Statistical assessment of numerical models

Montserrat Fuentes Peter Guttorp Peter Challenor

NRCSE

Technical Report Series

NRCSE-TRS No. 076

February 18, 2003

Statistical assessment of numerical models

Montserrat Fuentes[1] Peter Guttorp[2] and Peter Challenor[3]

ABSTRACT

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 one approach to assessment of model performance. For dry deposition fluxes and concentrations of air pollutants there is a very limited supply of evaluation data sets. We develop a formal method for evaluation of the performance of numerical models, which can be implemented even when the field measurements are very sparse. This approach is applied to a current U.S. Environmental Protection Agency air quality model. In other cases, exemplified by an ozone study from the California Central Valley, the observed field is relatively data rich, and more or less standard geostatistical tools can be used to compare model to data. Yet another situation is when the cost of model runs is prohibitive, and a statistical approach to approximating the model output is needed. We describe two ways of obtaining such approximations.

A common technical issue in the assessment of environmental numerical models is the need for tools to estimate nonstationary spatial covariance structures. We describe in detail two such approaches.

[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. This research was sponsored by a National Science Foundation grant DMS 0002790 and by a US EPA award R-8287801.

[2] Peter Guttorp is professor of Statistics, University of Washington, Seattle, WA 98195-4322. Tel.:(206) 543-6774, Fax: (206) 685-7419 , E-mail: peter@stat.washington.edu

[3] Peter Challenor is Head of Satellite Remote Sensing, James Rennell Division for Ocean Circulation and Climate, Southampton Oceanography Centre, Southampton, SO14 3ZH, UK. Tel: (+44)23 80596413, Fax: (+44) 23 80596400, E-mail: P.Challenor@soc.soton.ac.uk

1 Introduction

A major focus of the Clean Air Act, the United States main law on air pollution control, from its passage in 1970 to the 1990 amendments, has been the effect of atmospherically transported pollutants on terrestrial and aquatic ecosystems. The Clean Air Act Amendments (CAAA) of 1990 established emissions reductions to reduce risk to public health and to protect sensitive ecosystems. The CAAA also established a monitoring program to assess improvements in the Nation's air quality and overall environment. If a state or district in the US is found in violation of some of the air pollution standards in the Clean Air Act it is required to develop a plan for bringing the region back into compliance with the Act. If the violation is sufficiently substantial, the region must prove the effectiveness of the plan by developing a comprehensive deterministic air pollution model, describing emissions, air transport, chemical transformation, and deposition of the pollutant and its precursors. This model must be compared to observed data, found to describe these well (a process often called *model validation*, although a more appropriate term would be *model assessment*), and the proposed controls must, as shown by a model run under the modified conditions, bring the region back into compliance. These deterministic models produce predictions for grid squares over some temporal window. The data are obtained at individual points, and often have a different temporal resolution from the model output. Consequently, it is not possible to compare the data to the model output directly. Rather, some manipulation of the data (or the model output) is needed for comparability. The 1994 EPA Guidance on Urban Airshed Model Reporting Requirements for Attainment Demonstration suggests manipulating the model output:

> "... recommends the use of a four-cell weighted average to determine the predicted concentration to be used in comparison with observed values" [at monitoring sites].

Since the model output is already an average over the grid square, it seems inappropriate to smooth it further spatially in order to compare to non-smooth point measurements. Rather, we would be inclined to use the data to predict the model output, i.e., to predict the grid square values. However, this requires a rather data-rich situation, in which the prediction can be made with adequate precision. Furthermore notice that in order to do this, it would not be appropriate to model, statistically or stochastically, the air pollution transportation, transformation, emission and deposition processes. The deterministic model is built to model these processes, and to use another model of these processes for manipulating the data would confound the comparison between model and data.

The main objective of our work is a statistical assessment of the performance of complex air quality models. Numerical models of air pollution are a form of a highly complex scientific hypothesis concerning natural processes, that can be rejected through comparison with observations, but never confirmed. The objective comparison of modeled concentrations with observed field data provides a means for assessing model performance. Statistical evaluation of model performance is viewed as part of a larger process (which includes sensitivity analysis and other tools) that collectively is referred to as model evaluation.

Air quality simulation models have been used for many decades to characterize the transport and dispersion of material in the atmosphere. 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. The process of summarizing the overall performance of a model over the range of conditions experienced within a field experiment typically involves determining two points for the model evaluation objectives: estimate the bias in comparisons with observations, and study whether the differences seen in the comparisons are significant in light of the uncertainties in the observations.

