A Bayesian approach to combine disparate spatial data [1]

Montserrat Fuentes
North Carolina State University

and

Adrian E. Raftery
University of Washington

[1]Montserrat Fuentes is an assistant professor at the Statistics Department, North Carolina State University (NCSU), Raleigh, NC 27695-8203, and a visiting scientist at the US Environmental Protection Agency (EPA). Tel.:(919) 515-1921, Fax: (919) 515-1169, E-mail: fuentes@stat.ncsu.edu. Web: www.stat.ncsu.edu/ fuentes A. Raftery is Professor of Statistics and Sociology, University of Washington, Seattle, WA 98195-4320, Email: raftery@stat.washington.edu, Web: www.stat.washington.edu/raftery.
Fuentes' research was supported by a National Science Foundation grant DMS 0002790 and by a US EPA award R-8287801. This research was also sponsored by the Department of Defense Multidisciplinary Research Program of the University Research Initiative under grant number N-00014-01-10745. The authors would like to acknowledge the helpful insight and assistance provided by the US EPA Office of Research Development, RTP, NC, in particular by Peter Finkelstein, John Pleim and Robin Dennis. The authors are also grateful to Tilmann Gneiting for helpful comments.

ABSTRACT

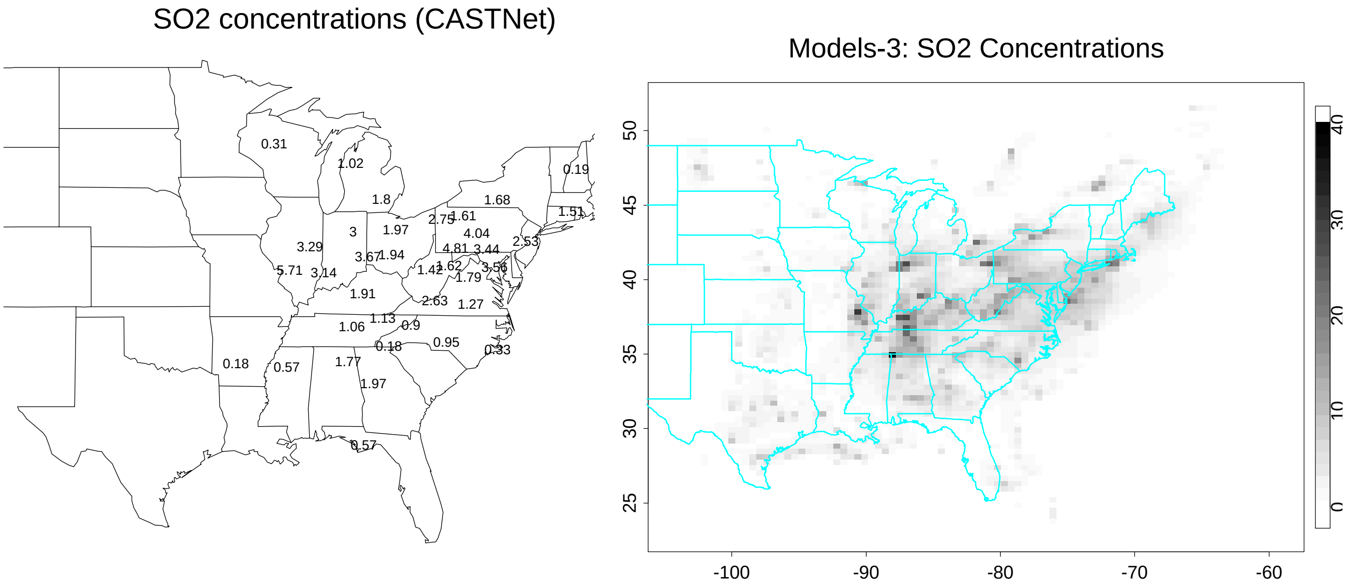
Constructing maps of pollution levels is vital for air quality management, and presents statistical problems typical of many environmental and spatial applications. Ideally, such maps would be based on a dense network of monitoring stations, but this does not exist. Instead, there are two main sources of information in the U.S.: 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 that is crucial to assist in control strategy selection.

Here 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 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 by estimating additive and multiplicative bias parameters in the model. We applied our methods to data on SO_2 concentrations, and we obtained high resolution SO_2 distributions by combining observed data with model output. We also concluded that the numerical models perform worse in areas closer to power plants, where the SO_2 values are overestimated by the models.

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 loadings of pollutants to aquatic and terrestrial ecosystems are useful over different geo-political boundaries, to discover when, where, and to what extent the pollution load is improving or declining. Ideally, such maps would be based on a dense network of monitoring stations, covering most of the U.S., at which fluxes 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 pollution levels in the U.S., and two resulting ways of constructing pollution maps. The first is a *sparse* set of about 50 irregularly spaced sites in the Eastern U.S., the Clean Air Status Trend Network (CASTNet), at which the EPA regularly measures concentrations and fluxes of different atmospheric pollutants (see Figure 1 (a)). It would be possible to use an interpolation method to produce 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 CASTNet 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 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 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 physics; they take the form of huge blocks of



(a)

(b)

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. The resolution is 36 km^2.

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 U.S. (see Figure 1 (b)).

The output of Models-3 generates averaged concentrations/fluxes over regions of dimensions 36km$\times$36km. This approach also may 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 validation, 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. The high risk of getting the wrong control strategy has costly economic and social consequences. The objective comparison of modeled concentrations with observed field data is an approach to assessment of model performance. Early evaluations of model performance usually relied on linear least-squares analysis of observed versus modeled values, using traditional scatter plots of the values, e.g. Clarke, (1964), Martin, (1971), and Hanna (1971). Further development of these proposed statistical evaluation procedures is needed, and we propose a Bayesian approach. The 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 large 36km $\times$ 36km 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 validation may be questionable.

A related problem is that the comparison does not take into account the uncertainty in the interpolated values. In this paper, we develop a new approach to the model validation 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 datafield 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 validation 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 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 is 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 was proposed by Best et al. (2000), who relate 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 unobservable process, but the statistical model we propose is novel. We present a new way to relate air pollution variables to an underlying process, with 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 our approach we have a simultaneous representation of the data in terms of the underlying truth. The way we estimate the 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, in Section 3 we show how to estimate it, and in Section 4 we show some of our results for model validation and map construction using the combined data for the air pollution problem.

2 The Statistical Model

2.1 General Framework

We do not treat CASTNet measurements as the "ground truth". Instead, we assume that there is some smooth underlying (but unobserved) field that measures the "true" concentration/flux of the pollutant at each location. CASTNet data are these "true values" plus some measurement error. The Models-3 output can also be written in terms of this true underlying (unobservable) process, with some parameters that explain the bias and microscale noise in Models-3. The *truth* is assumed to be a smooth underlying spatial process with some parameters that explain the large scale and short scale dependency structure of the air pollutants.

Our objectives are model validation and bias removal for Models-3, and construction of maps of air pollution combining Models-3 and CASTNet. We evaluate Models-3 by obtaining the posterior predictive distribution of CASTNet given Models-3 output. We remove the bias in Models-3 by obtaining the posterior distribution of the bias parameters given CASTNet data and Models-3 output. We construct maps of air pollutants simulating values from the posterior predictive distribution of the true values (underlying process) given CASTNet data and Models-3 output.

2.2 Statistical Models for CASTNet and Models-3 Output

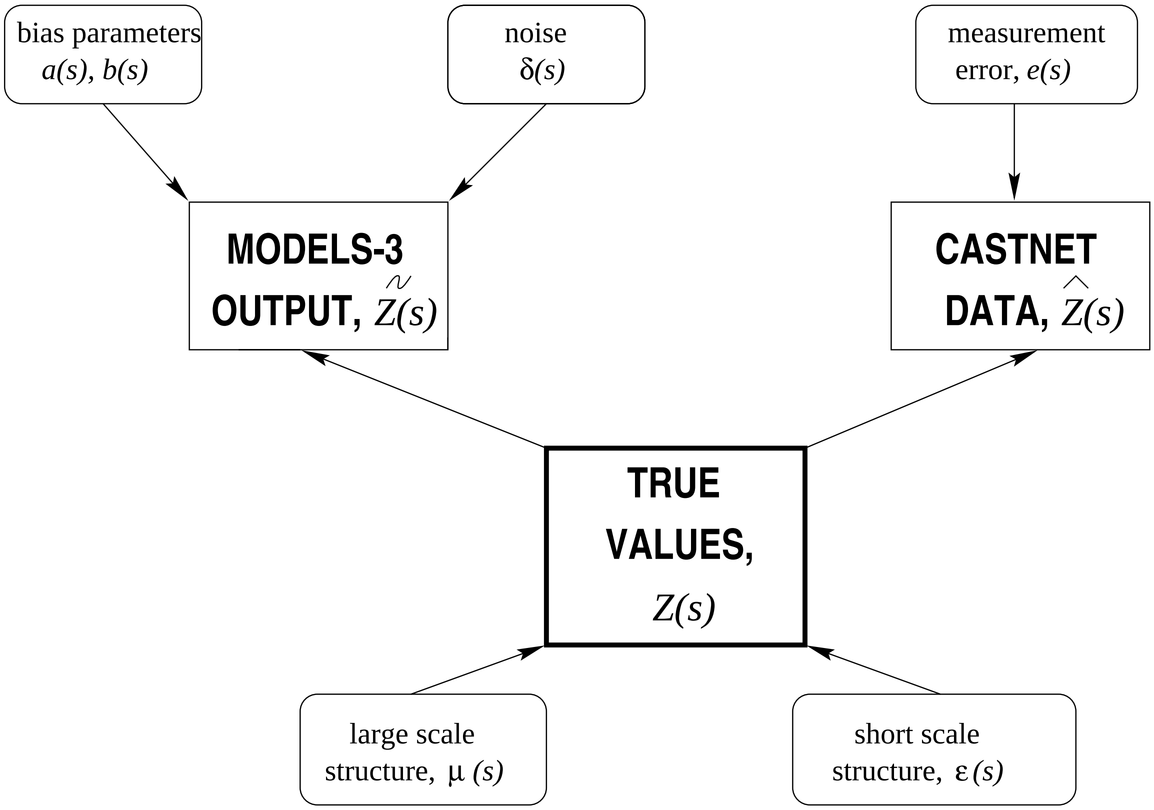


Figure 2: General Modeling Framework.

