Model Evaluation and Spatial Interpolation by Bayesian Combination of Observations with Outputs from Numerical Models

Montserrat Fuentes

Statistics Department, North Carolina State University, Raleigh, North Carolina 27695-8203, U.S.A.
email: fuentes@stat.ncsu.edu

and

Adrian E. Raftery

Statistics Department, University of Washington, Seattle, Washington 98195-4320, U.S.A.
email: raftery@stat.washington.edu

SUMMARY. Constructing maps of dry deposition pollution levels is vital for air quality management, and presents statistical problems typical of many environmental and spatial applications. Ideally, such maps would be based on a dense network of monitoring stations, but this does not exist. Instead, there are two main sources of information for dry deposition levels in the United States: one is pollution measurements at a sparse set of about 50 monitoring stations called CASTNet, and the other is the output of the regional scale air quality models, called Models-3. A related problem is the evaluation of these numerical models for air quality applications, which is crucial for control strategy selection. We develop formal methods for combining sources of information with different spatial resolutions and for the evaluation of numerical models. We specify a simple model for both the Models-3 output and the CASTNet observations in terms of the unobserved ground truth, and we estimate the model in a Bayesian way. This provides improved spatial prediction via the posterior distribution of the ground truth, allows us to validate Models-3 via the posterior predictive distribution of the CASTNet observations, and enables us to remove the bias in the Models-3 output. We apply our methods to data on SO_2 concentrations, and we obtain high-resolution SO_2 distributions by combining observed data with model output. We also conclude that the numerical models perform worse in areas closer to power plants, where the SO_2 values are overestimated by the models.

KEY WORDS: Air pollution; Deterministic simulation models; Environmental statistics; Geostatistics; Spatial statistics.

1. Introduction

Emission reductions were mandated in the Clean Air Act Amendments of 1990 with the expectation that they would result in major reductions in the concentrations of atmospherically transported pollutants. Maps of dry deposition and concentration levels of pollutants are useful for discovering when, where, and to what extent the pollution load is increasing or declining. Ideally, such maps would be based on a dense network of monitoring stations, covering most of the United States, at which dry deposition and concentrations of air pollutants would be measured on a regular basis. Unfortunately, such a network does not exist. Instead, there are two main sources of information about dry deposition pollution levels in the United States, and two resulting ways of constructing pollution maps. The first is a *sparse* set of about 50 irregularly spaced sites in the eastern United States, the Clean Air Status and Trends Network (CASTNet), at which the Environmental Protection Agency (EPA) regularly measures concentrations and fluxes of different atmospheric pollutants (see Figure 1A). It would be possible to use an interpolation method to pro-

duce a pollution map. However, the air pollutants' fluxes and concentrations are functions of terrain, atmospheric turbulence, vegetation, the rate of growth of the vegetation, and other soil and surface conditions. Because these factors vary abruptly in space and time and because the monitoring stations are too far from each other, interpolation of the CAST-Net monitoring data is recognized to be inadequate for the problem (Clarke and Edgerton, 1997).

The second source of information is pollution emissions data. The point and area sources of emissions are available from known sources of pollution such as chemical plants, generally in the form of annual totals. If the emissions data were accurate and available at a fine time resolution, and if we had precise information about local weather, land use and cover, and pollution transport dynamics, we could in principle work out pollution levels at each point in time and space quite accurately. This ideal is far from being attained. However, the available emissions data have been combined with numerical models of local weather (the Mesoscale Model version 5 [MM5]), the emissions process (the Sparse Matrix Operator
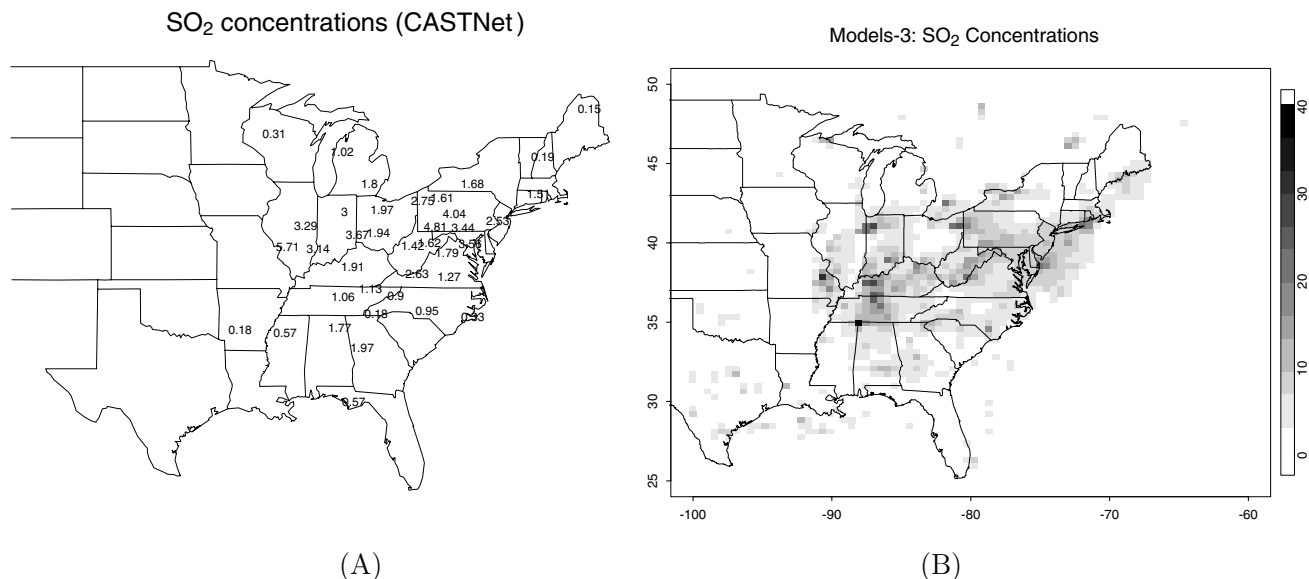
Figure 1. (A) Weekly average of SO$_2$ concentrations (parts per billion; ppb) at the Clean Air Status and Trends Network (CASTNet) sites, for the week of July 11, 1995. (B) Output of Models-3, weekly average of SO$_2$ concentrations (ppb), for the week of July 11, 1995.

Kernel Emissions [SMOKE] model), as well as information about land use and cover, to estimate pollution levels in space and time (the Community Multiscale Air Quality [CMAQ] output) and to produce maps (Dennis et al., 1996). These are not statistical models, but rather numerical deterministic simulation models based on systems of differential equations that attempt to represent the underlying chemistry; they take the form of huge blocks of computer code. The combination of these models is referred to as "Models-3" (models generation 3). The models are run by the EPA and individual U.S. states, and they provide estimates of pollutant concentrations and fluxes on regular grids in parts of the United States (see Figure 1B).

The output of Models-3 generates averaged concentrations/ fluxes over regions of size 36 × 36 km. This approach may also be unsatisfactory for two main reasons. First, the underlying emissions data are often not of high quality (Dolwick et al., 2001). Second, the underlying models may be inadequate in various ways. It seems clear that combining the two main approaches and sources of information, the model estimates and the point measurements, could lead to a better solution. So far, efforts to do this have focused on model evaluation, in which model predictions are compared with measurements, and the models are revised and the outputs adjusted if discrepancies are found (Dennis et al., 1990). The final maps are still based on the model output alone.