More generally, it can be helpful (Walden and Guttorp, 1987) to set up a framework for model assessment in which the predictions (model outputs) and observations each have a decomposition into different error types.

$$P_t = X_t + M_t + S_t + N_t \tag{1}$$

$$O_t = X_t + B_t + E_t \tag{2}$$

where

X_t is the true state of nature (a spatial field)

M_t stands for model error, caused by inadequate description of the physical system, simplifying assumptions, etc.

S_t is the smoothing error due to the modeling describing a discretized (rather than continuous) spatial and temporal field

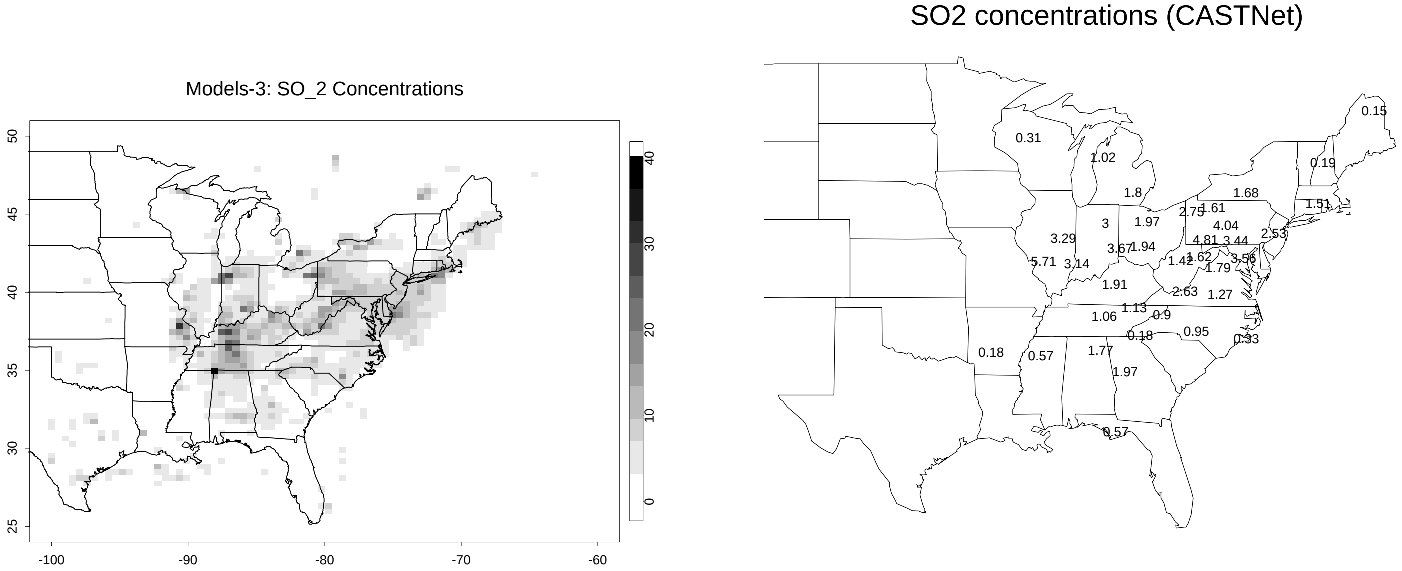
N_t describes numerical errors and/or approximations in the implementation of model predictions

B_t stands for measurement bias, caused by consistent operator error, properties of the measurement device, temporal averaging, etc.

E_t is the unavoidable measurement error, typically band-limited noise with a relatively flat spectrum

Notice that in this setup there is no "ground truth": both model output and observations have measurement error associated with them. In the decomposition we have, for simplicity, taken the components to be additive. Frequently there can be interactions between the components, such as having numerical error being a problem only when model error is large and smoothing error small. The discussion in Walden and Guttorp (op. cit.) illustrates how misleading it can be to use the correlation between model prediction and data to assess model quality: under realistic conditions, one can get high correlation between P_t and O_t if the smoothing error is a nearly linear function of the true process, i.e., when extreme episodes (something the model is developed to deal with) are poorly tracked.

In section 2, an approach to model assessment based on the Bayesian melding technique by Raftery and Poole (2000) is illustrated in a case where there is ample model output and sparse monitoring data. In section 3 we describe a geostatistical approach to model assessment, requiring monitoring data from a relatively dense network, but enabling detailed analysis of the model output. Section 4 describes a situation where it is difficult to get sufficient number of model runs, and a statistical approximation to the model output is compared to data. Finally, in section 5 we briefly compare the different scenarios that these types of assessment are suited for.



