more interest in smoothing than in measurement error, we place a rather informative prior on $\sigma^2$ specified so as to produce a posterior mean for $\sigma^2$ smaller, on average, than that of $\xi^2$, hence allowing for more variability in $V(B)$ than in $\eta(B)$.

For the GMRF smoothed downscaler model, we assign a vague normal prior to $\mu$ with prior mean equal to the average, over all grid cells $B$, of the numerical model output, $x(B)$. Similar definition was adopted in the spatio-temporal case; for each $t = 1,\ldots, T$, the prior mean for $\mu_t$ was taken to be equal to the average of $x(B, t)$.

Regarding priors for the decay parameters, it is not possible to consistently estimate all the spatial covariance parameters under weak priors (Zhang 2004). In light of this, we adopted the following strategy: we used a continuous prior—an inverse gamma—for the marginal variance, while we used a discrete prior placed on a grid of values for the decay parameters. That is, for $\varphi_0$ and $\varphi_1$, we placed a discrete uniform prior on the grid of values, 0.0015, 0.001, 0.01, 0.05, and 0.1, corresponding, respectively, to practical ranges of about 3000, 2000, 300, 60, and 30 km. In the spatio-temporal case, we assumed that for each $t$, $\varphi_{0,t}$, and $\varphi_{1,t}$ followed the same discrete uniform prior.

We could adopt the above specifications for the spatial decay parameters $\psi$ and $\varphi_Q$ in the downscaler with spatially varying random weights. However, we chose to keep them fixed.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

We set $\psi$ equal to 0.08 yielding an exponential kernel with a range of approximately 36 km, that is three 12-km grid cells. Essentially, only first, second, and third order grid cell neighbors contribute nonnegligibly to the weighted average $x(\tilde{\mathbf{s}})$ at $\mathbf{s}$. Experiments with $\psi$s in the neighborhood of 0.08 didn't reveal sensitivity in terms of predictive performance.

The parameter $\varphi_Q$ determines the smoothness of the process $Q(\mathbf{r})$ and, thus, affects the smoothness of the weights $w_k(\mathbf{s})$ as $\mathbf{s}$ moves across the spatial domain. It is clear that the smaller $\varphi_Q$ is, the smoother realizations of $Q(\mathbf{r})$ will be and thus, through (9), resulting weights may be less site-specific than we wish. So, we set $\varphi_Q$ equal to 0.12 (larger than $\psi$), which corresponds to a range of approximately 24 km, i.e., two 12-km grid cells. Experiments with daily decay parameters $\psi_t$ and $\varphi_{Q_t}$ yielded no distinguishable gain in model performance.

### 4.2 Computational details

We fit each of the downscaler models presented in Sections 3.1.1, 3.1.2, and 3.1.3 using a Markov Chain Monte Carlo (MCMC) algorithm. Previous experience with the downscaler (Berrocal et al. 2010b) suggested that we keep only the local intercept adjustment $\beta_0(\mathbf{s})$ different from zero. Thus, we have a single GP and all priors are conjugate.

In the smoothed downscaler using spatially varying random weights, the posterior distribution of the latent GP $Q(\mathbf{r})$ does not have a closed form. Hence, it is necessary to use a Metropolis–Hastings algorithm within the MCMC algorithm where we impose the sum to 0 constraint “on the fly” after each MCMC iteration (Besag et al. 1995). Moreover, we also face a dimensionality problem. To obtain a new sample of weights at each iteration, it is necessary to derive a realization of the latent process $Q(\mathbf{r})$ at the $g$ numerical model grid cell centroids $\mathbf{r}_k$. Given the size of $g$, to reduce the computational burden associated with the sampling of the weights in (9) and (20), respectively, we replace $Q(\mathbf{r}_k)$ and $Q(\mathbf{r}_k, t)$, respectively, with the predictive processes (Banerjee et al. 2008) $Q(\tilde{\mathbf{r}}_k)$ and $Q(\tilde{\mathbf{r}}_k, t)$. We used $m = 648$ knots thinned from the centroids of the overall set of 40,044 CMAQ grid cells by systematically selecting knots every 8 rows and 8 columns. So many knots selected in a space-filling fashion avoids concern regarding knot selection issues (see Banerjee et al. 2008). More details on predictive processes and our implementation are provided in the Appendix.

Extension of the MCMC algorithm to the space-time setting is straightforward when the time varying parameters are independent across time. However, we still fit a joint model due to the common variance parameters, e.g., $\tau^2$, $\sigma^2$, $\xi^2$, $\sigma_Q^2$. If some of the time varying parameters evolve dynamically in time, fitting requires embedding the Forward Filtering Backward Sampling (FFBS, Carter and Kohn 1994; West and Harrison 1999) algorithm within the MCMC algorithm. Finally, we accommodated missingness in the training dataset by using data augmentation to fill in the missing $Y(\mathbf{s}, t)$ at each MCMC iteration under the assumption that the missingness is ignorable.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