The evaluation of physically based computer models for air quality applications is crucial to assist in control strategy selection. Selecting the wrong control strategy has costly economic and social consequences. The objective comparison of modeled concentrations with observed field data is one approach to assessment of model performance. Early evaluations of model performance usually relied on linear least-squares analysis of observed versus modeled values, using scatterplots of the values.

Further development of these proposed statistical evaluation procedures is needed, and we propose a Bayesian approach. Statistical assessment is tricky in this case, because the model predictions and the observations do not refer to the same spatial locations, and indeed are on different spatial scales. The fact that they are on different spatial scales is called the "change of support" problem. The model predictions are averages over grid squares, while the observations are at points in space; the two are thus not directly comparable. One approach to making them comparable is to apply interpolation and extrapolation methods to the CASTNet point measurements so as to produce empirical estimates of grid square averages, and then compare those to the model predictions (Sampson and Guttorp, 1998). One difficulty with this is that the interpolated grid square averages can be poor because of the sparseness of the CASTNet network, and so treating them as ground truth for model evaluation is questionable.

A related problem is that the comparison does not take into account the uncertainty in the interpolated values. In this article, we develop a new approach to the model-evaluation problem, and show how it can also be used to remove the bias in model output and to produce maps that combine model predictions with observations in a coherent way. Combining both sources of information versus using just the sparse data field or the unreliable model output should lead to improved maps of air quality. We specify a simple model for both Models-3 predictions and CASTNet observations in terms of the unobserved ground truth, and estimate it in a Bayesian way. Solutions to all the problems considered here follow directly. Model evaluation then consists of comparing the CASTNet observations with their predictive distributions given the Models-3 output. Bias removal follows from estimation of the bias parameters in the model. Maps of pollution levels and of the uncertainty about them, taking into account all the available

information, are based directly on the posterior distribution of the (unobserved) ground truth. The resulting approach takes account of and estimates the bias in the atmospheric models, the lack of spatial stationarity in the data, the ways in which spatial structure and dependence change with locations, the change of support problem, and the uncertainty about these factors. It can be viewed as an instance of the Bayesian melding framework for inference about deterministic simulation models (Poole and Raftery, 2000), and its implementation is quite straightforward.

A similar approach has been developed by Cowles and Zimmerman (2002, 2003), who have used systematic sampling and standard numeric integration techniques rather than Monte Carlo integration to combine point and areal data. Another relevant method was proposed by Best, Ickstadt, and Wolpert (2000), who related different spatially varying quantities to an underlying unobservable random field for a regression analysis of health and exposure data. In our method we also have an underlying unobserved process, but the statistical model we propose is novel. We present a new way to relate air pollution variables to an underlying process representing the true air pollution values, as well as a new model for nonstationarity. Wikle et al. (2001) presented an approach to combining data from different sources to improve the prediction of wind fields. Wikle et al.'s approach is a conditional one in which all the spatial quantities are defined through a series of statistical conditional models. In their approach, the output of the numerical models is treated as a prior process. Here we present a simultaneous representation of the data and the output of numerical models in terms of the underlying truth. Our method is different from that of Wikle et al., because we do not treat the output of the numerical models as a prior process, but rather as another source of data. Therefore, we write the output of the models in terms of the underlying truth, taking into account the potential bias of the numerical models. The way we estimate these bias parameters is also novel. To our knowledge, this is the first time that statistical methods that take into account nonstationarity, change of support, and the uncertainty about these factors have been used for the evaluation of the regional scale air quality numerical models.

In Section 2 we describe the statistical model, and in Section 3 we show some of our results for model evaluation and map construction using the combined data for the air pollution problem. Section 4 presents some final remarks.

2. The Statistical Model

2.1 *Statistical Models for CASTNet and Models-3 Output*

We do not consider CASTNet measurements to be the "ground truth," because there is measurement error. Instead, we assume that there is an underlying (unobserved) field $Z(\mathbf{s})$, where $Z(\mathbf{s})$ measures the "true" concentration/flux of the pollutant at location $\mathbf{s}$. At station $\mathbf{s}$ we denote the CASTNet observation by $\hat{Z}(\mathbf{s})$, and we assume that

$$\hat{Z}(\mathbf{s}) = Z(\mathbf{s}) + e(\mathbf{s}), \qquad (1)$$

where $e(\mathbf{s})$ represents the measurement error at location $\mathbf{s}$ and $e(\cdot) \sim N(0, \sigma_e^2)$ is a white noise process. The process $e(\mathbf{s})$ is independent of $Z(\mathbf{s})$. In some applications, it might be necessary to add a term $B(\mathbf{s})$ in the observation equation

(1) to explain the potential measurement bias in the data. Here we considered this bias to be negligible and ignored it, following the recommendation of our EPA collaborators.

The true underlying process Z is assumed to follow the model

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + \epsilon(\mathbf{s}), \qquad (2)$$

where $Z(\mathbf{s})$ has a spatial trend, $\mu(\mathbf{s})$, that is a polynomial function of $\mathbf{s}$ with coefficients $\boldsymbol{\beta}$. We assume that $Z(\mathbf{s})$ has zero-mean correlated errors $\epsilon(\mathbf{s})$. The process $\epsilon(\mathbf{s})$ has a possibly nonstationary covariance with parameter vector $\boldsymbol{\theta}$ that might change with location.

We model the output of the EPA physical models as follows:

$$\tilde{Z}(\mathbf{s}) = a(\mathbf{s}) + b(\mathbf{s})Z(\mathbf{s}) + \delta(\mathbf{s}), \qquad (3)$$

where the parameter function $a(\mathbf{s})$ measures the additive bias of the air quality models at location $\mathbf{s}$ and the parameter function $b(\mathbf{s})$ accounts for the multiplicative bias in the air quality models. The process $\delta(\mathbf{s})$ explains the random deviation at location $\mathbf{s}$ with respect to the underlying true process $Z(\mathbf{s})$, and we take it to be a white noise process, namely, $\delta(\cdot) \sim N(0, \sigma_\delta^2)$. The process $\delta(\mathbf{s})$ is independent of $Z(\mathbf{s})$ and $e(\mathbf{s})$, which is the error term for CASTNet. Because the outputs of Models-3 are not point measurements but areal estimations in subregions $B_1, \ldots, B_m$ that cover the domain D, we have

$$\tilde{Z}(B_i) = \int_{B_i} a(\mathbf{s})\, d\mathbf{s} + b \int_{B_i} Z(\mathbf{s})\, d\mathbf{s} + \int_{B_i} \delta(\mathbf{s})\, d\mathbf{s} \qquad (4)$$

for $i = 1, \ldots, m$. According to the EPA modelers with whom we have communicated and to some preliminary analyses we have done, the bias is mostly additive, and their experience suggests treating the function $b(\mathbf{s})$ as constant over space. Therefore, we model $b(\mathbf{s})$ as an unknown constant term and the function $a(\mathbf{s})$ as a polynomial in $\mathbf{s}$ with a vector of coefficients, $\mathbf{a}_0$.