Our general modeling framework is shown in Figure 2. We do not consider CASTNet measurements to be the "ground truth", because there is measurement error. Thus, we assume 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 at station $\mathbf{s}$ by $\hat{Z}(\mathbf{s})$, and we assume that

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

where $e(\mathbf{s}) \sim N(0, \sigma_e^2)$ represents the measurement error at location $\mathbf{s}$. The process $e(\mathbf{s})$ is independent of $Z(\mathbf{s})$.

The true underlying process Z is a spatial process with a nonstationary covariance,

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

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 nonstationary covariance with parameter vector $\boldsymbol{\theta}$ that might change with location.

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

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

Here, 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 $\delta(\mathbf{s}) \sim N(0, \sigma_\delta^2)$ explains the random deviation at location $\mathbf{s}$ with respect to the underlying true process $Z(\mathbf{s})$. The process $\delta(\mathbf{s})$ is independent of $Z(\mathbf{s})$ and $e(\mathbf{s})$, which is the error term for CASTNet. Since 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

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

for $i = 1, \ldots, m$. According to the EPA modelers with whom we have communicated the bias is mostly additive, and their experience suggests treating the function $b(\mathbf{s})$ as constant over space. Therefore, we model the function $a(\mathbf{s})$ as a polynomial in $\mathbf{s}$ with a vector of coefficients, a_0, and b as a unknown constant term.

For spatial prediction we simulate values of Z from the posterior predictive distribution:

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

For model validation we simulate values of CASTNet given models-3, from the following posterior predictive distribution:

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

For bias removal we simulate values of the parameters a and b from the posterior distribution:

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

2.3 Change of Support

The change-of-support problem 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 with detail in Cressie (1993, p. 284). Here, we have point measurements at the CASTNet sites, and then we observe the output of Models-3 averaged over regions, $B_1, \ldots, B_m$, of dimensions 36km $\times$ 36km.

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 a point.

The covariance for the block averages is

$$\text{cov}(Z(B_i), Z(B_j)) = \int_{B_i} \int_{B_j} C(\mathbf{u}, \mathbf{v}) d\mathbf{u} d\mathbf{v} / |B_i||B_j|, \tag{8}$$

where

$$C(\mathbf{u}, \mathbf{v}) = \text{cov}(Z(\mathbf{u}), Z(\mathbf{v})),$$

C being a possibly nonstationary spatial function. If $B_i = \mathbf{s_i}$ (a point) the covariance is defined by

$$\text{cov}(Z(\mathbf{s}_i), Z(B_j)) = \int_{B_j} C(\mathbf{s}_i, \mathbf{v}) d\mathbf{v} / |B_j|. \tag{9}$$

In practice, for each pixel B_j we draw an independent set of locations $\mathbf{u}_{jr}$, $r = 1, 2, \ldots, L_j$ uniformly over B_j, and we approximate the integral in (9) with the following expression:

$$\widehat{\text{cov}}(Z(B_i), Z(B_j)) = L_i^{-1} L_j^{-1} \sum_r \sum_{r'} C(\mathbf{u}_{ir}, \mathbf{u}_{jr'}). \tag{10}$$

(e.g. Gelfand *et al* (2000)). We approximate the point prediction $E(Z(\mathbf{x}_0)|\mathbf{Z})$, where $\mathbf{Z} = (Z(B_1), \ldots, Z(B_N))$ with $\hat{E}(Z(\mathbf{x}_0)|\mathbf{Z})$, which is obtained using (10) as an approximation of (9). It is useful to note that, if we define $\hat{Z}(B_i) = L_i^{-1} \sum_r Z(\mathbf{u}_{ir})$ for $i = 1, \ldots, N$, then the solution $\hat{E}(Z(\mathbf{s}_0)|\mathbf{Z})$ is actually $E(Z(\mathbf{s}_0)|\hat{\mathbf{Z}})$.

In this paper, we define the process Z in terms of a pointwise covariance $C(\mathbf{u}, \mathbf{v})$, but we then use (9), approximated by (10), to derive the covariances of the block averages $Z(B_i)$, $i = 1, \ldots, N$, in terms of the pointwise covariance $C(\mathbf{u}, \mathbf{v})$. This is then used to define a likelihood function for the parameters of the covariance function for the process Z in terms of the observed block averages $Z(B_1), Z(B_2), \ldots, Z(B_N)$.

2.4 Methods for Combining Data with Different Spatial Resolutions

The spatial processes $\hat{Z}$ and $\tilde{Z}$ are measuring/estimating the same quantity of interest but they have different spatial resolution. The process $\hat{Z}$ is observed at n locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$, so we have n point measurements $\hat{\mathbf{Z}} = \left(\hat{Z}(\mathbf{s}_1), \ldots, \hat{Z}(\mathbf{s}_n)\right)^T$. The process $\tilde{Z}$ is averaged over the subgrids $B_1, \ldots, B_m$, so that we observe $\tilde{\mathbf{Z}} = \left(\tilde{Z}(B_1), \ldots, \tilde{Z}(B_m)\right)^T$. We model $\hat{Z}$ and $\tilde{Z}$ in terms of an underlying unobservable process Z, which is the true value of the quantity of interest. We take into account the measurement error and potential bias of $\hat{Z}$ and $\tilde{Z}$ as estimate of Z. Here we consider model (1) for $\hat{Z}$ and model (4) for the process $\tilde{Z}$.

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 write this distribution as a function of the parameters to calculate the MLE for the parameters in models (1) and (4). Since in practice this calculation will be hard, we present a Bayesian approach to estimate the parameters. We have

$$\begin{pmatrix} \hat{\mathbf{Z}} \\ \tilde{\mathbf{Z}} \end{pmatrix} \sim \mathcal{N}\left\{ \begin{pmatrix} \hat{\mu} \\ \tilde{a} + b\tilde{\mu} \end{pmatrix}, \begin{pmatrix} \Sigma_C & \Sigma_{CM} \\ \Sigma_{CM} & \Sigma_M \end{pmatrix} \right\}, \tag{11}$$

where

$$\hat{\mu} = \left(\mu(\mathbf{s}_1), \ldots, \mu(\mathbf{s}_n)\right)^T,$$
$$\tilde{a} = \left(\int_{B_1} a(\mathbf{s}) d\mathbf{s}, \ldots, \int_{B_m} a(\mathbf{s}) d\mathbf{s}\right)^T,$$

and

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

In (11) Σ_C is the covariance of $\hat{\mathbf{Z}}$ (CASTNet) which is the covariance of Z plus measurement error, Σ_M is the covariance of $\tilde{\mathbf{Z}}$ (Models-3), and Σ_{CM} is the cross-covariance between the point measurements $\hat{\mathbf{Z}}$ and the block averages $\tilde{\mathbf{Z}}$. We write Σ to denote the covariance of the multivariate normal distribution of $\left(\hat{\mathbf{Z}}^T, \tilde{\mathbf{Z}}^T\right)^T$, so that Σ is a $(n+m) \times (n+m)$ matrix given by

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

$$\Sigma_{n+j,i} = \Sigma_{i,n+j} = \text{cov}\left(\hat{Z}(\mathbf{s}_i), \tilde{Z}(B_j)\right) = b^2 \int_{B_j} C(\mathbf{s}_i, \mathbf{v}, \boldsymbol{\theta}) d\mathbf{v}/|B_j|, \qquad \text{for } i = 1, \ldots, n \text{ and } j = 1, \ldots, m,$$

$$\Sigma_{n+j_1,n+j_2} = \text{cov}\left(\tilde{Z}(B_{j_1}), \tilde{Z}(B_{j_2})\right) = b^4 \frac{\int_{B_{j_1}} \int_{B_{j_2}} C(\mathbf{u}, \mathbf{v}, \boldsymbol{\theta}) d\mathbf{u} d\mathbf{v}}{|B_{j_1}||B_{j_2}|} + \mathbf{1}_{\{j_1=j_2\}} \sigma_\delta^2 |B_{j_1}|, \qquad \text{for } j_1, j_2 = 1, \ldots, m,$$

where $C(\mathbf{u}, \mathbf{v}, \boldsymbol{\theta})$ is the covariance between $Z(\mathbf{u})$ and $Z(\mathbf{v})$, and $\boldsymbol{\theta}$ is the parameter associated with the covariance of the true underlying process Z, the function $\mathbf{1}_A(\mathbf{x})$ is an indicator for the set A, i.e. takes 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 all the parameters are known. We use the following classical result of multivariate analysis (see, e.g., Mardia, Kent and Bibby (1979), p. 63). If we consider a partitioned multivariate normal vector