## 5. Results

We discuss results for the spatio-temporal versions of the three downscaler models where all the time varying parameters are assumed to be independent across time along with an ordinary kriging model obtained from (10) by setting $\tilde{\beta}_1(\mathbf{s}, t)$ equal to 0. As mentioned in Section 2, we fit the models to 700 training sites and we assess the predictive performance of the various models at 100 hold-out sites. We evaluate the out-of-sample predictive performance of each model in terms of Predictive Mean Square Error (PMSE), averaged across space and time, Predictive Mean Absolute Error (PMAE), averaged across space and time, the average length of the 95% predictive interval, averaged across space time, and the empirical coverage of the 95% predictive interval. Table 1 presents results for these statistics for the four models. All downscaler models yield predictions with much lower PMSE and PMAE than an ordinary kriging model, demonstrating the benefit of using the information contained in the CMAQ output. In turn, both the GMRF smoothed downscaler and the smoothed downscaler using spatially varying random weights provide substantial improvement over the downscaler, supporting the need to account for error in the association that links the observation at a site $\mathbf{s}$, $Y(\mathbf{s})$, to the numerical model output at the grid cell $B$ that contains $\mathbf{s}$. In particular, the GMRF smoothed downscaler provides a 5.3% and a 1.9% reduction, respectively, in PMSE and PMAE over the downscaler, while the smoothed downscaler using spatially varying random weights yields a 14.5% and 5.7% improvement, respectively, in PMSE and PMAE. It is noteworthy that the improvement in PMSE and PMAE of the GMRF smoothed downscaler and of the smoothed downscaler with spatially varying weights over the downscaler model is larger at sites that are farther from monitoring training sites (figure available in online supplementary material).

Spatial plots of the posterior mean of $\{x(\tilde{\mathbf{r}}_k, t)\}_{k=1,\ldots,g}$, $\{\tilde{V}(B_k, t)\}_{k=1,\ldots,g}$ and of the CMAQ model output $\{x(B_k, t)\}_{k=1,\ldots,g}$ for July 4, 2001 for a subregion of the Eastern US are shown in Figure 3. Both the GMRF smoothed downscaler and the smoothed downscaler using spatially varying random weights yield surfaces that are smoother than the original CMAQ output and, as Table 1 indicates, are better associated with the monitoring data. All three downscaler models yield predictions of ozone concentration over the entire spatial domain. However, differently from the numerical model output, the three downscaler models produce predictions that are better calibrated than CMAQ itself.

Figure 4 shows the observed ozone concentration on August 9, 2001 at monitoring sites located in two subregions of the Eastern US along with the posterior predictive mean of ozone concentration resulting from the smoothed downscaler with spatially varying random weights. The predictive surfaces displayed in panels (b) and (d) reproduce the spatial gradient that is visible in the monitoring station data and tend to agree rather well with the observational data.

All three downscaler models provide information on the daily local and global bias of the CMAQ model output through $\beta_{0,t}$, $\beta_{1,t}$ and $\beta_0(\mathbf{s}, t)$. Though all the downscaler models have been fitted to ozone concentration on the square root scale, inspection of the posterior mean and 95% credible intervals for $\beta_{0,t}$ and $\beta_{1,t}$ clearly indicates that there is a need for calibration of CMAQ since for most days during the three months of June, July, and August

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

2010, all three models yield posterior estimates for $\beta_{0,t}$ and $\beta_{1,t}$ significantly different from 0 and 1, respectively.

Spatial plots of the weights $w_k(\mathbf{s}, t)$ for July 4, 2001, associated with four sites located in the Eastern US and depicted in Figure 5(a) are shown in Figures 5(b)–(e). In each plot, the location of the site within the numerical model grid cell is marked with a dot. As Figures 5(b)–(e) illustrate, the weights have a different orientation and magnitude from site to site.

In particular, sites $\mathbf{s}_1$, $\mathbf{s}_2$ and $\mathbf{s}_3$ assign a larger weight to the grid cell $B$ where they lie, while site $\mathbf{s}_4$ assigns similar weights to three numerical model grid cells, including the one in which it lies. In addition, the weights reveal different directionalities for the different sites. The weights $w_k(\mathbf{s}, t)$ not only vary in space, but they also vary in time. This is evident by inspecting the posterior mean of the weights $w_k(\mathbf{s}, t)$ at the same four sites using two additional days—July 20 and August 9, 2001 (figure available in the supplementary online material). These days were chosen because they are characterized by different conditions in terms of variability in the observed ozone concentration data as illustrated in Figure 2.

Finally, we investigated whether the new regressor, $x(\tilde{\mathbf{s}}, t)$, is better correlated with the monitoring data, i.e., it explains the observational data better than the numerical model output itself. In this regard, we have computed the daily correlation between $Y(\mathbf{s}, t)$ and, respectively, the numerical model output $x(B, t)$, the posterior mean of $\tilde{V}B, t)$, and the posterior mean of $x(\tilde{\mathbf{s}}, t)$ where $B$ still denotes the grid cell containing $\mathbf{s}$. Table 2 reports values of these correlations for the three selected days of July 4, July 20, and August 9, 2001. We see that $x(\tilde{\mathbf{s}}, t)$ has a higher correlation with $Y(\mathbf{s}, t)$ than $x(B, t)$ and $\tilde{V}(B, t)$.

## 6. Discussion

We have presented two extensions of our earlier downscaler approach that smooth the computer model output for insertion into a linear regression with spatially varying coefficients. These regressions provide daily interpolated exposure surfaces, assimilating the observed station data with the computer model output. Through a hold-out sample 5% and 15% improvement in predictive mean square error emerges for these new models relative to the original downscaler. Furthermore, greater improvement in predictive performance is found at sites that are far from the monitoring sites.