For spatial prediction we simulate values of Z from its posterior predictive distribution as

$$P(Z \,|\, \hat{Z}, \tilde{Z}). \qquad (5)$$

For model evaluation we simulate values of CASTNet given Models-3, assuming that Models-3 output is unbiased, i.e., from the following posterior predictive distribution:

$$P(\hat{Z} \,|\, \tilde{Z}, \mathbf{a}_0 = \mathbf{0}, b = 1). \qquad (6)$$

For bias removal, we use values of the parameters $\mathbf{a}_0$ and b estimated from their posterior distribution

$$P(\mathbf{a}_0, b \,|\, \hat{Z}, \tilde{Z}). \qquad (7)$$

2.2 *Methods for Combining Data at Different Spatial Resolutions*

We first describe the change of support problem that occurs when we combine data sources with different supports, or when the supports of predictand and data are not the same. This problem is treated in detail by Gotway and Young (2002). Here, we have point measurements at the CASTNet

sites, and then we observe the output of Models-3 averaged over grid cells, $B_1, \ldots, B_m$. In this section we discuss algorithms to calculate the covariance for areal measurements and the posterior predictive distribution of a random process at a point location $Z(\mathbf{x}_0)$ given data on block averages, $Z(B_1), \ldots, Z(B_m)$, where some of the blocks might be just points. The covariance for the block averages is

$$\mathrm{cov}(Z(B_i), Z(B_j)) = \int_{B_i} \int_{B_j} C(\mathbf{u}, \mathbf{v}, \boldsymbol{\theta}) \, d\mathbf{u} \, d\mathbf{v} / |B_i||B_j|, \quad (8)$$

where $C(\mathbf{u}, \mathbf{v}, \boldsymbol{\theta}) = \mathrm{cov}(Z(\mathbf{u}), Z(\mathbf{v}))$, C being a possibly nonstationary covariance spatial function. Gelfand, Zhu, and Carlin (2001) approximated the integral in (8) using a random sample over B_j. Here, we prefer a systematic sample because of the computational benefits of having the data on a regular grid.

We now deduce the joint distribution of $\hat{Z}$ and $\tilde{Z}$ conditioning on the value of the parameters in models (1) and (4). We could in principle write this distribution as a function of the parameters and calculate the maximum likelihood estimate (MLE) for the parameters in models (1) and (4). Because in practice this calculation will be hard, we present a Bayesian approach to estimate the parameters. We have

$$\begin{pmatrix} \hat{\mathbf{Z}} \\ \tilde{\mathbf{Z}} \end{pmatrix} \sim \mathcal{N} \left\{ \begin{pmatrix} \hat{\mu} \\ \tilde{a} + b\tilde{\mu} \end{pmatrix}, \begin{pmatrix} \Sigma_C & \Sigma_{CM} \\ \Sigma_{CM} & \Sigma_M \end{pmatrix} \right\}, \quad (9)$$

where

$$\hat{\mu} = (\mu(\mathbf{s}_1), \ldots, \mu(\mathbf{s}_n))^{\mathrm{T}},$$

$$\tilde{a} = \left(\int_{B_1} a(\mathbf{s}) \, d\mathbf{s}, \ldots, \int_{B_m} a(\mathbf{s}) \, d\mathbf{s} \right)^{\mathrm{T}},$$

and

$$\tilde{\mu} = \left(\int_{B_1} \mu(\mathbf{s}) \, d\mathbf{s}, \ldots, \int_{B_m} \mu(\mathbf{s}) \, d\mathbf{s} \right)^{\mathrm{T}}.$$

In (9) Σ_C is the covariance of $\hat{\mathbf{Z}}$ (CASTNet), which is the covariance of Z plus measurement error variance, Σ_M is the covariance of $\tilde{\mathbf{Z}}$ (Models-3), and Σ_{CM} is the crosscovariance between the point measurements $\hat{\mathbf{Z}}$ and the block averages $\tilde{\mathbf{Z}}$. We write Σ to denote the covariance matrix of $(\hat{\mathbf{Z}}^{\mathrm{T}}, \tilde{\mathbf{Z}}^{\mathrm{T}})^{\mathrm{T}}$, so that Σ is an $(n + m) \times (n + m)$ matrix with elements $\{\sigma_{i,j}\}$ given by

$$\sigma_{i_1, i_2} = \mathrm{cov}(\hat{Z}(\mathbf{s}_{i_1}), \hat{Z}(\mathbf{s}_{i_2}))$$
$$= C(\mathbf{s}_{i_1}, \mathbf{s}_{i_2}, \boldsymbol{\theta}) + \mathbf{1}_{\{i_1 = i_2\}} \sigma_e^2 \quad \text{for } i_1, i_2 \leq n,$$

$$\sigma_{n+j, i} = \Sigma_{i, n+j} = \mathrm{cov}(\hat{Z}(\mathbf{s}_i), \tilde{Z}(B_j))$$
$$= b \int_{B_j} C(\mathbf{s}_i, \mathbf{v}, \boldsymbol{\theta}) \, d\mathbf{v} / |B_j|,$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, m$, and

$$\sigma_{n+j_1, n+j_2} = \mathrm{cov}(\tilde{Z}(B_{j_1}), \tilde{Z}(B_{j_2}))$$
$$= b^2 \frac{\int_{B_{j_1}} \int_{B_{j_2}} C(\mathbf{u}, \mathbf{v}, \boldsymbol{\theta}) \, d\mathbf{u} \, d\mathbf{v}}{|B_{j_1}||B_{j_2}|} + \mathbf{1}_{\{j_1 = j_2\}} \sigma_\delta^2 |B_{j_1}|,$$

for $j_1, j_2 = 1, \ldots, m$, where the function $\mathbf{1}_A(\mathbf{x})$ is the indicator function of the set A, taking the value 1 when $\mathbf{x} \in \mathbf{A}$ and 0 otherwise.

The goal is to predict the value of Z at location $\mathbf{x}_0$ given the data. Thus we need the conditional distribution of $Z(\mathbf{x}_0)$ given the observations, assuming that all the parameters are known. We use classical results of multivariate analysis to derive the joint distribution of $Z(\mathbf{x}_0)$, and $\mathbf{Z} = (\hat{\mathbf{Z}}^{\mathrm{T}}, \tilde{\mathbf{Z}}^{\mathrm{T}})^{\mathrm{T}}$. We define $\tau = \mathrm{cov}(Z(\mathbf{x}_0), \mathbf{Z})$, an $(n + m)$-dimensional vector with components

$$\tau_i = \mathrm{cov}\{Z(\mathbf{x}_0), \mathbf{Z}_i\} = \mathrm{cov}\{Z(\mathbf{x}_0), \hat{Z}(\mathbf{s}_i)\}$$
$$= C(\mathbf{x}_0, \mathbf{s}_i, \boldsymbol{\theta}), \quad \text{for } i = 1, \ldots, n,$$

$$\tau_{n+j} = \mathrm{cov}\{Z(\mathbf{x}_0), \mathbf{Z}_{n+j}\} = \mathrm{cov}\{Z(\mathbf{x}_0), \tilde{Z}(B_j)\}$$
$$= b \int_{B_j} C(\mathbf{x}_0, \mathbf{v}, \boldsymbol{\theta}) \, d\mathbf{v} / |B_j|, \quad \text{for } j = 1, \ldots, m,$$

where $\mathbf{Z}_i$ denotes the ith component of $\mathbf{Z}$. We then deduce that the conditional distribution of $Z(\mathbf{x}_0)$ given $\{\hat{\mathbf{Z}}, \tilde{\mathbf{Z}}\}$ is normal with mean $\mu(\mathbf{x}_0) + \tau^{\mathrm{T}} \Sigma^{-1} (\mathbf{Z} - \boldsymbol{\mu})$, where $\boldsymbol{\mu} = (\hat{\mu}, \tilde{a} + b\tilde{\mu})^{\mathrm{T}}$, and variance $\sigma_0^2 - \tau^{\mathrm{T}} \Sigma^{-1} \tau$.