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N} \left\{ \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right\}, \tag{12}$$

then the conditional distribution of X_1 given X_2 is normal with mean

$$\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2)$$

and variance

$$\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

Applying this result with $X_1 = Z(\mathbf{x}_0)$, and $X_2 = \mathbf{Z} = (\hat{\mathbf{Z}}^T, \tilde{\mathbf{Z}}^T)^T$, then in our previous notation $\Sigma_{11} = \text{cov}(Z(\mathbf{x}_0), Z(\mathbf{x}_0)) = \sigma_0^2$, $\Sigma_{22} = \Sigma$, and $\Sigma_{12} = \text{cov}(Z(\mathbf{x}_0), \mathbf{Z}) = \tau$, where τ is a $(n+m)$ dimensional vector with components,

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

$$\tau_{n+j} = \text{cov}\left\{Z(\mathbf{x}_0), \mathbf{Z}_{n+j}\right\} = \text{cov}\left\{Z(\mathbf{x}_0), \tilde{Z}(B_j)\right\} = b^2 \int_{B_j} C(\mathbf{x}_0, \mathbf{v}, \boldsymbol{\theta}) d\mathbf{v}/|B_j|, \qquad \text{for } j = 1, \ldots, m,$$

and $\mathbf{Z}_i$ denotes the i^{th} component of $\mathbf{Z}$. We deduce then that the conditional distribution of $Z(\mathbf{x}_0)$ given $\{\hat{\mathbf{Z}}, \tilde{\mathbf{Z}}\}$ is normal with mean

$$\mu(\mathbf{x}_0) + \tau^T \Sigma^{-1}(\mathbf{Z} - \mu),$$

where

$$\mu = (\hat{\mu}, \tilde{a} + b\tilde{\mu})^T$$

and variance

$$\sigma_0^2 - \tau^T \Sigma^{-1} \tau.$$

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 $\mathbf{Z}$,

$$p(Z(\mathbf{x}_0)|\mathbf{Z}) \propto \int p(Z(\mathbf{x}_0)|\mathbf{Z}, \boldsymbol{\phi})\, p(\boldsymbol{\phi}|\mathbf{Z}) d\boldsymbol{\phi}. \tag{13}$$

A Gibbs sampling approach is used to simulate m values from the posterior of the vector parameter $\boldsymbol{\phi}$, where $\boldsymbol{\phi} = (\sigma_e^2, a_0, b, \sigma_\delta^2, \boldsymbol{\beta}, \boldsymbol{\theta})$. Thus, the predictive distribution is approximated by the *Rao-Blackwellized estimator*:

$$p(Z(\mathbf{x}_0)|\mathbf{Z}) = \frac{1}{m} \sum_{i=1}^{m} p(Z(\mathbf{x}_0)|\mathbf{Z}, \boldsymbol{\phi}^{(i)}), \tag{14}$$

where $\boldsymbol{\phi}^{(i)}$ is the i-th draw from the posterior distribution. We propose a Gibbs sampling approach, with three stages. In stage 1 we sample from the conditional posterior of the parameters $(\boldsymbol{\beta}, \boldsymbol{\theta})$, the parameters that represent the lack of stationary of Z. In the second stage we sample from the conditional posterior of the parameters that explain the bias of $\tilde{Z}$ and the measurement error of $\hat{Z}$ and $\tilde{Z}$, namely $(\sigma_e^2, a_0, b, \sigma_\delta^2)$. In stage three, we simulate values of Z (the unobserved true values) at the locations where we have measurements from the process $\hat{Z}$ and $\tilde{Z}$. We cycle through the three stages.

2.5 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. Therefore, the underlying process Z in (2) is nonstationary and the standard methods of spatial modeling and interpolation, kriging, are inadequate. In this section we give a new methodology for spatial modeling of nonstationary covariance. More specifically, we represent the process locally as a stationary isotropic random field with some parameters that describe the local spatial structure (Fuentes 2001, 2002). 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

$$Z(\mathbf{s}) = \int K(\mathbf{s} - \mathbf{u}) X(\mathbf{u}) d\mathbf{u},$$

with $K(\cdot)$ some kernel function and $X(\cdot)$ a constant-variance Gaussian white noise process. Higdon, Swall and Kern (1999) considered extensions of the form

$$Z(\mathbf{s}) = \int K_{\mathbf{s}}(\mathbf{u}) X(\mathbf{u}) d\mathbf{u}, \tag{15}$$

where the kernel $K_{\mathbf{s}}$ depends on position $\mathbf{s}$. The idea of Higdon *et al.* was to model $K_{\mathbf{s}}(\mathbf{u})$ as an unknown function in terms of specific parameters which can then be estimated in a hierarchical Bayes framework. In the case where $K_{\mathbf{s}}$ is a Gaussian kernel for each $\mathbf{s}$, this leads to tractable expressions for the covariance function and hence the likelihood function for the process. Ver Hoef and Barry (1998) used a similar moving average specification of a Gaussian process for cokriging and multivariable spatial prediction.

This approach is quite different from earlier approaches for nonstationarity; see for example Sampson and Guttorp (1992), Guttorp and Sampson (1994), Haas (1995), Nychka *et al.* (1999). It has the disadvantage of not being easily related to traditional spatial models. The development of Higdon *et al.* relies heavily on the Gaussian form of kernel function and it is not clear how restrictive this is. Our own approach also uses kernel representations, but it has a different motivation. We model the nonstationary process Z locally as a stationary isotropic random field with some parameters that describe the local spatial structure.

Consider a Gaussian spatial process $Z(\mathbf{x})$, where $\mathbf{x}$ varies over a domain D contained in a d-dimensional Euclidean space $\mathbb{R}^d$ for some $d > 1$. Typically, $d = 2$. We represent Z as a convolution of local stationary processes (Fuentes and Smith, 2001):

$$Z(\mathbf{x}) = \int_D K(\mathbf{x} - \mathbf{s}) Z_{\boldsymbol{\theta}(\mathbf{s})}(\mathbf{x}) d\mathbf{s}, \tag{16}$$

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 (16) is defined as a limit (in mean square) of approximating sums (e.g., Cressie, 1993, p. 107, Yaglom, 1962, p. 23). 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,

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

where μ could be a polynomial function of $\mathbf{x}$ with coefficients $\boldsymbol{\beta}$.

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

$$\operatorname{cov}\{Z_{\boldsymbol{\theta}(\mathbf{s})}(\mathbf{s_1}), Z_{\boldsymbol{\theta}(\mathbf{s})}(\mathbf{s_2})\} = C_{\boldsymbol{\theta}(\mathbf{s})}(\mathbf{s_1} - \mathbf{s_2}).$$

The process $Z_{\boldsymbol{\theta}(\mathbf{s})}$ could have a Matérn stationary covariance (Matérn, 1960):

$$C_{\boldsymbol{\theta}(\mathbf{s})}(\mathbf{x}) = \frac{\sigma_s}{2^{\nu_s-1}\Gamma(\nu_s)}(2\nu_s^{1/2}|\mathbf{x}|/\rho_s)^{\nu_s}\mathcal{K}_{\nu_s}(2\nu_s^{1/2}|\mathbf{x}|/\rho_s), \tag{17}$$

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. $\sigma_s = \operatorname{var}(Z_{\boldsymbol{\theta}(\mathbf{s})}(\mathbf{x}))$, where the covariance parameter σ_s is 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; e.g. when $\nu_s = \frac{1}{2}$, we get the exponential covariance function. In the limit as $\nu_s \to \infty$ we get the Gaussian covariance

$$C_{\boldsymbol{\theta}(\mathbf{s})}(\mathbf{x}) = \sigma_s e^{-|\mathbf{x}|^2/\rho_s^2}.$$

The covariance $C(\mathbf{s_1}, \mathbf{s_2}; \theta)$ of Z is a convolution of the local covariances $C_{\boldsymbol{\theta}(\mathbf{s})}(\mathbf{s_1} - \mathbf{s_2})$,

$$C(\mathbf{s_1}, \mathbf{s_2}; \boldsymbol{\theta}) = \int_D K(\mathbf{s_1} - \mathbf{s})K(\mathbf{s_2} - \mathbf{s})C_{\boldsymbol{\theta}(\mathbf{s})}(\mathbf{s_1} - \mathbf{s_2})d\mathbf{s}. \tag{18}$$

The processes $Z_{\boldsymbol{\theta}(\mathbf{s}_i)}$ are stationary with autocovariances $C_{\boldsymbol{\theta}(\mathbf{s}_i)}$, then they can be represented in the form:

$$Z_{\boldsymbol{\theta}(\mathbf{s}_i)}(\mathbf{x}) = \int_{\mathbb{R}^2} K_{\mathbf{s}_i,\mathbf{x}}(\mathbf{u} - \mathbf{x})X_{\mathbf{s}_i}(\mathbf{u})d\mathbf{u} \tag{19}$$

where $K_{\mathbf{s}_i,\mathbf{x}}$ is a kernel for each $\mathbf{s}_i$ and $\mathbf{x}$, and $X_{\mathbf{s}_i}$ is for each $\mathbf{s}$ a white noise processes. The white noise processes $X_{\mathbf{s}_i}$ are independent, for different values of $\mathbf{s}_i$. A direct attempt to rewrite (16) in the form (15) does not work because the process $X_{\mathbf{s}_i}$ is an independent white noise process for each $\mathbf{s}_i$, and we conjecture that there is no way of reducing the present model to one of the structure of (15), with a common underlying white noise process X. There is a possible alternative interpretation of (16), in which the process $X_{\mathbf{s}_i}$, instead of being an independent white noise process for each $\mathbf{s}_i$, is a common process for all $\mathbf{s}$, so the processes $Z_{\boldsymbol{\theta}(\mathbf{s}_i)}$ are correlated. In that case it is straightforward to rewrite (16) in the form of (15), though the kernel $K_{\mathbf{s}_i}(\mathbf{u})$ would be difficult to evaluate in practice. Because of its additional computational complexity, we have not pursued that approach in the present paper.