These new models are more demanding to fit than the original downscaler. However, the GMRF smoothed downscaler can take advantage of the associated convenient full conditional updating while the smoothed downscaler with spatially varying random weights can be implemented using dimension reduction through predictive processes. User-friendly software for fitting the latter model is available on request. Again, fitting a daily fusion model that requires stochastic integration is infeasible with more than 40, 000 grid cells much less across thetimeofanentireozone season.Furthermore, despite being uncalibrated, CMAQ contains useful information for predicting ozone concentration at unmonitored sites: all downscaler models yielded a substantial improvement in out-of-sample predictive performance over an ordinary kriging model.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Both extensions of the downscaler model can be used in conjunction with an environmental exposure analysis. If the health data is point-referenced, then the health outcome at **s** could be modeled as a function of ozone concentration at **s**, where the ozone concentration at **s** is the posterior predictive mean obtained from the model. Similarly, if the health outcome is aggregated over an area, then it could be regressed on the posterior predictive mean of the average ozone concentration over the area. Arguably better is a joint Bayesian approach that models exposure and health outcome jointly and induces a conditional model for outcome given exposure.

Further work is following two tracks. One notes that the primary US EPA air quality standard for ozone is specified in terms of the fourth highest daily maximum across the year exceeding a particular threshold. Using these new models, we would like to provide predictive distributions to assess compliance with respect to this criterion. A second track seeks to develop conditional CAR models using weights that are driven by a GP. That is, given the GP realization, we will define the weights to yield a valid Gaussian CAR model. This enables CAR models with random, spatially varying weights. Allowing directionality in weights enables improved reconstruction of blurred images using GMRF's. We will report on this in a forthcoming manuscript.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Appendix

## A.1 Predictive Process Models

In this appendix we provide a brief review on predictive process models and we illustrate how predictive processes can be employed in the smoothed downscaler model with spatially varying random weights to ease computation.

Let $Y(\mathbf{s})$, $\mathbf{s} \in \mathcal{D}$, denote a spatial process, not necessarily Gaussian. The classical geostatistical model decomposes $Y(\mathbf{s})$ as follows:

$$Y(\mathbf{s}) = \mu(\mathbf{s}) + w(\mathbf{s}) + \varepsilon(\mathbf{s}) \quad \varepsilon(\mathbf{s}) \overset{ind}{\sim} N(0, \tau^2), \quad \text{(A.1)}$$

where $\mu(\mathbf{s})$ represents the mean structure of $Y(\mathbf{s})$, and $w(\mathbf{s})$ is modeled as a mean-zero GP with covariance function $C(\cdot, \cdot; \theta)$. Inference on the covariance parameter $\theta$ is computationally challenging if a large number of observations is available.

The predictive process model addresses this problem by replacing $w(\mathbf{s})$ in (A.1) with the predictive process $w(\tilde{\mathbf{s}})$ defined as follows. Let $\mathbf{r}_1^*, \ldots, \mathbf{r}_m^*$ be $m$ pre-specified sites, or knots, in $\mathcal{D}$ and let $\mathbf{w}^*$ denote the $m \times 1$ vector $(w(\mathbf{r}_1^*), \ldots, w(\mathbf{r}_m^*))'$. Then, for each $\mathbf{s} \in \mathcal{D}$, the predictive spatial process $w(\tilde{\mathbf{s}})$ derived from the parent process $w(\mathbf{s})$ is defined as

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\tilde{w}(\mathbf{s}) = c'(\mathbf{s};\boldsymbol{\theta}) \cdot \mathbf{C}^{*-1}(\boldsymbol{\theta}) \cdot \mathbf{w}^*, \quad \text{(A.2)}$$

where $c(\mathbf{s}; \boldsymbol{\theta})$ is the $m \times 1$ vector with $i$th component equal to $C(\mathbf{s}, \mathbf{r}_i^*;\boldsymbol{\theta}), i=1,\ldots,m$, $\mathbf{C}^*(\boldsymbol{\theta})$ is the $m \times m$ matrix with $(i, j)$th element equal to $C(\mathbf{r}_i^*, \mathbf{r}_j^*;\boldsymbol{\theta})$ and $\mathbf{w}^* \sim MVN_m(\mathbf{0}, \mathbf{C}^*(\boldsymbol{\theta}))$.

In other words, the predictive process $w(\tilde{\mathbf{s}})$ is the projection of the original spatial process $w(\mathbf{s})$ onto the $m$-dimensional space generated by $\mathbf{w}^*$. Simulation experiments assessing the performance of predictive process models on knots location has shown little sensitivity of results on knots selection, particularly if the knots are chosen on a regular grid with small spacing relatively to the range of the parent process $w(\mathbf{s})$.

In the smoothed downscaler model with spatially varying weights, we introduce predictive processes to alleviate computation when working with the complete CMAQ model output consisting of $g$=40,044 grid cells. In this situation, as mentioned in Section 4.2, rather than working with the full 40,044-dimensional vectors $(Q[\mathbf{r}_k])_{k=1,\ldots,g}$ and $(Q[\mathbf{r}_k, t])_{k=1,\ldots,g}$, we replace $Q(\mathbf{r}_k)$ and $Q(\mathbf{r}_k, t)$, respectively, with $Q(\tilde{\mathbf{r}}_k)$ and $Q(\tilde{\mathbf{r}}_k, t)$ defined using (A.2) and $m$(=648) dimensional vectors $\mathbf{Q}^*$ and $\mathbf{Q}_t^*$. Then, at each MCMC iteration, we update the $m$-dimensional random vector $\mathbf{Q}^* = (Q[\mathbf{r}_1^*], \ldots, Q[\mathbf{r}_m^*])'$ and $\mathbf{Q}_t^* = (Q[\mathbf{r}_1^*, t], \ldots, Q[\mathbf{r}_m^*, t])'$, respectively, by block-updating, using four blocks of dimension 162 and a random-walk multivariate normal proposal with diagonal covariance matrix and proposal variance appropriately chosen to achieve an acceptance rate of 25 − 40%. We then update the predictive process $Q(\tilde{\mathbf{r}}_k)$, $k = 1,\ldots, g$ (respectively, $Q[\tilde{\mathbf{r}}_k, t]$) and the corresponding new sets of weights $w_k(\mathbf{s})$, $k = 1,\ldots, g$ (respectively, $w_k(\mathbf{s}, t)$, $k = 1,\ldots, g$) for each site $\mathbf{s}$, and we accept or reject the new proposed value following the conventional Metropolis-Hastings scheme. The conditional distributions of the remaining parameters are sampled directly.