When the goal is to predict Z at a location $\mathbf{x}_0$, the Bayesian solution is the predictive distribution of $Z(\mathbf{x}_0)$ given the observations $\mathbf{Z}$,

$$p(Z(\mathbf{x}_0) \,|\, \mathbf{Z}) = \int p(Z(\mathbf{x}_0) \,|\, \mathbf{Z}, \boldsymbol{\phi}) \, p(\boldsymbol{\phi} \,|\, \mathbf{Z}) \, d\boldsymbol{\phi}, \quad (10)$$

where $\boldsymbol{\phi} = (\sigma_e^2, \mathbf{a}_0, b, \sigma_\delta^2, \boldsymbol{\beta}, \boldsymbol{\theta})$. A Gibbs sampling approach is used to simulate m values from the posterior distribution of the vector parameter $\boldsymbol{\phi}$. The predictive distribution is approximated by the *Rao-Blackwellized estimator* as follows:

$$p(Z(\mathbf{x}_0) \,|\, \mathbf{Z}) = \frac{1}{T} \sum_{t=1}^{T} p(Z(\mathbf{x}_0) \,|\, \mathbf{Z}, \boldsymbol{\phi}^{(t)}), \quad (11)$$

where $\boldsymbol{\phi}^{(t)}$ is the tth draw from the posterior distribution.

2.3 Modeling a Nonstationary Covariance

The spatial patterns shown by the air pollutant fluxes and concentrations change with location, in the sense that the spatial covariance is different at different locations, as shown by Guttorp and Sampson (1994), Haas (1995), and Holland et al. (1999), among others. Therefore, the underlying process Z in (2) is nonstationary and the standard methods of spatial modeling and interpolation are inadequate. In this section we review the methodology of Fuentes (2001, 2002) and Fuentes and Smith (2001), which we will use to model the covariance of the process Z. We represent the process locally as a stationary isotropic random field with some parameters that describe the local spatial structure. These parameters are allowed to vary across space and reflect the lack of stationarity of the process.

A broad class of stationary Gaussian processes may be represented in the form

$$Z(\mathbf{x}) = \int_D K(\mathbf{x} - \mathbf{s}) Z_{\boldsymbol{\theta}(\mathbf{s})}(\mathbf{x}) \, d\mathbf{s}, \quad (12)$$

where K is a kernel function and $Z_{\boldsymbol{\theta}}(\mathbf{x}), \mathbf{x} \in D$ is a family of (independent) stationary Gaussian processes indexed by $\boldsymbol{\theta}$. The parameter $\boldsymbol{\theta}$ is allowed to vary across space to reflect the

lack of stationarity of the process. The stochastic integral (12) is defined as a limit (in mean square) of approximating sums (e.g., Cressie, 1993, p. 107). Each stationary process $Z_{\boldsymbol{\theta}(\mathbf{s})}(\mathbf{x})$ has a mean function $\mu_\mathbf{s}$ that is constant, i.e., $\mu_\mathbf{s}$ does not depend on $\mathbf{x}$. We propose a parametric model for the mean of Z,

$$E\{Z(\mathbf{x})\} = \mu(\mathbf{x}; \boldsymbol{\beta});$$

here we take μ to be a polynomial function of $\mathbf{x}$ with coefficients $\boldsymbol{\beta}$, but other choices are possible.

The covariance of $Z_{\boldsymbol{\theta}(\mathbf{s})}$ is stationary with parameter $\boldsymbol{\theta}(\mathbf{s})$,

$$\mathrm{cov}\{Z_{\boldsymbol{\theta}(\mathbf{s})}(\mathbf{s}_1), Z_{\boldsymbol{\theta}(\mathbf{s})}(\mathbf{s}_2)\} = C_{\boldsymbol{\theta}(\mathbf{s})}(\mathbf{s}_1 - \mathbf{s}_2).$$

We take the process $Z_{\boldsymbol{\theta}(\mathbf{s})}$ to have a Matérn stationary covariance (Matérn, 1960):

$$C_{\boldsymbol{\theta}(\mathbf{s})}(\mathbf{x}) = \frac{\sigma_s}{2^{\nu_s - 1}(\nu_s)}\left(2\nu_s^{1/2}|\mathbf{x}|/\rho_s\right)^{\nu_s} \mathcal{K}_{\nu_s}\left(2\nu_s^{1/2}|\mathbf{x}|/\rho_s\right), \quad (13)$$

where $\mathcal{K}_{\nu_s}$ is a modified Bessel function and $\boldsymbol{\theta}(\mathbf{s}) = (\nu_s, \sigma_s, \rho_s)$. The parameter ρ_s measures how the correlation decays with distance; generally this parameter is called the *range*. The parameter σ_s is the variance of the random field, i.e., $\sigma_s = \mathrm{var}(Z_{\boldsymbol{\theta}(\mathbf{s})}(\mathbf{x}))$, usually referred to as the *sill*. The parameter ν_s measures the degree of smoothness of the process $Z_{\boldsymbol{\theta}(\mathbf{s})}$. The higher the value of ν_s the smoother $Z_{\boldsymbol{\theta}(\mathbf{s})}$ would be.

The covariance of Z, $C(\mathbf{s}_1, \mathbf{s}_2; \theta)$, is a convolution of the local covariances $C_{\theta(\mathbf{s})}(\mathbf{s}_1 - \mathbf{s}_2)$,

$$C(\mathbf{s}_1, \mathbf{s}_2; \boldsymbol{\theta}) = \int_D K(\mathbf{s}_1 - \mathbf{s})K(\mathbf{s}_2 - \mathbf{s})C_{\boldsymbol{\theta}(\mathbf{s})}(\mathbf{s}_1 - \mathbf{s}_2)\, d\mathbf{s}. \quad (14)$$

In (14) every entry requires an integration. Because each such integration is actually an expectation with respect to a uniform distribution, we propose Monte Carlo integration. We draw a systematic sample of locations $\mathbf{v}_j$, $j = 1, 2, \ldots, M$ over D. Hence, we replace $C(\mathbf{s}_1, \mathbf{s}_2; \boldsymbol{\theta})$ with

$$\begin{aligned} &C_M(\mathbf{s}_1, \mathbf{s}_2; \boldsymbol{\theta}) \\ &= M^{-1} \sum_{j=1}^M K(\mathbf{s}_1 - \mathbf{v}_j)K(\mathbf{s}_2 - \mathbf{v}_j)C_{\boldsymbol{\theta}(\mathbf{v}_j)}(\mathbf{s}_1 - \mathbf{s}_2). \end{aligned} \quad (15)$$

This is a Monte Carlo integration, which can be made arbitrarily accurate and does not involve the data $\mathbf{Z}$. The sampling points $\mathbf{v}_j$, $j = 1, 2, \ldots, M$, determine subregions of local stationarity for the process Z. We increase the value of M until convergence is achieved.