(a)

(b)

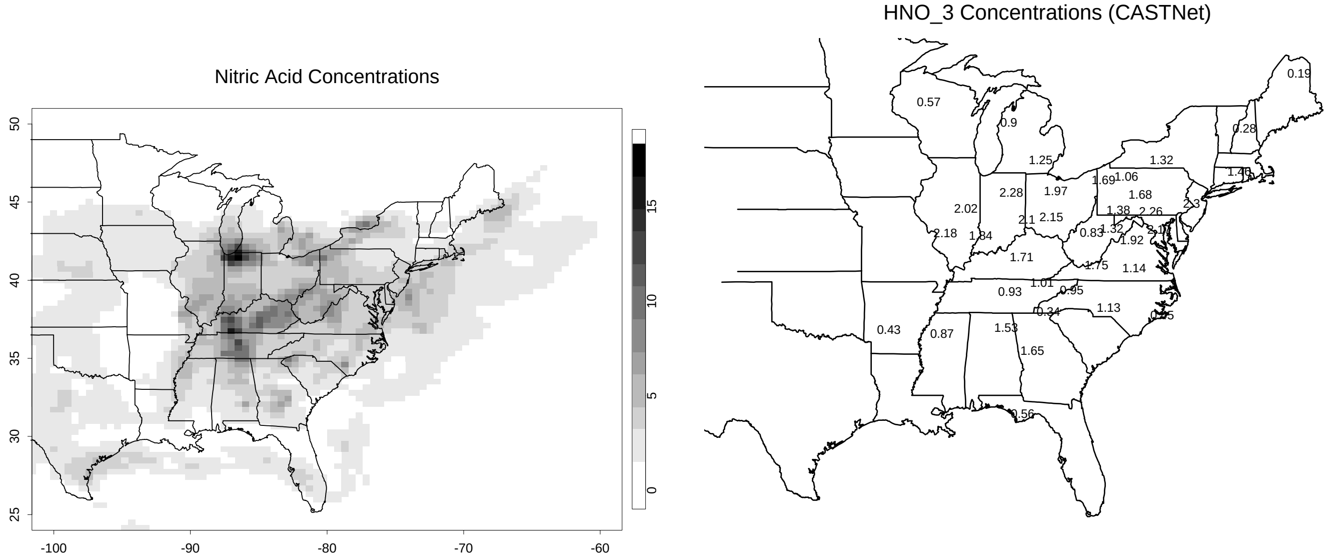
Figure 1: (a): SO_2 concentrations (ppb) from Models-3 for the week of July 11, 1995. The resolution is $36km \times 36km$. (b): SO_2 concentrations (ppb) at the CASTNet sites for the same week.

2 A Bayesian framework for model evaluation

2.1 Statistical Models for CASTNet and Models-3 Output

Maps of loadings of pollutants to aquatic and terrestrial ecosystems are needed over different geo-political boundaries, to discover when, where, and to what extent the pollution load is improving or declining. One important source of information on pollution loads over large areas are the regional scale air quality models. These models, e.g. Models-3, are run by EPA and the U.S. States and provide air pollution concentrations and fluxes in regular grids in parts of the US (see Figure 1 (a) and Figure 2 (a)). The current resolution of Models-3 is 36 km × 36 km. 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 regulation process of the air pollutants standards. EPA provides point measurements at 50 irregularly spaced sites in the eastern U.S. known as the Clean Air Status and Trends network (CASTNet) (see Figure 1 (b) and Figure 2 (b)). At each site, EPA measures dry deposition fluxes and concentrations of different atmospheric pollutants.

Models-3 is used to examine the response of the air pollution network to different control strategies under



(a) (b)

Figure 2: (a): HNO_3 concentrations (ppb) from Models-3 for the week of July 11, 1995. The resolution is $36km \times 36km$. (b): HNO_3 concentrations (ppb) at the CASTNet sites for the same week.

various high-pollution scenarios. To establish its credibility, however, it is essential that it should accurately reproduce observed measurements when applied to ground data. Models-3 uses as inputs metereological data, emissions data and boundary values of air pollution. The available emissions data are combined with numerical models of local weather (the Mesoscale Model version 5 (MM5)), the emissions process (the Sparse Matrix Operator Kernel Emissions (SMOKE)), 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 produce maps (Dennis *et al*, 1996). Models-3. are not statistical models but numerical deterministic simulation models based on systems of differential equations that attempt to represent the underlying physics, and take the form of huge blocks of computer code. To statistically assess the performance of Models-3 we need measures of how well Models-3 output and real data agree. An approach to evaluation of model performance 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). 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 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 evaluation would be questionable.