In (18) every entry requires an integration. Since each such integration is actually an expectation with respect to a uniform distribution, we propose Monte Carlo integration. We propose to draw a systematic sample of locations $\mathbf{s}_m$, $m = 1, 2, ..., M$ over D. Hence, we replace $C(\mathbf{s_1}, \mathbf{s_2}; \boldsymbol{\theta})$ with

$$C_M(\mathbf{s_1}, \mathbf{s_2}; \boldsymbol{\theta}) = M^{-1} \sum_{m=1}^{M} K(\mathbf{s}_1 - \mathbf{s}_m) K(\mathbf{s}_2 - \mathbf{s}_m) C_{\boldsymbol{\theta}(\mathbf{s}_m)}(\mathbf{s_1} - \mathbf{s_2}). \tag{20}$$

This is a Monte Carlo integration which can be made arbitrarily accurate and has nothing to do with the data Z. The sampling points $\mathbf{s}_m$, $m = 1, 2, ..., M$, determine subregions of local stationarity for the process Z. We increase the value of M until convergence is achieved.

3 Estimation

In this section we explain how to efficiently implement our algorithm for spatial prediction combining observations with the output from numerical models.

3.1 Algorithm

1. Posterior predictive values for Z

   For spatial prediction the quantity of interest is the predictive distribution for $Z(\mathbf{x}_0)$ given the observed values $\mathbf{Z}$. We use expression (14) to approximate the predictive distribution with a Rao-Blackwellized estimator, conditioning on the posterior simulated values for all the parameters, using for this simulation the following Gibbs sampling algorithm.

2. Algorithm for Gibbs sampling

   We now discuss how to sample from the posterior distribution of the parameters. In our Gibbs sampling approach there are three stages. We alternate between the parameters that measure the lack of stationarity, $(\boldsymbol{\beta}, \boldsymbol{\theta})$ (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).

   Gibbs sampling: Stage 1.

We obtain the conditional posterior for the parameters that measure the lack of stationarity, $(\boldsymbol{\beta}, \boldsymbol{\theta}(\mathbf{s}))$, conditioning on the values of Z that are updated in Stage 3. The posterior of $(\boldsymbol{\beta}, \boldsymbol{\theta}(\mathbf{s}))$ will be completely specified once we define the priors for $(\boldsymbol{\beta}, \boldsymbol{\theta}(\mathbf{s}))$, because we have that $[Z|\boldsymbol{\beta}, \boldsymbol{\theta}]$ is Gaussian, where the brackets $[]$ are used here to denote densities. We consider a Metropolis-Hasting step for blocks of parameters, after performing partial marginalisations of the full conditionals. We treat as a block $\boldsymbol{\theta}$, which are the three covariance parameters, the sill, the range and the smoothness. In our experience this scheme produce a chain with far better mixing properties than the one that is obtained by independent sampling of the full conditionals of the sill, range and smoothness parameters. We use a gamma prior for all the covariance parameters, except for the sill parameter that we use a uniform prior for the logarithm transformation of the sill. For the $\boldsymbol{\theta}$ parameter we use uniform priors.

Gibbs sampling: Stage 2.

We obtain the conditional posterior for the parameters a_0, b, σ_δ^2 and σ_e^2 that measure the bias and uncertainty of Models-3, and the measurement error of CASTNet.

The posterior of σ_e^2 given the n values of $\hat{Z}$ and Z at the CASTNet sites (updated in Stage 3), can be easily obtained, because we have the following regression problem:

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

where σ_e^2 is the variance of the error term $e(\mathbf{s})$, and $Z(\mathbf{s})$ is independent of $e(\mathbf{s})$. We have that $[\hat{Z}(\mathbf{s})|Z(\mathbf{s}), \sigma_e^2]$ is normal with mean $Z(\mathbf{s})$ and variance σ_e^2. Then, the posterior of σ_e^2 is proportional to

$$[\hat{Z}(\mathbf{x}_1), \ldots, \hat{Z}(\mathbf{x}_n)|Z(\mathbf{x}_1), \ldots, Z(\mathbf{x}_n), \sigma_e^2][\sigma_e^2]$$

where $[\sigma_e^2]$ denotes the prior distribution for σ_e^2, and $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are the n CASTNet sites.

The posterior distributions of a_0, b, σ_δ^2 given the values of $\tilde{Z}$ and Z (updated in Stage 3) at the m blocks, can be easily calculated, because we have the following regression problem:

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

where σ_δ^2 is the variance of the error term $\delta(\mathbf{s})$, and $Z(\mathbf{s})$ is independent of $\delta(\mathbf{s})$. It follows that

$$[\tilde{Z}(B_1), \ldots, \tilde{Z}(B_m)|Z(B_1), \ldots, Z(B_m), a_0, b, \sigma_\delta^2]$$

is normal with mean $\mathbf{a} + b\{Z(B_1), \ldots, Z(B_m)\}$, where $\mathbf{a} = \left\{\int_{B_1} a(\mathbf{x})d\mathbf{x}, \ldots, \int_{B_m} a(\mathbf{x})d\mathbf{x}\right\}$, and a diagonal covariance matrix with diagonal elements $\sigma_\delta^2|B_i|$. Thus, the posterior of a_0, b, σ_δ^2 is proportional to

$$[\tilde{Z}(B_1), \ldots, \tilde{Z}(B_m)|Z(B_1), \ldots, Z(B_m), a_0, b, \sigma_\delta^2][a_0, b, \sigma_\delta^2].$$

<u>Gibbs sampling: Stage 3.</u>

We simulate values of Z (the unobserved true values) at the n locations where we have measurements for $\hat{Z}$, and also at the m blocks where we observe $\tilde{Z}$, conditioning on the values of $\boldsymbol{\beta}, \boldsymbol{\theta}$ (updated in Stage 1) and $\mathbf{Z}$. The simulated values at the m blocks are obtained by simulating values of Z at a sample of locations within each block. Then $Z(B_i)$ is approximated by $L^{-1}\sum_{k=1}^{L} Z(\mathbf{s}_{i_k})$, where $\mathbf{s}_{i_1}, \ldots, \mathbf{s}_{i_L}$ is a centered systematic sample in B_i.

For model validation, we simulate values from the posterior distribution of CASTNet given Models-3,

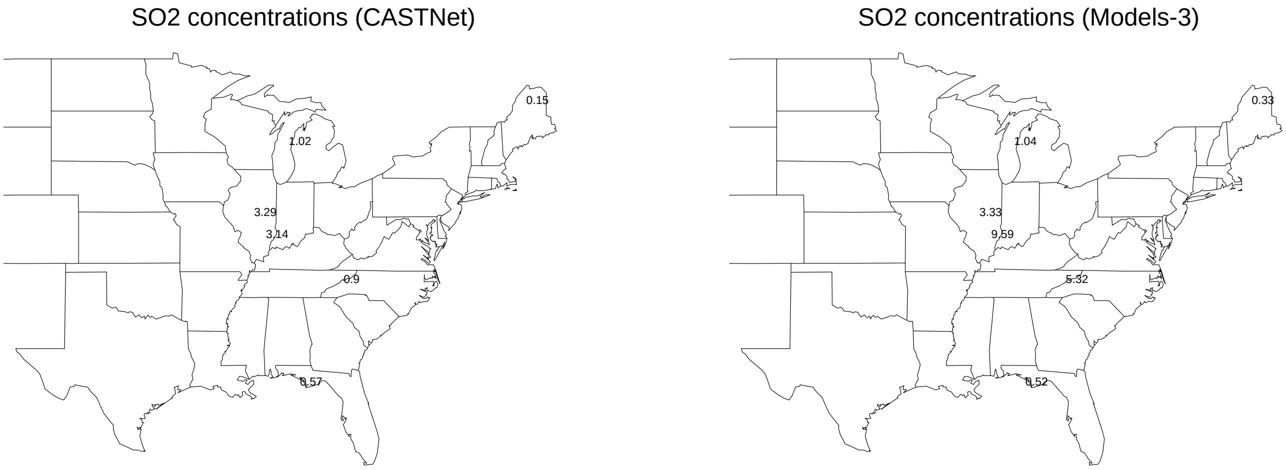
$$P(\hat{Z}|\tilde{Z}, a = 0, b = 1),$$

and we compare the actual observations with this simulated posterior predictive distribution.