## A.2 Full Conditionals

Here we provide full conditionals for $\mathbf{V}_t$ and $\mathbf{Q}_t^*$ in, respectively, the GMRF smoothed downscaler and the smoothed downscaler with spatially varying random weights. To facilitate exposition, we present full conditionals for the case of no missing data. Extension to handle missing data is straightforward.

For each $t$, let $\mathbf{Y}_t = (Y[\mathbf{s}_1, t],\ldots, Y[\mathbf{s}_n, t])'$ denote the $n \times 1$ vector of observations of ozone concentrations for day $t$ and $\boldsymbol{\beta}_{0t}$ denote the $n \times 1$ vector $\boldsymbol{\beta}_{0t} = (\beta_0[\mathbf{s}_1, t],\ldots, \beta_0[\mathbf{s}_n, t])'$. Let $\mathbf{Y}_t^{(B_i)}$ and $\boldsymbol{\beta}_{0t}^{(B_i)}$ denote, respectively, the $n_i \times 1$ vectors $(Y[\mathbf{s}_j, t])_{j=1,\ldots,n_i}$ and $(\beta_0[\mathbf{s}_j, t])_{j=1,\ldots,n_i}$ with $\mathbf{s}_j \in B_i$. For each $t$ and $i = 1,\ldots, g$, from (14), (15), and (16), it follows that the full conditional distribution $[V(B_i, t)|\{V(B_j, t), j \neq i\}, \mu_t, \beta_{0,t}, \beta_{1,t}, \boldsymbol{\beta}_{0t}, \xi^2, \tau^2, \sigma^2, \mathbf{Y}_t, x(B_i, t)]$ is a $N(\gamma_i, \varsigma_i^2)$ distribution with

$$\varsigma_i^2 = \left(\frac{n_i \beta_{1,t}^2}{\tau^2} + \frac{m_i}{\xi^2} + \frac{1}{\sigma^2}\right)^{-1}$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

and

$$\gamma_i = \varsigma_i^2 \cdot \left\{ \frac{\beta_{1,t}}{\tau^2} \mathbf{1}' \left( \boldsymbol{Y}_t^{(B_i)} - \beta_{0,t}\mathbf{1} - \beta_{1,t}\mu_t \mathbf{1} - \boldsymbol{\beta}_{0t}^{(B_i)} \right) + \left( \sum_{j \in \delta B_i} \frac{V(B_j, t)}{\xi^2} \right) + \left( \frac{x(B_i, t) - \mu_t}{\sigma^2} \right) \right\},$$

where $\mathbf{1}$ is a $n_i \times 1$ vector of all 1's.

Now we derive the full conditional distribution of $\boldsymbol{Q}_t^*$ in the smoothed downscaler model using spatially varying random weights. For each $t = 1,\ldots, T$, let $Q(\tilde{\mathbf{r}}_k, t)$, $k = 1,\ldots, g$ denote the predictive spatial process derived from the parent process $Q(\mathbf{r}, t)$. Then, from (A.2), for each $k = 1,\ldots, g$

$$\tilde{Q}(\mathbf{r}_k, t) = c' \left( \sigma_Q^2, \phi_{Q_t} \right) \cdot \mathbf{C}^{*-1} \left( \sigma_Q^2, \phi_{Q_t} \right) \cdot \mathbf{Q}_t^* . \quad \text{(A.3)}$$

We indicate with $\mathbf{e}\tilde{\mathbf{Q}}_t^d$ the $g \times g$ diagonal matrix with $(k, k)$th element $exp(Q(\tilde{\mathbf{r}}_k, t))$, and with $\mathbf{J}$ the $g \times g$ matrix of all 1's. Let $\mathbf{K}_t$ indicate the $n \times g$ matrix $\mathbf{K}_t = (\mathcal{K}[\mathbf{s}_i - \mathbf{r}_j; \psi_t])_{i=1,\ldots,n; j=1,\ldots,g}$ and let $\mathbf{W}_t$ denote the $n \times g$ matrix of spatially varying weights $\mathbf{W}_t = (w_j[\mathbf{s}_i, t])_{i=1,\ldots,n; j=1,\ldots,g}$. Then, from (20) it follows that

$$\mathbf{W}_t = \frac{\mathbf{K}_t \cdot e\tilde{\mathbf{Q}}_t^d}{\left( \mathbf{K}_t \cdot e\tilde{\mathbf{Q}}_t^d \right) \circ \mathbf{J}}, \quad \text{(A.4)}$$

where $\circ$ indicates the Schur product of matrices and the above division of matrices is element-wise.