A related problem is that the comparison does not take into account the uncertainty in the interpolated values. In this section, 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. 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. 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. Fuentes and Raftery (2001) used this Bayesian framework for inference about deterministic simulation with the goal of combining data from different sources as well as for numerical model evaluation. The approach presented here could be considered an instance of the Bayesian melding approach (Poole and Raftery, 2000).

Our general modeling framework is shown in Figure 3. 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 make an observation $\hat{Z}(\mathbf{s})$, corresponding to the CASTNet observation at this station, and we assume that

$$\hat{Z}(\mathbf{s}) = Z(\mathbf{s}) + e(\mathbf{s}), \tag{3}$$

where $e(\mathbf{s}) \sim N(0, \sigma_e^2)$ represents the measurement error (nugget) 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,

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + \epsilon(\mathbf{s}), \tag{4}$$

where $Z(\mathbf{s})$ has a spatial trend, $\mu(\mathbf{s})$, that is a polynomial function of $\mathbf{s}$ with coefficients $\boldsymbol{\beta}$. If additional information about covariates is available, the spatial trend of $Z(\mathbf{s})$ can be also modeled as a function of

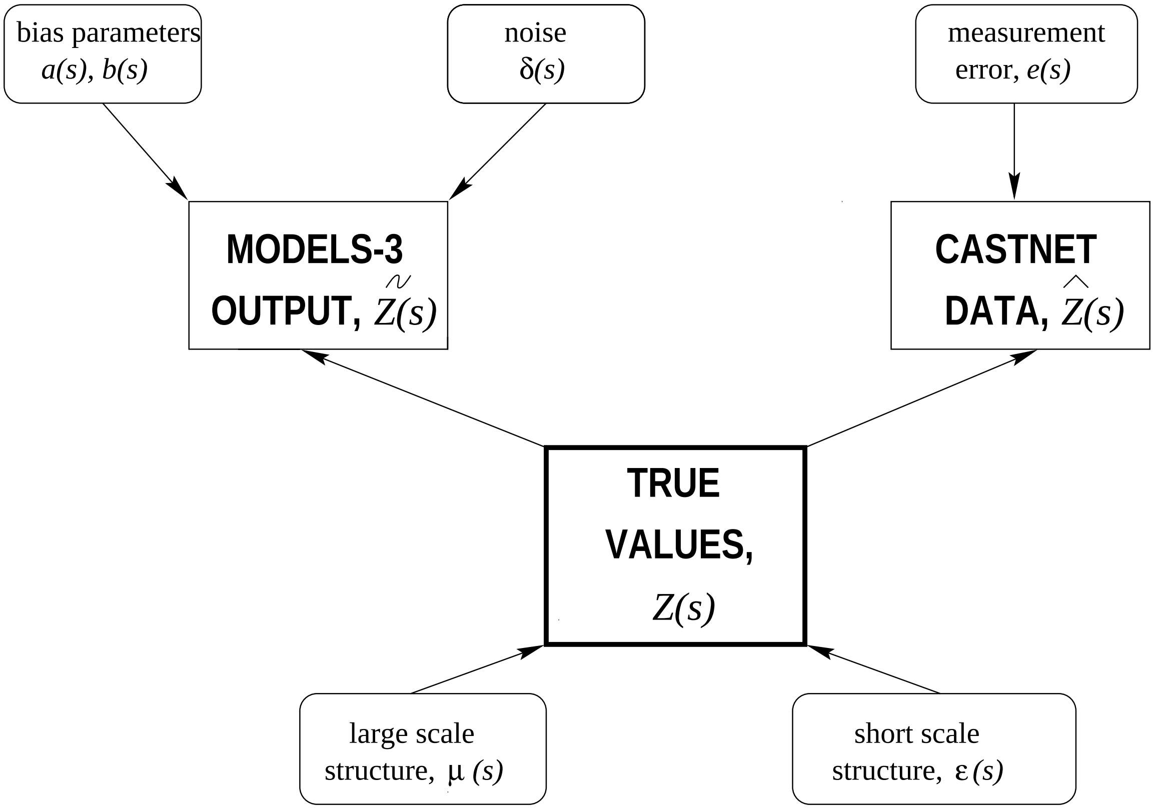


Figure 3: General Modeling Framework.

some metereological and geographic covariates $f_1, \ldots, f_p$ that are know functions at some locations $\mathbf{s}$, with unknown coefficients $\boldsymbol{\beta}$:

$$\mu(\mathbf{s}) = \sum \beta_i f_i(\mathbf{s}).$$

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:

$$\tilde{Z}(\mathbf{s}) = a(\mathbf{s}) + b(\mathbf{s})Z(\mathbf{s}) + \delta(\mathbf{s}). \tag{5}$$

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

$$\tilde{Z}(B_i) = \int_{B_i} a(\mathbf{s})d\mathbf{s} + b\int_{B_i} Z(\mathbf{s})d\mathbf{s} + \int_{B_i} \delta(\mathbf{s})d\mathbf{s} \tag{6}$$

for $i = 1, \ldots, m$. We model the function $a(\mathbf{s})$ as a polynomial in $\mathbf{s}$ with a vector of coefficients, a_0, and b is a unknown constant term.