4 Application: Air Pollution Data

The regional scale air quality models (Models-3) run by the U.S. EPA estimate hourly concentrations and fluxes of different air pollutants. The primary objective of Models-3 is to improve the environmental management community's ability to evaluate the impact of air quality management practices for multiple pollutants at multiple scales, as part of the process of regulating air pollution. The spatial domain, D, is a regular grid (81×87) and the dimensions of each pixel in the grid are 36km × 36km. Models-3 provides hourly concentrations for each pixel. As an example we examine sulfur dioxide. Figure 1 (b) shows the weekly averaged concentrations of SO_2 for the week starting July 11, 1995. Our first objective is the validation of Models-3.

The Clean Air Status and Trends Network (CASTNet) measures SO_2 weekly averaged concentrations and fluxes at 50 sites (see Figure 1 (a)). We propose to combine both sources of information, namely the monitoring data and the output of the air quality models, to get more reliable maps of air pollutants concentrations and fluxes.



(a) (b)

Figure 3: (a): Weekly average of SO_2 concentrations (ppb) at some selected CASTNet sites, for the week of July 11, 1995. (b): Weekly average of SO_2 concentrations (ppb) from Models-3 interpolated at 6 selected locations, for the week of July 11, 1995, for Models-3 validation.

In Figure 3(a) we show the CASTNet values at 6 selected sites that are representative of different metereological, land use, and altitude conditions. For validation of Models-3, in Figure 3(b) we show the modes of the posterior predictive distribution (eq. (6)) of CASTNet given Models-3 at the 6 selected sites. There is clear evidence that Models-3 is overestimating the concentrations of SO_2. We modeled the covariance for Models-3 using equation (18), taking into account the lack of stationarity and the change-of-support problem. We calculated the covariances involving block averages by drawing a set of L locations in each pixel: this is a Monte Carlo 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 convergence.

Now, we use the methodology presented in this paper to estimate the bias in Models-3 for bias removal, using expression (7). We modeled Models-3 in terms of an underlying unobservable process Z with the true values of SO_2, but we added an additive constant bias, a multiplicative constant bias, and a measurement error term. We also modeled CASTNet in terms of the "true" process Z and we added a measurement error term (see Section 2.2).

Figure 4 (a) show the posterior distributions of some covariance parameters for the underlying process Z at the selected sites shown in Figure 3 (a). We used gamma priors for all the Matérn covariance parameters, except for the sill parameter for which we used

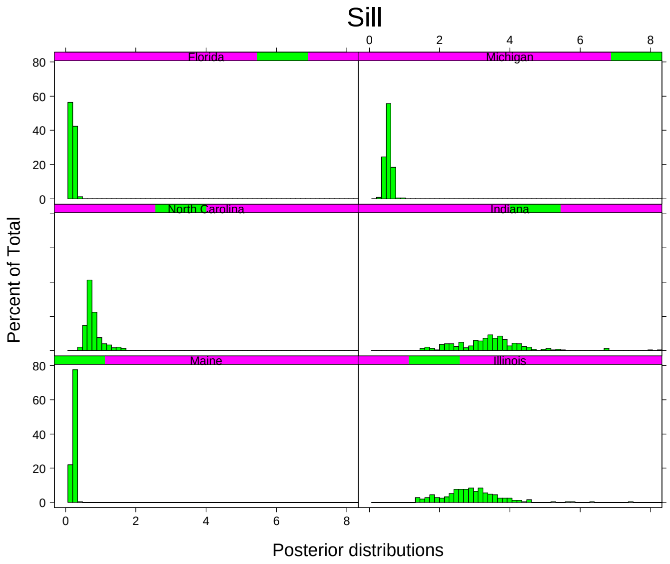
$$p(\sigma) \propto \sigma^{-1},$$

which is a uniform prior for $\log(\sigma)$. The mean and variance of the gamma priors were determined using previous knowledge and results from the analysis of similar data. For the smoothness parameter the prior mean was .5 and the variance 1, and for the range parameter the prior mean was 100 and the variance 2025.

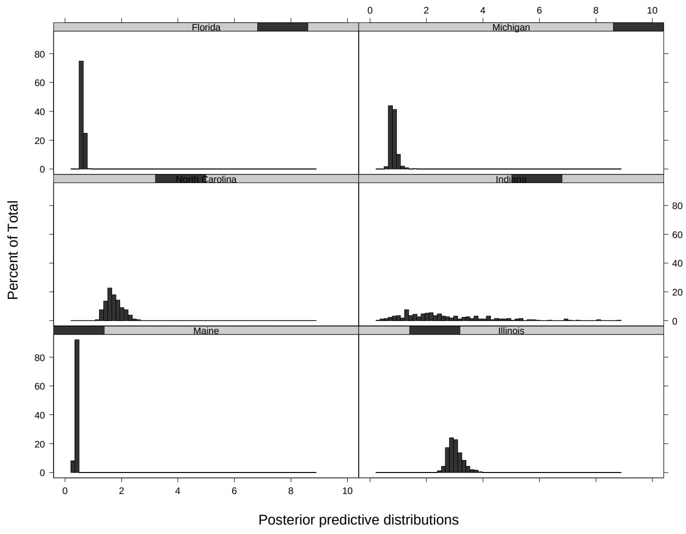
In Figure 4 (a) we plot the posterior distribution of $\sigma(\mathbf{x})$ at some locations $\mathbf{x}$, where we have CASTNet sites. These spatial locations at which we examine the covariance parameters in Figure 4 have nothing to do with the election of the coordinates of the nodes $\mathbf{s}_1, \ldots, \mathbf{s}_M$ in equation (20). We are assuming that the covariance parameters are approximately constant in between nodes, otherwise we increase M (the number of nodes). Thus, the value of $\boldsymbol{\theta}$ at some location $\mathbf{x}$ of interest is calculated as the value of $\boldsymbol{\theta}$ at the node $\mathbf{s}_i$ that is the closest to $\mathbf{x}$. An alternative approach is to model $\boldsymbol{\theta}$ using splines, then we would have a continuous function of space, but when M is large enough we get similar results using splines as with the simpler approach we are presenting here. The sill parameter changes with location as illustrated by the variation in the distributions in Figure 4 (a). Thus, this indicates a lack of stationarity. The mean and standard deviation of the posterior distributions for the sill parameter at six selected sites located in Maine, Illinois, North 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), 0.54 (0.08). The plots were obtained using 100 posterior simulated values. We used every tenth of the last 1000 iterations from the corresponding chains. The range parameter does not change much with location. The mean and standard deviation of the posterior distributions for the range parameter at the same six sites are 97.73 (12.77), 95.74 (28.47), 119.03 (35.77), 91.89 (27.30), 114.92 (21.89), 96.18 (15.50).

The smoothing parameter does not change with location either, and is always close to 1/2 (exponential).

We implemented the nonstationary model (16) with weight function $K(\mathbf{u}-\mathbf{s}) = \frac{1}{h^2} K_0\left(\frac{\mathbf{u}-\mathbf{s}}{h}\right)$, where



(a)



(b)

Figure 4: (a): Posterior distributions for the sill parameter of the Matérn covariance for the SO_2 concentrations of Z, for the week starting July 11, 1995, at the 6 selected locations showed in Figure 3 (a). (b): Predictive posterior distributions for the SO_2 concentrations of Z at the 6 selected locations showed in Figure 3, for the week starting July 11, 1995. The predictive posterior values are for the spatial underlying process Z given CASTNet and Models-3.

$K_0(\mathbf{u})$ is the quadratic weight function

$$K_0(\mathbf{u}) = \frac{3}{4}(1 - {u_1}^2)_+ \frac{3}{4}(1 - {u_2}^2)_+, \tag{21}$$

for $\mathbf{u} = (u_1, u_2)$. The bandwidth parameter h is defined as $l/2 + l/2\epsilon$, where l is the distance between the sample points $\mathbf{s_1}, \ldots, \mathbf{s_M}$ in (20), and ϵ is a value between 0 and 1. For ϵ we used a uniform prior in the interval $[0, 1]$. The parameter ϵ determines the amount of overlapping between the subregions of stationarity centered at the sampling points $\mathbf{s_1}, \ldots, \mathbf{s_M}$, and h can be interpreted as the diameter of the subregions of stationarity.

The mode of the posterior distribution for the parameter that measures the measurement error for CASTNet is .8, and for Models-3 it is .1. The mode of the posterior distribution for the parameter that measures the multiplicative bias for Models-3 is .5 with a standard error of .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) for model selection. In this application $q = 4$ seemed to be large enough.

We use a Bayesian approach for spatial prediction. We sample from the predictive distribution (eq. (5)) of the true underlying process Z given the observations and the model output, taking into account the lack of stationarity and the change-of-support problem (we calculated the block averages covariances drawing a set of 4 locations in each pixel). We use posterior predictive checks (PPC) as suggested by Rubin (1984) and Raftery (1988) for validation of the proposed model for Z. Thus, we compare the posterior predictive distributions of the "true" process Z at different locations to the observed data, and we judge whether the generated data are similar to the CASTNet data (Figure 3(a)). Figure 4 (b) shows simulated values from the posterior predictive distribution (eq. (5)) of SO_2 weekly average concentrations at the CASTNet selected sites.