2.4 *Algorithm for Estimation and Prediction*

In our Gibbs sampling approach there are three stages. We alternate between the parameters that measure the lack of stationarity, $(\boldsymbol{\beta}, \boldsymbol{\theta}(\mathbf{s}))$ (Stage 1), the parameters that measure the bias of Models-3 and the measurement error of CASTNet (Stage 2), and the unobserved true values of Z at all the CASTNet sites and at the blocks where we have the Models-3 output (Stage 3). We obtain the conditional posterior distribution of the parameters that measure the lack of stationarity, $(\boldsymbol{\beta}, \boldsymbol{\theta}(\mathbf{s}))$, given the values of Z that are updated in Stage 3. The posterior distribution of $(\boldsymbol{\beta}, \boldsymbol{\theta}(\mathbf{s}))$ will be completely specified once we define the priors for $(\boldsymbol{\beta}, \boldsymbol{\theta}(\mathbf{s}))$, because $(Z \mid \boldsymbol{\beta}, \boldsymbol{\theta})$ is Gaussian.

We use a Metropolis–Hastings step for blocks of parameters, after performing partial marginalizations of the full conditionals. We treat as a block $\boldsymbol{\theta}$, which are the three covariance parameters, namely, the sill, the range, and the smoothness. In our experience, this scheme produces a chain with better mixing properties than that obtained by independent sampling of the full conditional distributions of the sill, range, and smoothness parameters. We use a gamma-prior distribution for all the covariance parameters, except for the sill parameter; we use a uniform prior for the logarithm of the sill. For the $\boldsymbol{\beta}$ parameter, we use uniform priors.

3. Application: Air Pollution Data

We model the covariance function for the process Z, the true SO_2 values, using equation (14). We estimate the covariance parameters of the process Z, given the CASTNet data and Models-3 output, taking into consideration the change-of-support problem. We calculate the covariances involving block averages by drawing a systematic sample of L locations in each pixel, so that M is equal to L times the number of pixels. This is an approximation that can be made arbitrarily accurate by increasing the value of L. In this application $L = 4$ seemed to be large enough to achieve a sufficiently good approximation. We sample systematically rather than randomly because it allows us to use the fast Fourier transform, with which the covariance calculations are faster and easier.

We implement the nonstationary model (12) with weight function $K(\mathbf{u} - \mathbf{s}) = \frac{1}{h^2}K_0(\frac{\mathbf{u}-\mathbf{s}}{h})$, where $K_0(\mathbf{u})$ is the quadratic weight function

$$K_0(\mathbf{u}) = \tfrac{3}{4}\left(1 - u_1^2\right)_+ \tfrac{3}{4}\left(1 - u_2^2\right)_+, \quad (16)$$

for $\mathbf{u} = (u_1, u_2)$. The bandwidth parameter h is defined as $(1 + \epsilon)\, l/2$, where l is the distance between neighboring sample points $\mathbf{v}_1, \ldots, \mathbf{v}_M$ in (15), and ϵ is a value between 0 and 1. For ϵ we use a uniform prior in the interval $[0, 1]$. The parameter ϵ determines the overlap between the subregions of stationarity centered at the sampling points $\mathbf{v}_1, \ldots, \mathbf{v}_M$, and h can be interpreted as the diameter of the subregions of stationarity. We use gamma priors for all the Matérn covariance parameters, except for the sill parameter for which we use $p(\sigma) \propto \sigma^{-1}$, which is a uniform prior for $\log(\sigma)$. The mean and variance of the gamma priors are determined using previous knowledge and results from the analysis of similar data. For the smoothness parameter the prior mean is 0.5 and the variance 1, and for the range parameter the prior mean is 100 km and the prior standard deviation is 45 km.

In Figure 2 we plot the posterior distribution of $\sigma(\mathbf{x})$ at some CASTNet sites $\mathbf{x}$. The spatial locations at which we examine the covariance parameters in Figure 2 have nothing to do with the choice of the nodes $\mathbf{v}_1, \ldots, \mathbf{v}_M$ in equation (15). We are assuming that the covariance parameters are approximately constant between nodes; otherwise we increase the number of nodes, M. Thus, the value of $\boldsymbol{\theta}$ at some location $\mathbf{x}$ of interest is calculated as the value of $\boldsymbol{\theta}$ at the node $\mathbf{v}_j$ that is closest to $\mathbf{x}$.

The sill parameter changes with location as illustrated by the variation in the distributions in Figure 2. The mean and standard deviation of the posterior distributions for the sill parameter at six selected sites located in Maine, Illinois, North
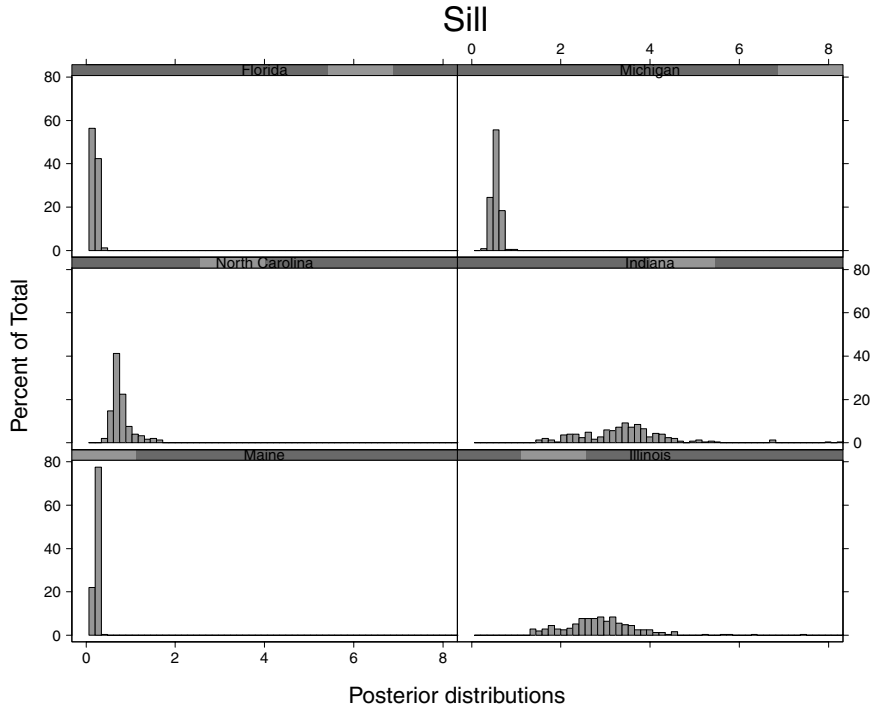
Figure 2. Posterior distributions of the sill parameter of the Matérn covariance for the SO_2 concentrations of Z, for the week starting July 11, 1995, at six selected locations.

Carolina, Indiana, Florida, and Michigan, are, respectively, 0.23 (0.03), 2.89 (0.85), 0.78 (0.23), 3.40 (0.99), 0.21 (0.04), and 0.54 (0.08). The range parameter does not change much with location. The means and standard deviations of the posterior distributions for the range parameter at the same six sites are 97.7 (12.8), 95.7 (28.5), 119.0 (35.8), 91.9 (27.3), 114.9 (21.9), and 96.2 (15.5). The smoothing parameter does not change much with location either, and is always close to 1/2 (exponential). The mode of the posterior distribution of the parameter that measures the measurement error for CASTNet is 0.8, and for Models-3 it is 0.1.