The Gaussian prior specification for the parent process $Q(\mathbf{r}, t)$, and the likelikood in (18) imply that the full conditional distribution $[\mathbf{Q}_t^* | \beta_{0,t}, \beta_{1,t}, \boldsymbol{\beta}_{0t}, \tau^2, \sigma_Q^2, \phi_{Q_t}, \psi_t, \boldsymbol{Y}_t]$, $\mathbf{t} = \mathbf{1},\ldots, \mathbf{T}$ is

$$\begin{aligned}
\left[ \mathbf{Q}_t^* | \beta_{0,t}, \beta_{1,t}, \boldsymbol{\beta}_{0t}, \tau^2, \sigma_Q^2, \phi_{Q_t}, \psi_t, \boldsymbol{Y}_t \right] &\propto \frac{1}{(\tau^2)^{\frac{n}{2}}} \\
&\times \exp \left\{ -\frac{1}{2} (\boldsymbol{Y}_t - \beta_{0,t}\mathbf{1} - \beta_{1,t}\mathbf{W}_t\mathbf{X}_t - \boldsymbol{\beta}_{0t})' (\tau^2 \mathbf{I})^{-1} (\boldsymbol{Y}_t \right. \\
&- \beta_{0,t}\mathbf{1} - \beta_{1,t}\mathbf{W}_t\mathbf{X}_t \\
&\left. - \boldsymbol{\beta}_{0t}) \right\} \cdot \frac{1}{\left| \mathbf{C}^* \left( \sigma_Q^2, \phi_{Q_t} \right) \right|^{\frac{1}{2}}} \exp \left\{ \right. \\
&\left. - \frac{1}{2} \left( \mathbf{Q}_t^{*'} \mathbf{C}^{*-1} \left( \sigma_Q^2, \phi_{Q_t} \right) \mathbf{Q}_t^* \right) \right\},
\end{aligned}$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

where $\mathbf{X}_t$ is the $g \times 1$ vector $\mathbf{X}_t = (x(B_1, t)\ldots x(B_g, t))'$, $\mathbf{I}$ is the $n \times n$ identity matrix and $\mathbf{W}_t$ is as in (A.4).

## References


Banerjee, S.; Carlin, BP.; Gelfand, AE. Hierarchical Modeling and Analysis for Spatial Data. Boca Raton, FL: Chapman & Hall/CRC; 2004.

Banerjee S, Gelfand AE, Finley AO, Sang H. Gaussian predictive process models for large spatial datasets. Journal of the Royal Statistical Society, Series B. 2008; 70:825–848.

Berrocal VJ, Gelfand AE, Holland DM. A bivariate spatio-temporal downscaler under space and time misalignment. Annals of Applied Statistics. 2010a; 4:1942–1975. [PubMed: 21853015]

Berrocal VJ, Gelfand AE, Holland DM. A spatio-temporal downscaler for outputs from numerical models. Journal of Agricultural, Biological and Environmental Statistics. 2010b; 15:176–197.

Besag J. Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society, Series B. 1974; 136:192–236.

Besag J, Green PJ, Higdon D, Mengersen K. Bayesian computation and stochastic systems. Statistical Science. 1995; 10:3–41.

Byun D, Schere KL. Review of the governing equations, computational algorithms, and other components of the Models-3 Community Multiscale Air Quality (CMAQ) Modeling System. Applied Mechanics Reviews. 2006; 59:51–77.

Carter C, Kohn R. On Gibbs sampling for state space models. Biometrika. 1994; 81:541–553.

Cressie, NAC. Statistics for Spatial Data. 2nd. New York: Wiley; 1993.

Fuentes M, Raftery AE. Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. Biometrics. 2005; 61:36–45. [PubMed: 15737076]

Gelfand AE, Schmidt AM, Banerjee S, Sirmans CF. Nonstationary multivariate process modeling through spatially varying coregionalization. TEST. 2004; 13:263–312.

Gelfand AE, Banerjee S, Gamerman D. Spatial process modelling for univariate and multivariate dynamic spatial data. Environmetrics. 2005; 16:465–479.

Gotway CA, Young LJ. Combining incompatible spatial data. Journal of the American Statistical Association. 2002; 97:632–648.

Guillas S, Bao J, Choi Y, Wang Y. Statistical correction and downscaling of chemical transport model ozone forecasts over Atlanta. Atmospheric Environment. 2008; 42:1338–1348.

Higdon DM. A process-convolution approach to modeling temperatures in the north Atlantic Ocean. Journal of Environmental and Ecological Statistics. 1998; 5:173–190.

Kyung M, Ghosh SK. Bayesian inference for directional conditionally autoregressive models. Bayesian Analysis. 2010; 4:675–706.

Liu Z, Le N, Zidek JV. Combining measurements and physical model outputs for the spatial prediction of hourly ozone space-time fields. The University of British Columbia Department of Statistics Technical Report no 239. 2008

Lu H, Carlin BP. Bayesian areal wombling for geographical boundary analysis. Geographical Analysis. 2005; 36:265–285.

McMillan N, Holland DM, Morara M, Feng J. Combining numerical model output and particulate data using Bayesian space-time modeling. Environmetrics. 2010; 21:48–65.

Sahu SK, Gelfand AE, Holland DM. High resolution space-time ozone modeling for assessing trends. Journal of the American Statistical Association. 2007; 102:1221–1234. [PubMed: 19759840]

Wackernagel, H. Multivariate Geostatistics: an introduction with applications. 3rd. Berlin: Springer; 2003.

West, M.; Harrison, J. Bayesian Forecasting and Dynamic Models. 2nd. New York: Springer-Verlag; 1999.