Figure 5 is a comparison of the generated data through the Bayesian melding approach proposed in this paper to CASTNet, taking into account the uncertainties about CASTNet and Models-3. The graph on the right in Figure 5 shows the CASTNet measurements for the week starting July 11, 1995, versus the modes and 90% credible intervals of the predictive Bayesian distributions (eq. (5)) derived from the Bayesian melding approach at the CASTNet locations. The dotted lines indicate a 90% credible region for the CASTNet values. All the modes fall within the credible region for CASTNet, so that the generated data are similar to the observed data. On the other hand, the graph on the left in Figure 5 shows CASTNet measurements versus the modes and 90% credible intervals of the predictive Bayesian distributions (eq. (6)) derived only from Models-3 (without combining Models-3 with CASTNet) at the CASTNet locations for validation of Models-3. The modes in the latter plot do not fall within the credible bands for CASTNet. This figure shows the improvement obtained in the prediction of SO_2 by combining CASTNet and Models-3 through the Bayesian melding approach presented in this paper.

Site	Mode	S.E.	90%	C. I.	CASTNet	Models-3	90%	C. I.
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	0.29	0.98	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

TABLE 1. Columns 2-5 in this table show the modes, standard errors and 90% credible intervals of the posterior predictive distribution (eq. (5)) for the underlying process Z measuring the true SO_2 concentrations at the 6 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 (eq. (6)) for model validation.

In Table 1 we have the modes, sample standard deviations, and 90% credible intervals of the posterior predictive distribution (5) for SO_2 at the 6 selected sites. As expected, 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 on 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 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 if the generated data are similar to the CASTNet data. The CASTNet values in table 1 at the 6 sites represent the 5th, 91st, 1st, 68th, 14th, and 79th percentiles of the corresponding posterior predictive distributions from the Bayesian melding approach, for NC the CASTNet values are low relative to the posterior ppd distribution. This is due to the higher altitude (and lower pollution levels) of this particular site in relation the nearby locations that have much higher pollution levels because of the presence of power plants at lower altitudes.

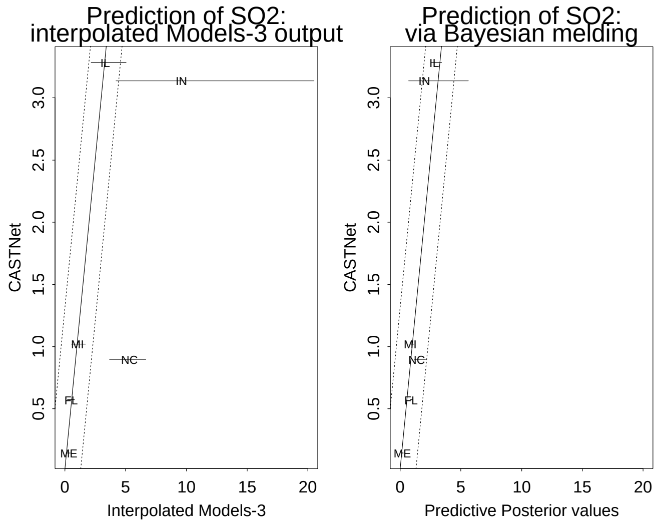


Figure 5: The graph on the left shows CASTNet measurements for the week starting July 11, 1995, versus the modes and 90% credible intervals of the predictive Bayesian distribution ($p(\hat{Z}|\tilde{Z}, a = 0, b = 1)$) given Models-3 at the CASTNet locations for Models-3 validation. The graph on the right shows the CASTNet measurements versus the modes and 90% credible intervals of the predictive Bayesian distributions ($p(Z|\hat{Z}, \tilde{Z})$) derived from a Bayesian melding approach to combine observations and Models-3 output. The dotted lines indicate a 90% credible region for the CASTNet values.

The last 3 columns in Table 1 are the modes and the corresponding 90% credible intervals of the posterior predictive distribution (6) for model validation, which are also plotted in Figure 3 (b). 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 (s.e. 3.00 ppb) while the CASTNet value is only 0.90 ppb (see Figure 5). We could 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 a_0) and the multiplicative bias ($b \sim N(.5, .5)$). We calculated at each site the posterior means of $a(\mathbf{x})$ and b, using the posterior distribution (7), and we obtained the following adjusted Models-3 values (adjusted value = $((\text{Models 3}) - \tilde{a})/\tilde{b}$) at the 6 selected sites: 0.12, 2.88, 1.09, 3.12, 0.44, and 1.01, where $\tilde{a}$ and $\tilde{b}$ are the posterior means. These adjusted values are very similar to CASTNet. Therefore, the methodology presented here for combining spatial data is not only useful for improving the prediction of air quality but it is also a tool for validating air quality numerical models and for bias removal.

Figure 7 (a) shows the predicted values of SO_2 at different locations in a regular grid, using a Bayesian melding approach for prediction to combine CASTNet and Models-3 data. The predicted values in Figure 7 (a) are the means of the posterior predictive distribution (eq. (5)) for the SO_2. This graph looks similar to the output of Models-3 shown in Figure 1 (b). However, the SO_2 values in Figure 7 (a) are between 0ppb and 8ppb, while in Figure 1 (b) the SO_2 values were between 0ppb and 40ppb. The range of values in Figure 7 (a) is more reasonable, and it is closer to the range of values for CASTNet shown in Figure 1 (a). This illustrates the effectiveness of the Bayesian melding approach for correcting the bias in Models-3. In Figure 6 we have 6 ensembles of SO_2 concentrations, where each ensemble is a simulation from the posterior predictive distribution (eq. (5)) of the SO_2. The variability from ensemble to ensemble is due to the uncertainty in the prediction. Figure 7 (b) 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.

5 Discussion

In this paper we have introduced a new statistical methodology for validation and adjustment of numerical models, and for spatial prediction combining data with the output from numerical models. In the application presented in this paper, we assume there is some smooth underlying (but unobserved) spatial field that measures the "true" concentration/flux of the pollutant at each location. The data collected at the monitoring sites are considered the "true values" plus some measurement error. The output of the air quality numerical models can be also written in terms of the true underlying, but unobservable, process, with some parameters that explain the bias and microscale noise in the numerical models. The *truth* is assumed to be a smooth underlying spatial process with some parameters that explain the large scale and short scale dependency structure of the air pollutants. This spatial field is represented locally as a stationary isotropic random field, but the parameters of the stationary random field are allowed to vary across space. Kernel functions are used to ensure that the field is well-defined but also continuous. In this paper we also take into account the change of support problem that occurs when combining data with different spatial resolution.