We use a uniform prior distribution for all the additive bias parameters, and a normal prior for the multiplicative bias parameter b. The mode of the posterior distribution for the parameter that measures the multiplicative bias for Models-3

is 0.5 with a standard error of 0.5, and for the additive bias we have a polynomial of degree q. We treat q as a hyperparameter and use a reversible jump Markov Chain Monte Carlo (Green, 1995; Denison, Mallick, and Smith, 1998) algorithm for model selection. In this application $q = 4$ seemed to be large enough.

In Table 1 we show the modes, standard deviations, and 90% credible intervals of the posterior predictive distribution (5) for SO_2 at the six selected sites. We get very high variability at the Indiana site. This site is very close to several coal power plants, and so the SO_2 levels can be very high or very low depending on wind speed, wind direction, and the atmospheric stability. The sites in Maine and Florida have the lowest SO_2 levels and variability. The agricultural site in Illinois and the site in North Carolina have similar behavior in terms of SO_2 levels. The site in North Carolina is not far

Table 1

Columns 2–5 in this table show the modes, standard deviations, and 90% credible intervals of the posterior predictive distribution (equation [5]) for the underlying process Z measuring the true SO_2 concentrations at the six selected sites. Column 6 shows the CASTNet values ($\hat{Z}$). Columns 7–9 show the modes and the corresponding 90% credible intervals of the posterior predictive distribution (equation [6]) for model evaluation.

Site	Mode	SD	90%	CI	CASTNet	Models-3	90%	CI
Maine	0.18	0.02	0.15	0.25	0.15	0.33	0.10	0.43
Illinois	2.80	0.25	2.55	3.41	3.29	3.33	2.17	5.03
North Carolina	1.98	0.29	1.18	2.08	0.90	5.32	3.67	6.67
Indiana	0.98	1.59	0.74	5.63	3.14	9.59	4.20	20.50
Florida	0.56	0.05	0.50	0.69	0.57	0.52	0.20	0.80
Michigan	0.83	0.12	0.79	1.14	1.02	1.04	0.53	1.70

Figure 3. Map of the estimated additive spatial bias of Models-3. The bias parameters are estimated by the means of their posterior distributions.

from the Tennessee power plants, and the site in Illinois is also relatively close to some midwestern power plants. The site in Michigan, which is very close to Lake Michigan and relatively far from power plants also has low SO_2 levels.

In Table 1 we also show the CASTNet values, to judge whether the generated data are similar to the CASTNet data. The CASTNet values in Table 1 at the six sites represent the 5th, 91st, 1st, 68th, 14th, and 79th percentiles of the corresponding posterior predictive distributions from the Bayesian melding approach. For the North Carolina site the CASTNet values are low relative to the posterior predictive distribution. This is due to the higher altitude (and lower pollution levels) of this particular site in relation to the nearby locations.

The last three columns in Table 1 are the modes and the corresponding 90% credible intervals of the posterior predictive distribution (6) for model evaluation. We can clearly appreciate the bias in the numerical models by comparing these values with CASTNet. For instance, for the site in North Carolina, the interpolated value of Models-3 is 5.32 ppb (SE = 3.00 ppb), while the CASTNet value is only 0.90 ppb. We can remove the bias in these interpolated Models-3 values by taking into account the additive bias measured by $a(\mathbf{x})$ (a polynomial of degree 4 with coefficients $\mathbf{a}_0$) and the multiplicative bias. At each site we calculate the posterior means of $\mathbf{a}_0$ and b, using the posterior distribution (7), and we obtain the following adjusted Models-3 values (adjusted value(s) = $((\text{Models-3})(\mathbf{s}) - \tilde{a}(\mathbf{s}))/\tilde{b}$) at the six selected sites: 0.12, 2.88, 1.09, 3.12, 0.44, and 1.01, where $\tilde{a}$ is a fourth-degree polynomial with coefficients that are the posterior means of $\mathbf{a}_0$, and $\tilde{b}$ is the posterior mean of b. These adjusted values are very similar to CASTNet. Figure 3 shows a contourplot for the additive bias of Models-3, with the bias parameters, $\mathbf{a}_0$, estimated with the mean of their posterior distribution. This bias seems to be larger in a north–south ridge covering some of the midwestern and the southeastern parts of our domain. The week of July 11 (1995) was very dry with a heat wave affecting the midwest and southeast of the United States, this probably affected the ability of the numerical models to

estimate correctly the mixing heights (the depth of the unstable air in the boundary layer) that determines pollutant trajectories.

Figure 4A shows the predicted values of SO_2 at different locations on a regular grid, using our Bayesian melding approach for prediction to combine CASTNet and Models-3 data. The predicted values in Figure 4A are the means of the posterior predictive distribution (equation [5]) for the SO_2 data. This graph looks similar to the output of Models-3 shown in Figure 1B. However, the SO_2 values in Figure 4A are between 0 and 8 ppb, while in Figure 1B the SO_2 values are between 0 and 40 ppb. The range of values in Figure 4A is more reasonable, and it is closer to the range of values for CASTNet shown in Figure 1A. This illustrates the effectiveness of the Bayesian melding approach for correcting the bias in Models-3. Figure 4B shows the standard error of the posterior predictive distribution at each point. There is higher uncertainty in areas close to power plants, where there is more variability due in part to the effect of metereological variables such as wind.

Figure 5 is an illustration of the performance of our Bayesian melding approach. In this figure we plot the CASTNet values $\hat{Z}_i$ (at each site i) versus the mean of the posterior predictive distribution of the truth Z at site i given the models output $\tilde{Z}$ and the CASTNet values $\hat{Z}_{-i}$, at all sites except at the site i that we are predicting. The dotted lines show the 90% pointwise credible bounds for CASTNet, and the solid line has slope 1 and intercept 0. The average length of the 90% credible intervals for the SO_2 values using the Bayesian melding approach over all CASTNet sites is 7 ppb and the standard deviation of the length of the intervals is 8 ppb. There is quite a lot of variability in the lengths of the intervals, mainly due to the lack of stationarity of the spatial covariance. We compare this mean interval length to that of the 90% credible intervals for the SO_2 values at all CASTNet sites, this time given only the Models-3 output, and not conditional on the CASTNet values. This is only 3.5 ppb, half of the average interval length given Models-3 and the
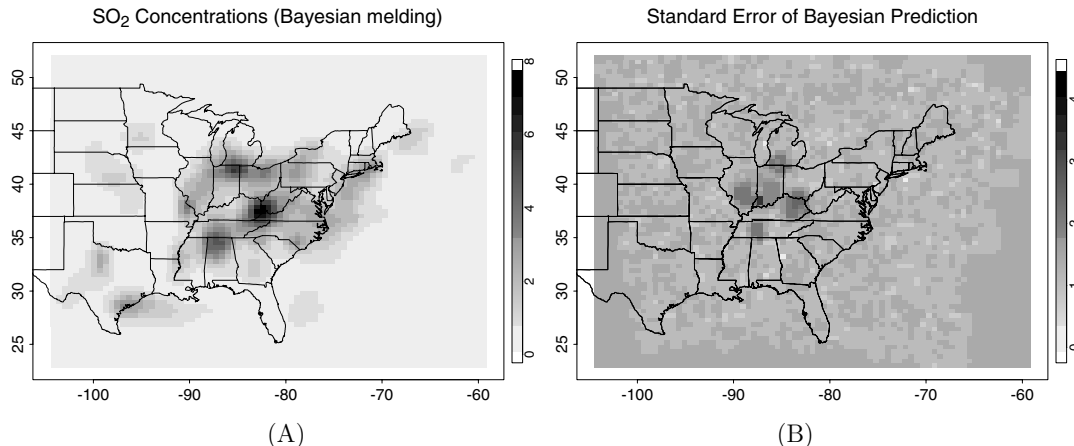
Figure 4. Predicted SO_2 concentrations using the Bayesian melding approach to combine CASTNet with Models-3 data: (A) mean and (B) standard deviation of the posterior predictive distribution of the underlying process Z given CASTNet and Models-3.