Zhang H. Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. Journal of the American Statistical Association. 2004; 99:250–261.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

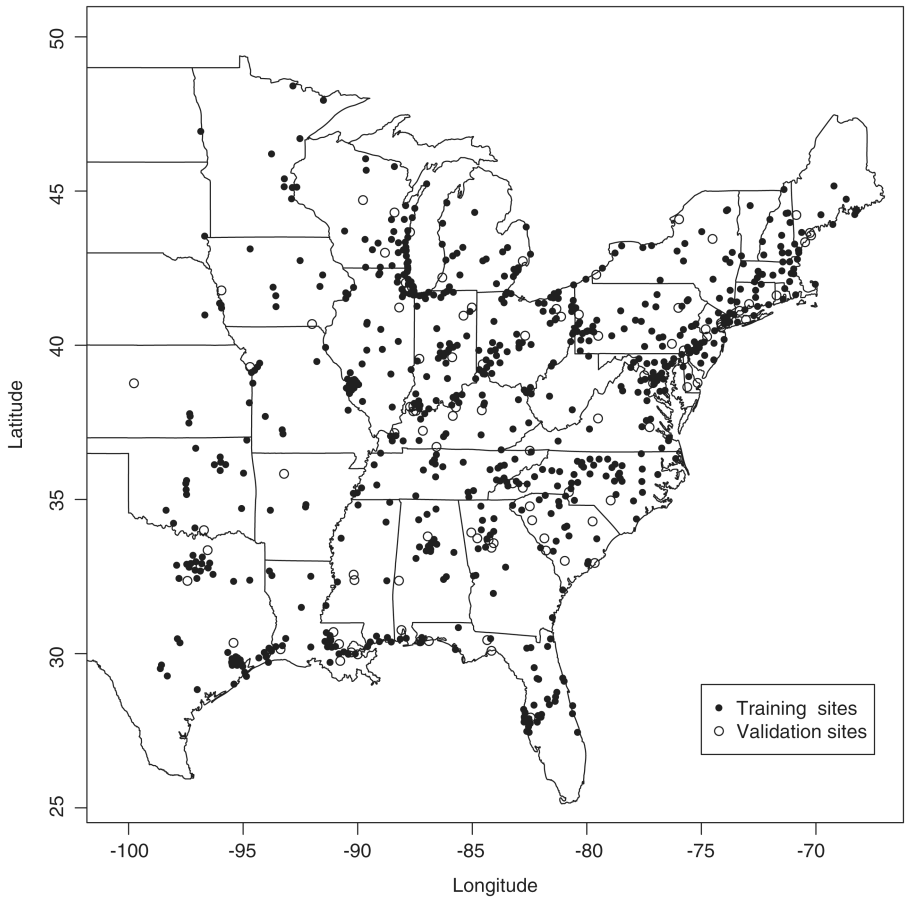


**Figure 1.**

Training and validation sites used to fit and assess the out-of-sample predictive performance of the ordinary kriging model, the downscaler, the GMRF smoothed downscaler and the smoothed downscaler using spatially varying random weights.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

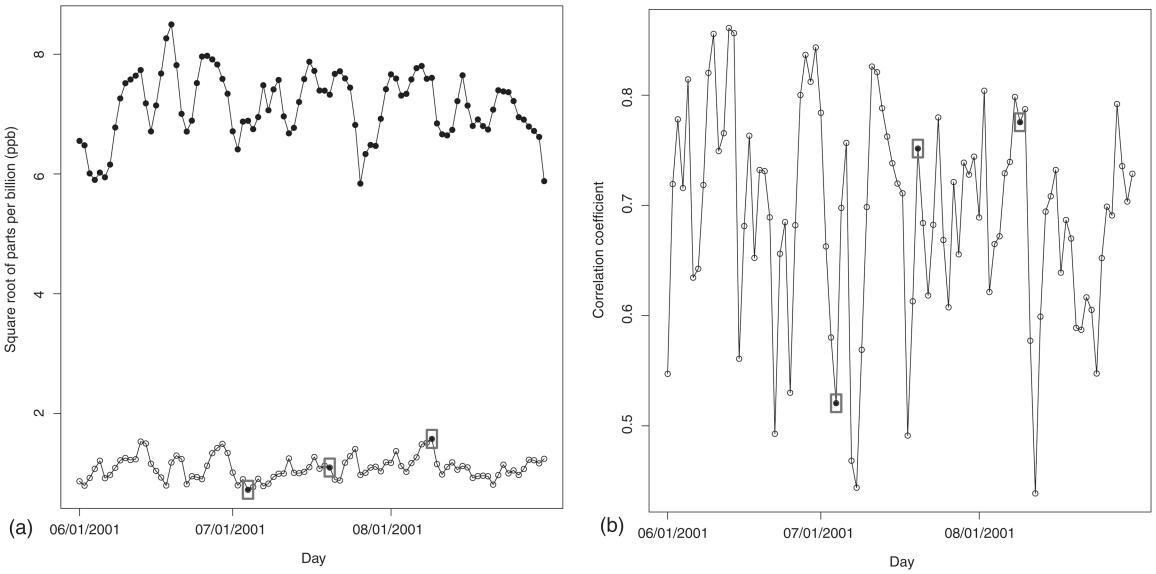


**Figure 2.**

(a) Daily mean (filled circles) and standard deviation (empty circles) of square root of observed ozone concentration at all the 800 monitoring sites. (b) Daily correlation betweeen square root of observed ozone concentration and square root of CMAQ output of ozone concentration. In both plots, the three days for which we will present results in Section 5 are surrounded by a box. They are, respectively, July 4, July 20, and August 9, 2001.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