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

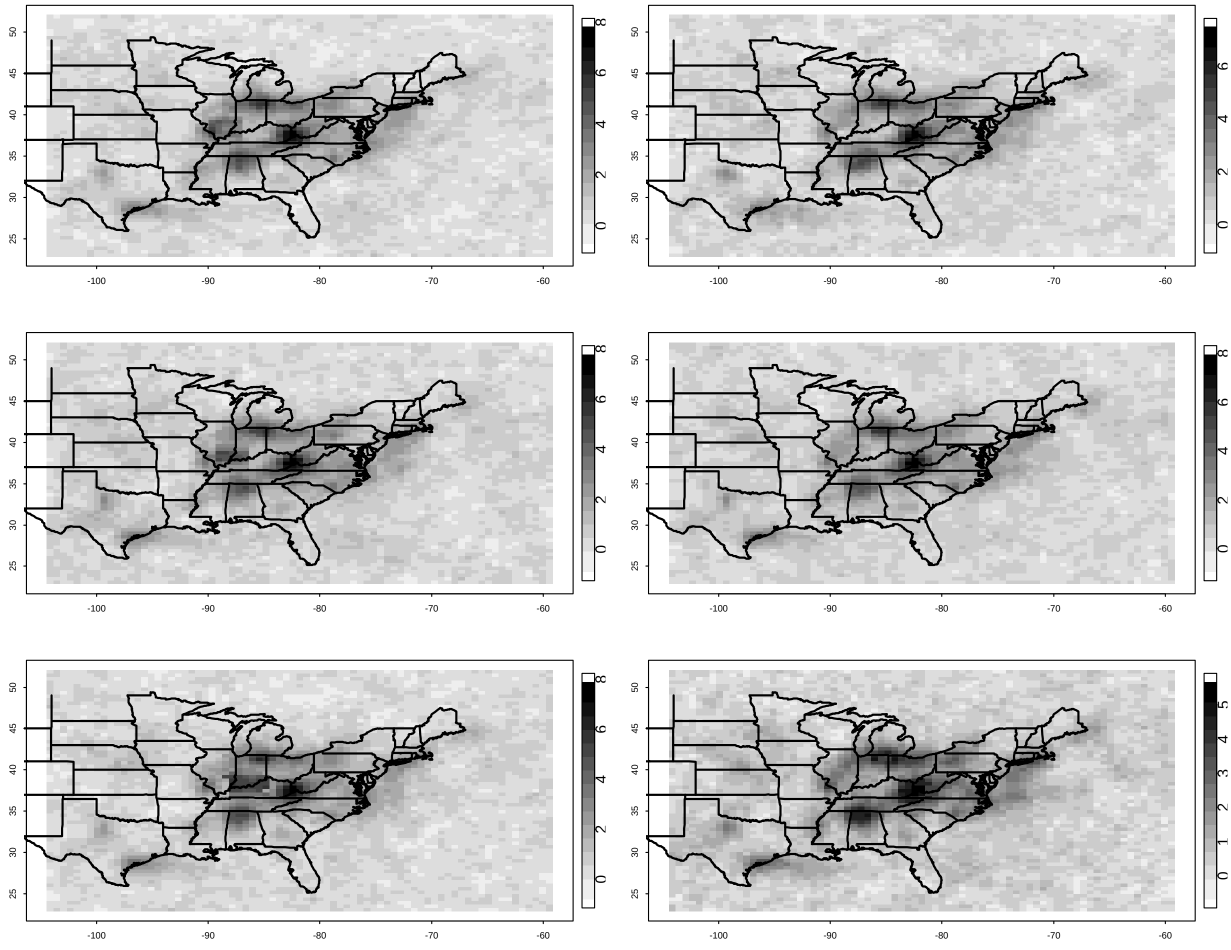
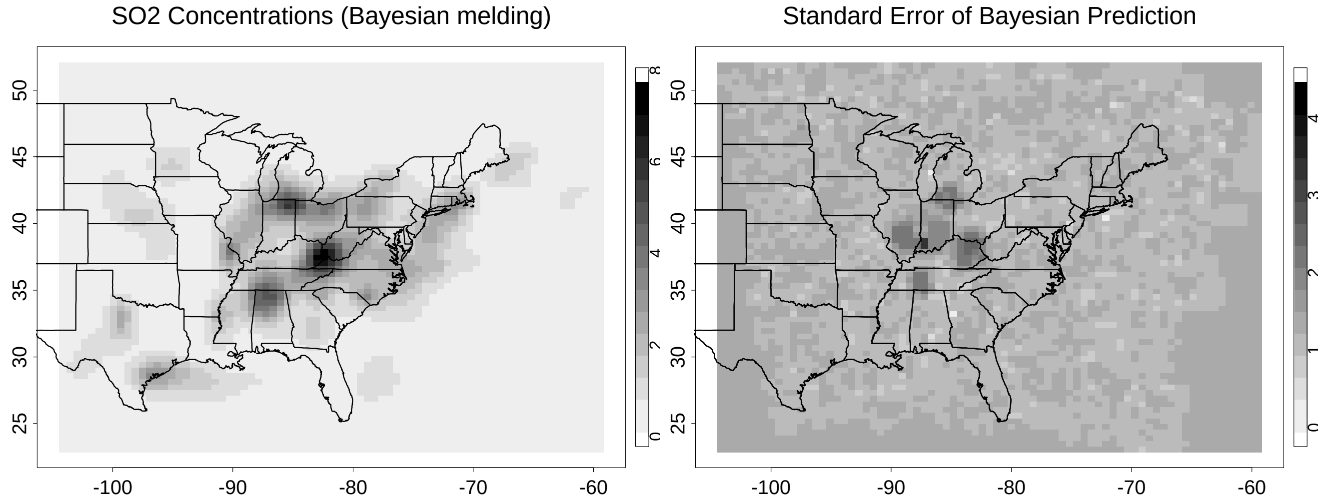


Figure 6: Simulated SO_2 concentrations from the predictive distribution for the SO_2, using a Bayesian melding approach for prediction to combine CASTNet and Models-3 data. The simulated values are from the posterior predictive distribution for the underlying process Z given CASTNet and Models-3

that they should accurately reproduce observed measurements when applied to ground data. Our objectives are model validation and bias removal 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 validate 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.

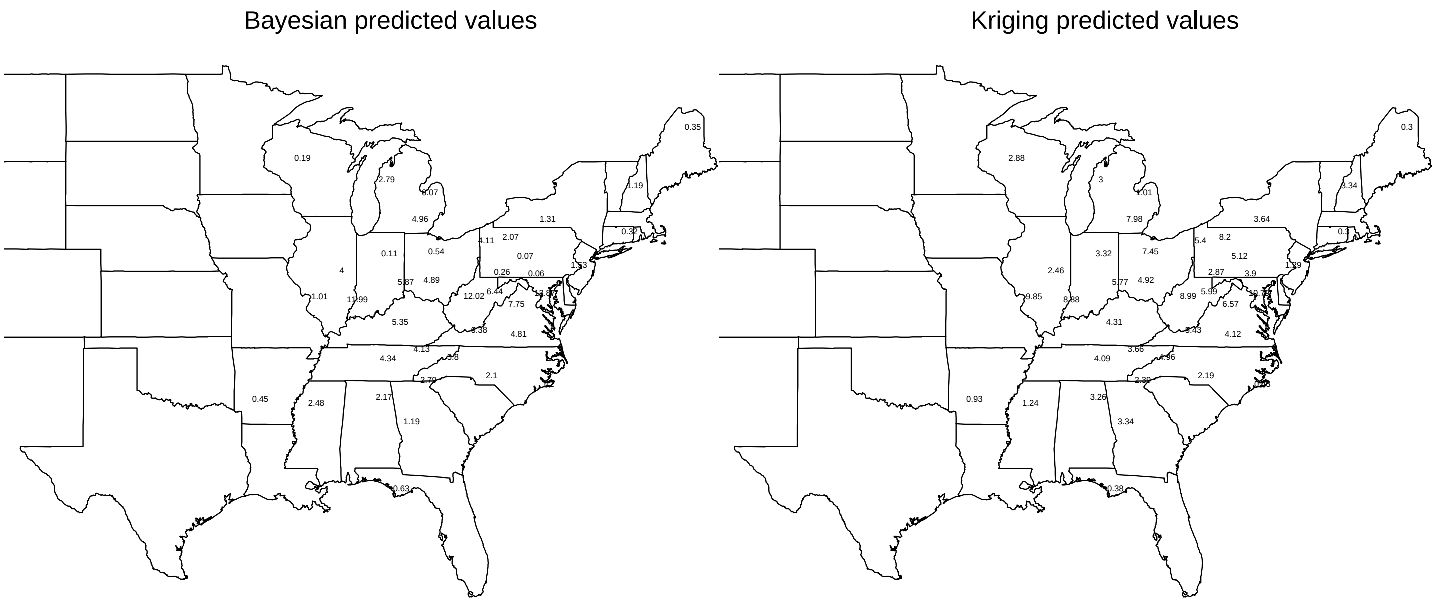


(a) (b)

Figure 7: (a): Predicted SO_2 concentrations via a Bayesian melding approach to combine CASTNet and Models-3 data. This graph shows the mean of the posterior predictive distribution for the underlying process Z given CASTNet and Models-3. (b): Standard error of the posterior predictive distribution for the SO_2 concentrations of Z, using a Bayesian melding approach for prediction to combine CASTNet and Models-3 data.

The Bayesian approach is ideal for this application because it provides a natural framework to combine data from very different sources taking into account different uncertainties, and it also provides posterior distributions of quantities of interest that can be used for scientific inference. However, our approach could be also 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 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 are used to estimate the bias and covariance parameters, using a likelihood approach. And we iterate between steps 1 and 2.

Another approach to model validation is to use spatio-temporal models for monitoring data to provide estimates of average concentrations over grid cells corresponding to model prediction (Dennis et al. 1990, Sampson and Guttorp 1998, Davis et al. 2000, Chu 1995, Chu 1996, Chu and Cox 1995). This approach is reasonable when the monitoring data are dense enough that we can fit an appropriate spatio-temporal model to the data. In situations like the one presented here, with few and sparse data points that show a lack of

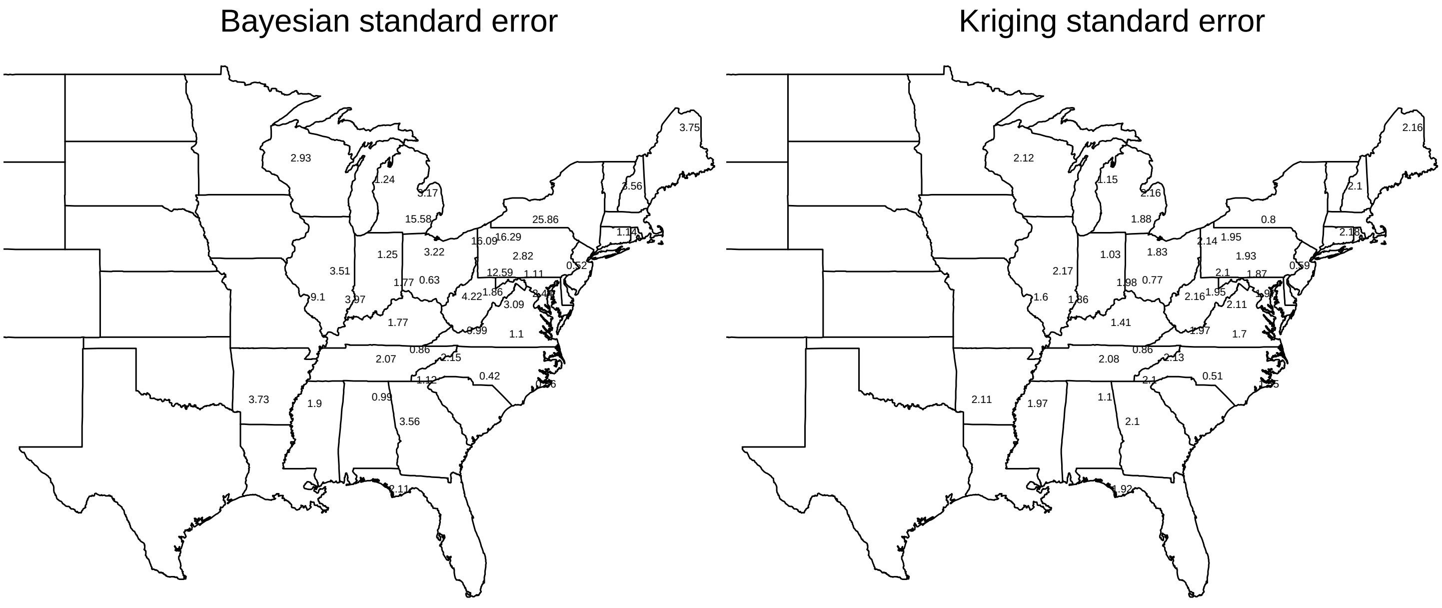


(a) (b)

Figure 8: (a): Mean of the predictive posterior distributionof Models-3 at each CASTNet site. (b): Kriging predictor of Models-3 at each CASTNet location.

stationary, 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 validation would be questionable.