CASTNet values. These intervals are too overoptimistic, mainly due to the fact that the output of Models-3 is smoother than the real data. The calibration of the intervals based on Models-3 alone is poor: for Models-3 alone, only 55% of the 90% credible intervals contain the CASTNet values. We improve the calibration to 80% by including CASTNet sites (see Figure 5).

4. Discussion

Air quality numerical models are used to examine the response of the air pollution network to different control strategies under various high-pollution scenarios. To establish their credibility, however, it is essential that they should accurately reproduce observed measurements when applied to ground data. Our objectives are model evaluation and bias removal

Figure 5. CASTNet values of SO_2 versus the mean of the predictive posterior distribution of the true SO_2 values given Models-3 and CASTNet values at all sites, except for the site at which the predictions are being made. The dotted lines show the 90% pointwise crossvalidation predictive credible bands for CASTNet.

for the air quality numerical models, and construction of reliable maps of air pollution combining the output of numerical models with air pollution measurements at monitoring sites. We evaluate air quality models by obtaining the posterior predictive distribution of the measurements at the monitoring sites given the numerical models output. We remove the bias in the air quality models by obtaining the posterior distribution of the bias parameters given the measurements at the monitoring sites and the numerical models output. We construct maps of air pollutants simulating values from the posterior predictive distribution of the true values (underlying process) given the measurements at the monitoring sites and the numerical models output.

The Bayesian approach provides a natural way to combine data from different sources taking into account the different uncertainties, and it also provides posterior distributions of quantities of interest that can be used for scientific inference. However, our approach could also be implemented using a geostatistical approach by predicting the truth with the optimal predictor, namely, the mean of the conditional distribution of the truth given the data and the parameters. For this, we would use an iterative approach with two steps. In step 1, we would obtain the predictor of the truth given the data and some initial values of the parameters. In step 2, the predicted values obtained in step 1 combined with the data would be used to estimate the bias and covariance parameters, using a likelihood approach. We would then iterate between steps 1 and 2.

Another approach to model evaluation is to use spatiotemporal models for monitoring data to provide estimates of average concentrations over grid cells corresponding to model prediction (Dennis et al., 1990; Chu, 1996; Sampson and Guttorp, 1998; Davis, Nychka, and Barley, 2000). This approach is reasonable when the monitoring data are dense enough that we can fit an appropriate spatiotemporal model to the data. In situations such as the one presented here, with few and sparse data points that show a lack of stationarity, the interpolated grid square averages would be poor because of the sparseness of the CASTNet network, and so treating them as ground truth for model evaluation would be questionable.
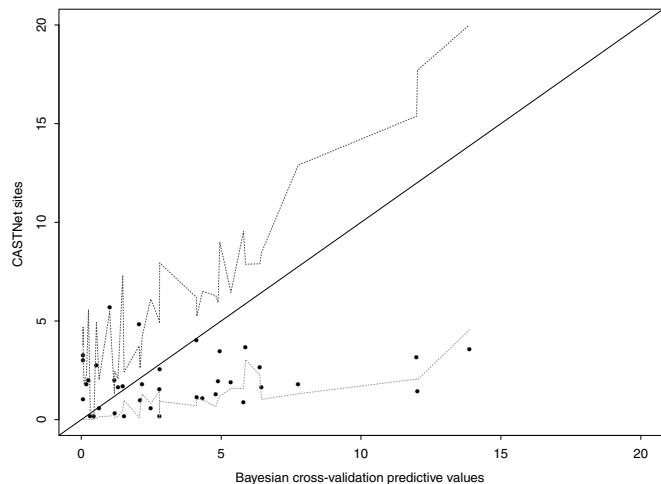
We model the underlying process Z, with the true values of fluxes/concentrations of air pollution, as a spatial process with nonstationary covariance. Misspecification of the covariance would lead to poor estimates of the predictive standard deviation, and possibly also biased predictions. For instance, suppose we assume that the sill is constant, while it is actually changing with location. Then, we would tend to overestimate the prediction error in areas with smaller sill than the proposed fixed value, and we would probably underestimate the prediction error in areas with larger sill than the assumed fixed value. Here we represent the process locally as a stationary isotropic random field, but allow the parameters of the stationary random field to vary across space. With this model we are able to make inferences about the nonstationary random field with only one realization of the process. We have used a model that assumes the spatial covariance structure to be approximately piecewise constant between nodes. An alternative approach would be to model the spatial covariance parameter θ using splines, when we would have a continuous function of space. However, when the number of nodes M is large enough, our results using splines are similar to those from the simpler approach we have presented here.

We have overcome some limitations of earlier approaches to the evaluation of numerical models and mapping of air quality, but other limitations remain. In our analysis we have modeled the true underlying pollution process as a Gaussian field, which seemed to be a reasonable assumption because we were working with weekly averages of hourly values. However, for hourly pollution levels the logarithmic transformation could be more appropriate. Based on the experience of our EPA collaborators, the main problem with air quality models seems to be the presence of additive bias. This is the reason for our representation of the numerical models as a linear function of the truth. In other applications, one might need to write the output of a numerical model as a nonlinear function of the truth. We have also ignored the potential additive bias of the observations. Preliminary analysis showed that this bias was negligible compared to the bias of the Models-3 output. We assumed independence between the truth and the measurement error. To our knowledge, this assumption is reasonable for SO_2 concentrations. But, it is more questionable for other air pollution variables, such as fluxes and deposition velocities. Thus, when modeling other pollution variables, a more complicated dependency structure might be needed.

The only monitoring network for dry deposition and concentrations is CASTNet, and the CASTNet monitoring sites are all in rural areas. Because Models-3 are regional models, the main interest is to capture regional patterns, rather than isolated hot spots (probably located in urban areas). Therefore, CASTNet seems to be an appropriate network for evaluation of these regional air quality models, despite the fact that all sites are rural. However, we plan to use in the future urban sites whenever they become available.

In this article we combine data by relating the spatially varying variables to an underlying unobservable true air pollution process and we predict this latent process. This is different from the traditional geostatistical approaches of cokriging or kriging with external drift (KED) for spatial interpolation with auxiliary covariates, as used by Phillips et al. (1997). In cokriging or KED the data are represented as a function of

covariates using a regression model. Therefore, if the observed data and the output of numerical models are measured at different spatial resolutions, we cannot compare the observed data directly to the output of numerical models. In this article we overcome that problem by relating each source of information to the unobserved ground truth.