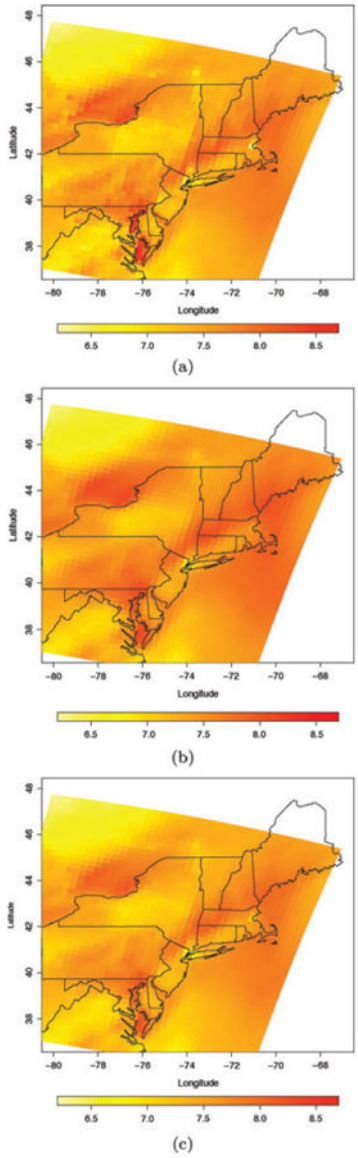


**Figure 3.**

Spatial maps of: (a) the square root of the CMAQ output, $x(B, t)$, (b) the posterior mean of $\tilde{V}(B, t)$, and (c) the posterior mean of $x(\tilde{B}, t)$, for July 4, 2001 for a subregion in the Northeast.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

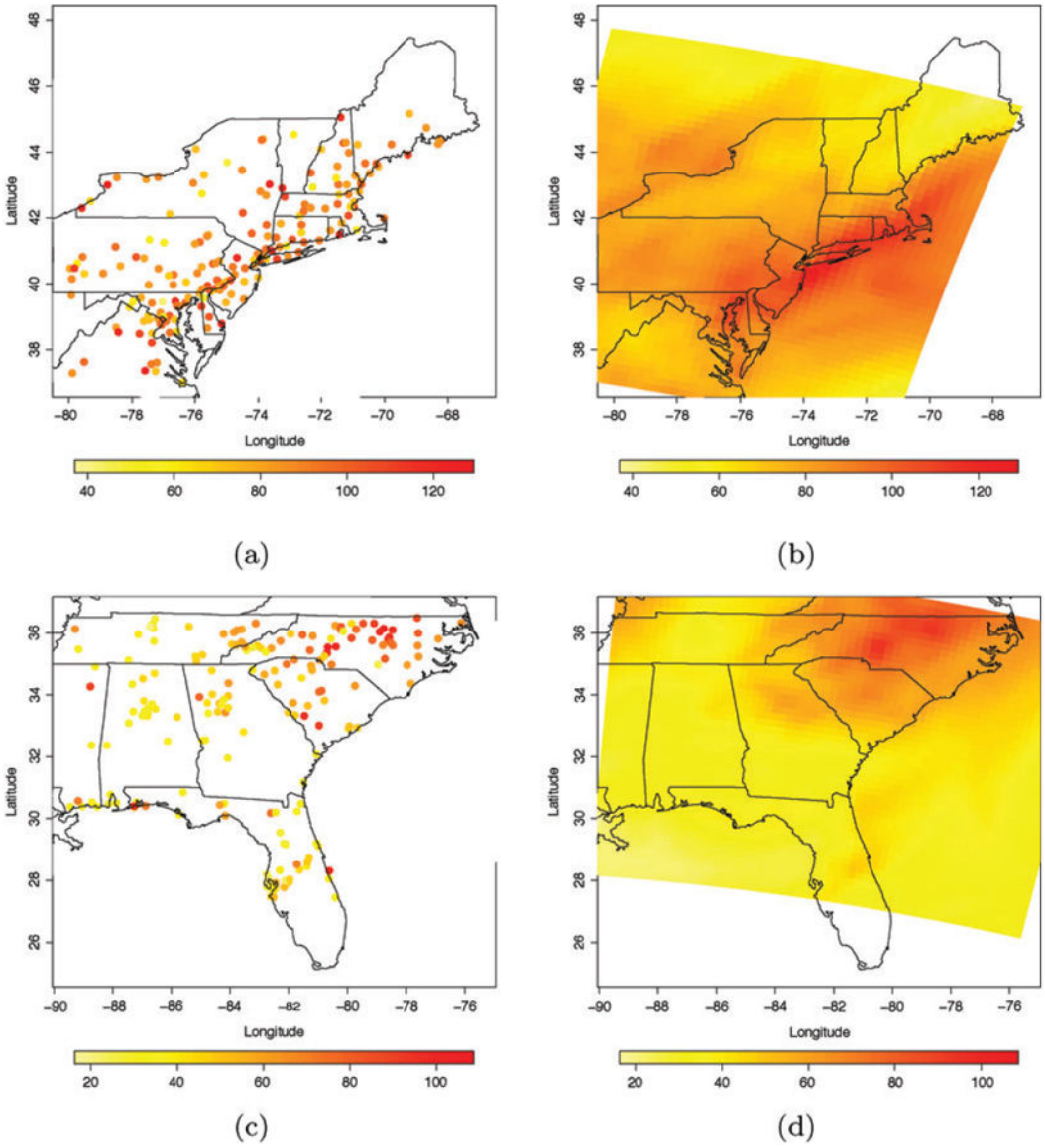


**Figure 4.**

(a)–(c) Observed ozone concentration (ppb) on August 9, 2001 in two subregions of the Eastern US. (b)–(d) Posterior predictive mean of ozone concentration on August 9, 2001 as yielded by the smoothed downscaler with spatially varying random weights.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