The spatial dependency structure shown by the air pollutant fluxes and concentrations changes with location. Therefore, we need to model the underlying process Z, with the true values of fluxes/concentrations of air pollution, as a spatial process with a nonstationary covariance. Misspecification of the covariance will lead to wrong predictions and unreliable estimates of the standard errors in the prediction. 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. In recent years, probably the most extensively studied method for nonstationary spatial processes is the deformation approach due to Sampson and Guttorp (1992); see also Guttorp and Sampson (1994), and Guttorp, Meiring and Sampson (1994). Maximum likelihood versions of the method were developed by Mardia and Goodall (1993) and Smith (1996). In a series of papers best represented by Haas (1995), T. Haas has proposed an approach to nonstationary spatial kriging based on moving windows. Higdon, Swall and Kern (1999) give a



(a) (b)

Figure 9: (a): Standard error of the predictive posterior distribution at each CASTNet site. (b): Standard error of the Kriging predictor at each CASTNet location.

model for accounting for heterogeneity in the spatial covariance function of a spatial process, using a moving average specification of a Gaussian process. Another approach has been developed by Nychka and Saltzman (1998) and Holland *et al.* (1999), that extends the "empirical orthogonal functions" (EOF) approach that is popular among atmospheric scientists. Here we use a new model for nonstationary processes proposed by Fuentes (2001, 2002), and further developed by Fuentes and Smith (2001). In this model the process is represented locally as a stationary isotropic random field, but the parameters of the stationary random field are allowed 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.

The methodology presented in this paper has overcame the inherent limitations of earlier approaches to evaluation of numerical models and mapping of air quality. However, there are also some assumptions and limitations in our approach. In our analysis we have modeled the true underlying pollution process as a Gaussian field, which seemed to be a reasonable assumption since we were working with weekly averages of hourly values. However, for hourly pollution levels the log transformation could be more appropriate. Based on the experience of our EPA collaborators, the main problem with air quality models seem to be be the presence of additive bias. This is the reason and motivation 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 also assumed independence between the truth and the measurement error. This assumption, to our knowledge, is valid and 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.

In this paper 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), and Davis et al. (2000). In cokriging or KED we write the data in a regression model as a function of some covariates. Therefore, if the observed data and the output of numerical models exhibit different spatial resolution, we could not write directly the observed data as a function of the output of numerical models or vice versa. In this paper we overcome that problem by writing each source of information in terms of the truth, without assuming we observe it.

For the evaluation of the models, an alternative approach would be to use kriging to predict the value of the numerical models at the locations where we have CASTNet. We compared our proposed modeling approach with this traditional stationary kriging prediction method, that ignores the uncertainty in the covariance parameters (Figures 8 and 9). Figure 8 shows the kriging predicted values and also the mean of the ppd using our Bayesian approach. Both methods tend to give similar predicted values, the main difference is in the prediction error (Figure 9), which is the critical factor for any inference done with the data. Figure 9 shows that our approach gave larger standard errors than kriging for the Models-3 predicted values, specially in areas close to power plants. This reflects the difficulty in estimating the covariance parameters in these areas. The kriging approach reported uniform 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 for the prediction we underestimated considerably the length of the confidence intervals leading to wrong conclusions about the performance of Models-3.

Some of the main sources of uncertainty that affect the performance of the models are the following: the photochemistry model parameterizations (which is the treatment of photochemistry phenomena varying at scales smaller than the grid size of the model), the boundary conditions, the treatment of the land use/land cover at smaller scales than the grid size, the quality of the emissions input that goes into the air quality models, and the air dispersion modeling of pollution plumes (at smaller scales than the grid size). The present analysis showed that the air quality models perform worse in areas closer to power plants, suggesting that the dispersion modeling of pollution plumes in that area needs to be improved. The Models-3 output used for the analysis in this paper assumes that the SO_2 diffuses uniformly within each grid cell. The Models-3 group, encouraged by the findings here, is currently considering new dispersion models for the air quality models. For more information about dispersion modeling of pollution plumes, see e.g. Beychok (1995).

The approach presented in this paper gave us a good understanding of the spatial structure of the "true" concentrations of SO_2. This information can be useful for designing future data collection. Part of our future work is to use the findings in this paper for monitoring network design.

In this paper we have focused on air quality applications, but the methods presented here could be extended to other problems where we need to combine data from various sources and with different spatial/temporal resolutions. For instance, combining spatial data from models and observations is a frequent problem in climate and weather prediction.

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.

Beychok, M. R. (1995). *Fundamentals of Stack Gas Dispersion*, 3rd Edition. Published by author, Irvine, California.

Chu, S.-H. (1995). Partition of phase and magnitude contributions in residual variance analysis. *Journal of Geophysical Research,* **100**, 18,639-18,649.

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.

Chu, S.-H., Cox, W. M. (1995). Effects of emissions reductions on ozone predictions by the Regional Oxidant Model during the July 1988 episode. *Journal of Applied Meteorology,* **34**, 679-693.

Clarke, J. F., Edgerton, E. S., (1997). Dry deposition calculations for the clean air status and trends network. *Atmospheric Environment,* **31**, 3667-3678.

Cressie, N. A. (1993). *Statistics for spatial data.* Revised Edition. Wiley, New York.

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.

Dennis, R. L., Barchet, W. R., Clark, T. L., 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, Vol 1, National Acid Precipitation Assessment Program, Washington, DC.

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.

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.

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. 94^{th} Annual

A&WMA Conference and Exhibition, Orlando, FL.

Fuentes, M. (2001). A new high frequency kriging approach for nonstationary environmental processes. *Envirometrics,* **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. Tech. report at North Carolina State University, Institute of Statistics Mimeo Series #2534.

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

Guttorp, P., Meiring, W. and Sampson, P. (1994), A space-time analysis of ground-level ozone data. *Environmetrics* **5**, 241–254.

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

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.

Higdon, D., Swall, J. and Kern, J. (1999), Non-stationary spatial modeling. In *Bayesian Statistics 6*, eds. J.M. Bernardo *et al.*, Oxford University Press, pp. 761–768.

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*, eds. Gómez-Hernández, J., Soares, A. and Froidevaux, R., Kluwer, Dordrecht, 65–76.

Matérn, B. (1960). *Spatial Variation.* Meddelanden från Statens Skogsforskningsinstitut, **49**, No. 5. Almaenna Foerlaget, Stockholm. Second edition 91986), Springer-Verlag, Berlin.

Mardia, K.V. and Goodall, C.R. (1993), Spatial-temporal analysis of multivariate environmental monitoring data. In *Multivariate Environmental Statistics*, eds. G.P. Patil and C.R. Rao, Elsevier Science Publishers, pp. 347–386.

Mardia, K.V., Kent, J. T, and Bibby, J. M. (1979). *Multivariate Analysis.* New York: Academic Press.

Nychka, D. and Saltzman, N. (1998), Design of air quality networks. In *Case Studies in Environmental*

Statistics, eds. D. Nychka, W. Piegorsch and L.H. Cox, Lecture Notes in Statistics number 132, Springer Verlag, New York, pp. 51–76.

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.

Raftery, A.E. (1988). Analysis of a simple debugging model. *Journal of the Royal Statistical Society, series C - Applied Statistics*, 37, 12-22.

Rubin, D. B.(1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician, *Annals of Statistics*, **12**, 1151-1172.

Sampson, P.D. and Guttorp, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association,* **87**, 108-119.

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

Smith, R.L. (1996), Estimating nonstationary spatial correlations. Preprint, University of North Carolina.

Stein, M. L. (1999). Predicting random fields with increasingly dense observations. *Ann. Appl. Probab.,***9**, 242-273.

Ver Hoef, J.M. and R.D. Barry. (1998). Modeling crossvariograms for cokriging and multivariable spatial prediction. *Journal of Statistical Planning and Inference,* **69**, 275-294.

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.

Yaglom, A. M. (1962). *An introduction to the theory of stationary random functions.* Prentice-Hall, NJ.