An alternative approach to model evaluation would use kriging to predict the output of the numerical models at the CASTNet locations. We compared our proposed modeling approach with this traditional stationary kriging prediction method that ignores the uncertainty in the covariance parameters. Both methods tend to give similar predicted values; the main difference is in the prediction error, which is the critical factor for any inference done with the data. Our approach gave larger standard errors than kriging for the Models-3-predicted values, especially in areas close to power plants. This reflects the difficulty in estimating the covariance parameters in these areas. The kriging approach reported the same standard errors everywhere. For the evaluation of the physical models, we studied where CASTNet values lay with respect to a confidence interval for the SO_2 predicted values from Models-3. Using kriging we considerably underestimated the lengths of the prediction intervals leading to wrong conclusions about the performance of Models-3.

Résumé

La construction de cartes de niveaux de pollution par dépôts secs est fondamentale pour la gestion de la qualité de l'air, et présente des problèmes statistiques typiques de nombreuses applications spatiales et environnementales. Ces cartes devraient idéalement reposer sur un réseau dense de stations de suivi, mais ce n'est jamais le cas. Aux Etats-Unis, il existe deux principales sources d'information sur les dépôts secs: le réseau de stations de mesures de pollution CASTNet (réseau lâche de 50 stations environ), et les sorties d'un modèle régional de prédiction de la qualité de l'air appelé Models-3. L'évaluation de ces modèles numériques pour leurs applications en qualité de l'air constitue un problème annexe mais crucial pour le choix de stratégies de contrôle. Nous développons simultanément des méthodes formelles de combinaison de sources d'information à différentes résolutions spatiales et d'évaluation des modèles numériques. Nous écrivons un modèle simple combinant les sorties de Models-3 et les observations de CASTNet sous forme d'une "donnée terrain réelle non observée," et nous estimons le modèle par une

méthode bayésienne. Ceci permet (1) une prédiction spatiale améliorée, (2) la validation des prédictions de Models-3 via la distribution prédictive a posteriori des observations de CAST-Net, et (3) l'élimination du biais dans les sorties de Models-3. Nous appliquons notre méthode à des données de concentration de SO2, ce qui nous permet de produire des cartes de distribution à haute résolution du SO2 en combinant les données et les prédictions du modèle numérique. Nous concluons aussi que le modèle numérique se comporte moins bien à proximité des centrales thermiques, où il surestime les valeurs de SO2.

References

Best, N. G., Ickstadt, K., and Wolpert, R. L. (2000). Spatial Poisson regression for health and exposure data measured at disparate resolutions. *Journal of the American Statistical Association* **95,** 1076–1088.

Chu, S.-H. (1996). ROM2.2 model performance in the July 3–11, 1988 ozone episode. Transactions of the 1996 AMS-AWMA Joint Meeting on ROM Evaluation.

Clarke, J. F. and Edgerton, E. S. (1997). Dry deposition calculations for the Clean Air Status and Trends Network. *Atmospheric Environment* **31,** 3667–3678.

Cowles, M. K. and Zimmerman, D. L. (2002). Combining Snow Water Equivalent data from multiple sources to estimate spatio-temporal trends and compare measurement systems. *Journal of Agricultural, Biological and Environmental Statistics* **7,** 536–557.

Cowles, M. K. and Zimmerman, D. L. (2003). A Bayesian space-time analysis of acid deposition data combined from two monitoring networks. *Journal of Geophysical Research* **108,** 9006.

Cressie, N. A. (1993). *Statistics for Spatial Data*, revised edition. New York: Wiley.

Davis, J. M., Nychka, D., and Bailey, B. (2000). A comparison of regional oxidant model (ROM) output with observed ozone data. *Atmospheric Environment* **34,** 2413–1423.

Denison, D. G. T., Mallick, B. K., and Smith, A. G. M. (1998). Automatic Bayesian curve fitting. *Journal of the Royal Statistical Society B* **60,** 333–350.

Dennis, R. L., Barchet, W. R., Clark, T. L., and Seilkop, S. K. (1990). Evaluation of regional acidic deposition models (Part 1), NAPAP SAS/T Report 5. In *National Acid Precipitation Assessment Program: State of Science and Technology*, Volume 1. Washington, D.C.: National Acid Precipitation Assessment Program.

Dennis, R. L., Byun, D. W., Novak, J. H., Galluppi, K. L., Coats, C. J., and Vouk, M. A. (1996). The next generation of integrated air quality modelling: EPA's Models-3. *Atmospheric Environment* **30,** 1925–2938.

Dolwick, P. D., Jang, C., Possiel, N., Timin, B., Gipson, G., and Godowitch, J. (2001). Summary of results from a series of Models-3/CMAQ simulations of ozone in the Western United States. 94th Annual A & WMA Conference and Exhibition, Orlando, Florida.

Fuentes, M. (2001). A new high frequency kriging approach for nonstationary environmental processes. *Environmetrics* **12,** 469–483.

Fuentes, M. (2002). Spectral methods for nonstationary spatial processes. *Biometrika* **89,** 197–210.

Fuentes, M. and Smith, R. (2001). *A new class of nonstationary models*. Technical report, Institute of Statistics Mimeo Series 2534, North Carolina State University.

Gelfand, A. E., Zhu, L., and Carlin, B. P. (2001). On the change of support problem for spatio-temporal data. *Biostatistics* **2,** 31–45.

Gotway, C. A. and Young, L. J. (2002). Combining incompatible spatial data. *Journal of the American Statistical Association* **97,** 632–648.

Green, P. J. (1995). Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82,** 711–732.

Guttorp, P. and Sampson, P. D. (1994). Methods for estimating heterogeneous spatial covariance functions with environmental applications. In *Handbook of Statistics 12*, G. P. Patil and C. R. Rao (eds), 661–689. Amsterdam: Elsevier Science B.V.

Haas, T. C. (1995). Local prediction of a spatio-temporal process with an application to wet sulfate deposition. *Journal of the American Statistical Association* **90,** 1189–1199.

Holland, D., Saltzman, N., Cox, L. H., and Nychka, D. (1999). Spatial prediction of sulfur dioxide in the eastern United States. In *geoENV II: Geostatistics for Environmental Applications*, J. Gómez-Hernández, A. O. Soares, and R. Froidevaux (eds), 65–76. Dordrecht: Kluwer.

Matérn, B. (1960). Spatial variation. In *Meddelanden från Statens Skogsforskningsinstitut*, Volume 49, No. 5. Stockholm: Almaenna Foerlaget. Berlin: Springer-Verlag, 2nd edition, 1986.

Phillips, D. L., Lee, H. E., Herstrom, A. A., Hogsett, W. E., and Tingey, D. T. (1997). Use of auxiliary data for spatial interpolation of ozone exposure in southeastern forests. *Environmetrics* **8,** 43–61.

Poole, D. and Raftery, A. E. (2000). Inference for deterministic simulation models: The Bayesian melding approach. *Journal of the American Statistical Association* **95,** 1244–1255.

Sampson, P. D. and Guttorp, P. (1998). Operational Evaluation of Air Quality Models. Proceedings of a Novartis Foundation Symposium on Environmental Statistics.

Wikle, C. K., Milliff, R. F., Nychka, D., and Berliner, M. (2001). Spatiotemporal hierarchical Bayesian modeling: Tropical ocean surface winds. *Journal of the American Statistical Association* **95,** 1076–1987.