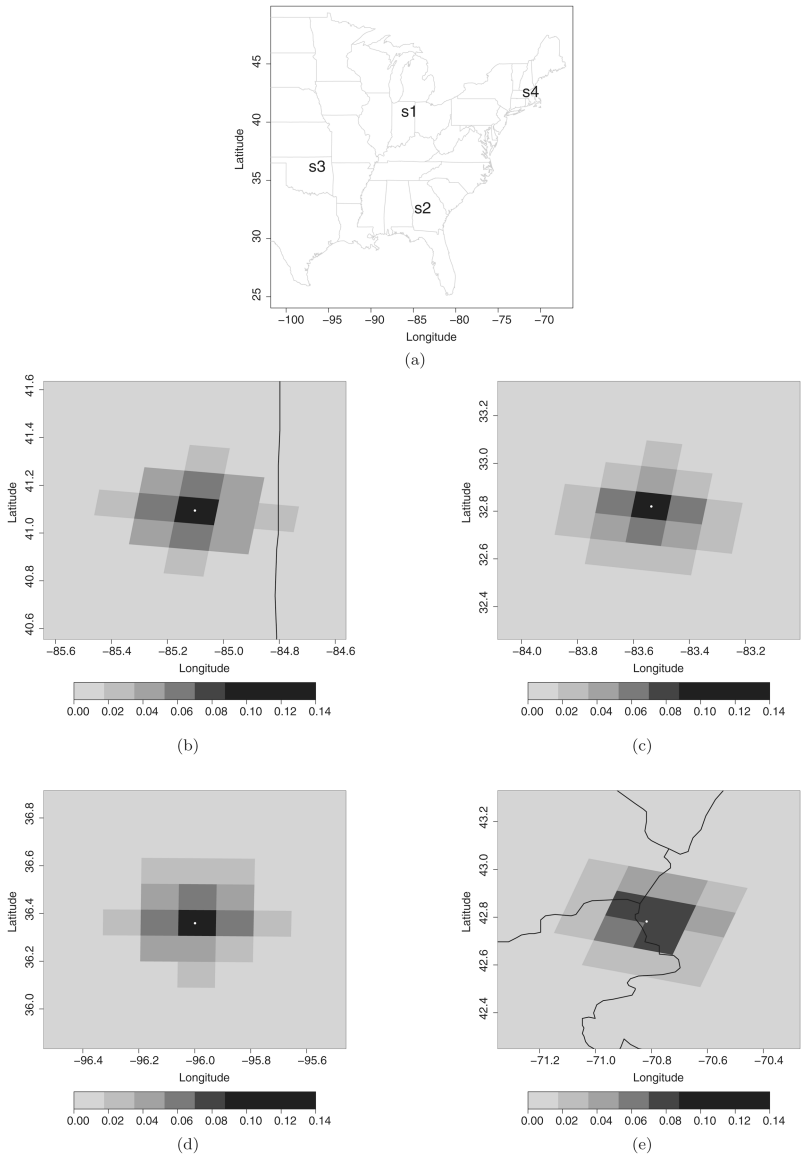


**Figure 5.**

(a) Location of the four sites for which we are displaying the posterior predictive mean of the spatially varying random weights $w_k(\mathbf{s}, t)$. (b)–(e) Posterior predictive mean of the spatially varying random weights $w_k(\mathbf{s}, t)$ for sites: (b) $\mathbf{s}_1$; (c) $\mathbf{s}_2$; (d) $\mathbf{s}_3$; and (e) $\mathbf{s}_4$ on July 4, 2001.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

**Table 1**

Predictive Mean Squared Error (PMSE), Predictive Mean Absolute Error (PMAE), average length of the 95% predictive interval, and empirical coverage of the 95% predictive interval for the numerical model CMAQ, the new regressor $x(\tilde{\mathbf{s}}, t)$, an ordinary kriging model, the downscaler, the GMRF smoothed downscaler and the smoothed downscaler with spatially varying random weights

| Model | PMSE | PMAE | Average length of 95% PI | Empirical coverage of 95% PI |
|---|---|---|---|---|
| CMAQ model | 135.9 | 9.1 | — | — |
| $x(\tilde{\mathbf{s}}, t)$ | 124.2 | 8.7 | — | — |
| Ordinary kriging | 60.9 | 5.8 | 30.6 | 94.8% |
| Downscaler | 53.1 | 5.3 | 30.4 | 94.9% |
| GMRF smoothed downscaler | 50.3 | 5.2 | 29.4 | 94.9% |
| Smoothed downscaler with spatially varying random weights | 45.4 | 5.0 | 27.7 | 95.0% |

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

**Table 2**

Correlation between the square root of the observed ozone concentration at the training sites, $Y(\mathbf{s}, t)$, and the square root of the CMAQ output, $x(B, t)$, the posteriormeanof $\tilde{V}(B, t)$ where $\mathbf{s} \in B$, and the posterior mean of $\tilde{x}(\mathbf{s}, t)$ for three days in the summer of 2001: July 4, July 20, and August 9, 2001

| Day | $x(B, t)$ | Posterior mean of $\tilde{V}(B, t)$ | Posterior mean of $x(\tilde{s}, t)$ |
|---|---|---|---|
| 07/04/2001 | 0.52 | 0.59 | 0.61 |
| 07/20/2001 | 0.76 | 0.80 | 0.81 |
| 08/09/2001 | 0.78 | 0.85 | 0.86 |