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

$$P(\hat{Z}|\tilde{Z}, a = 0, b = 1). \tag{7}$$

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

$$P(a, b|\hat{Z}, \tilde{Z}). \tag{8}$$

2.2 Nonstationary Covariance

The spatial patterns shown by the air pollutant fluxes and concentrations change with location, so that the underlying process Z with the true values of fluxes/concentrations of air pollution is nonstationary and standard methods of spatial modeling and interpolation are inadequate. 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), which also contains descriptions of other approaches in the literature up to that time. A Bayesian framework for the deformation approach was introduced by Damian, Sampson, and Guttorp, P. (2000), and also by Schmidt and O'Hagan (2000). Maximum likelihood versions of the method were developed by Mardia and Goodall (1993) and Smith (1996). This approach requires repeated observations of the underlying field, and is therefore well suited for application to monitoring data. In a series of papers best represented by Haas (1995), Haas has proposed an approach to nonstationary spatial estimation based on moving windows. Another approach has been developed by Nychka and Saltzman (1998) and Holland *et al.* (1999), extending the "empirical orthogonal functions" (EOF) approach that is popular among atmospheric scientists.

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. The motivation for defining a spatial process as an integral of white noise can be said to go back to Whittle (1954), who gave a similar representation for discrete spatial processes. Matérn (1986) used this representation to derive a wide class of stationary spatial processes. Higdon, Swall and Kern (1999) considered extensions of the form

$$Z(\mathbf{s}) = \int K_{\mathbf{s}}(\mathbf{u})X(\mathbf{u})d\mathbf{u}, \tag{9}$$

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. This approach is promising, and a quite different idea from earlier approaches for nonstationary processes, but 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 has a quite different motivation.

Another model for nonstationary processes was 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. In this section we use this approach by Fuentes (2002) to model nonstationary covariance. 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):

$$Z(\mathbf{x}) = \int_D K(\mathbf{x} - \mathbf{s})Z_{\boldsymbol{\theta}(\mathbf{s})}(\mathbf{x})d\mathbf{s}, \tag{10}$$

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 stationary of the process. The stochastic integral (10) 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 is a polynomial function of $\mathbf{x}$ (or known functions of $\mathbf{x}$) with coefficients $\boldsymbol{\beta}$. The covariance of $Z_{\boldsymbol{\theta}(\mathbf{s})}$ is stationary with parameter $\boldsymbol{\theta}(\mathbf{s})$,

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

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

$$C_{\boldsymbol{\theta}(\mathbf{s})}(\mathbf{x}) = \frac{\sigma_s}{2^{\nu_s - 1}\Gamma(\nu_s)}(2\nu_s^{1/2}|\mathbf{x}|/\rho_s)^{\nu_s}\mathcal{K}_{\nu_s}(2\nu_s^{1/2}|\mathbf{x}|/\rho_s), \tag{11}$$

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. $\sigma_s = \text{var}(Z_{\boldsymbol{\theta}(\mathbf{s})}(\mathbf{x}))$, where the covariance parameter σ_s is usually refereed 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. The covariance $C(\mathbf{s_1}, \mathbf{s_2}; \theta)$ of Z is a convolution of the local covariances $C_{\boldsymbol{\theta}(\mathbf{s})}(\mathbf{s_1} - \mathbf{s_2})$,

$$C(\mathbf{s_1}, \mathbf{s_2}; \boldsymbol{\theta}) = \int_D K(\mathbf{s_1} - \mathbf{s})K(\mathbf{s_2} - \mathbf{s})C_{\boldsymbol{\theta}(\mathbf{s})}(\mathbf{s_1} - \mathbf{s_2})d\mathbf{s}. \tag{12}$$

.

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

$$Z_{\boldsymbol{\theta}(\mathbf{s})}(\mathbf{x}) = \int_{\mathbb{R}^2} K_{\mathbf{s}}(\mathbf{u} - \mathbf{x})X_{\mathbf{s}}(\mathbf{u}) \tag{13}$$

where $K_{\mathbf{s}}$ is a kernel for each $\mathbf{s}$, and $X_{\mathbf{s}}$ are independent white noise processes for each $\mathbf{s}$. In the present paper, we consider a model in which $X_{\mathbf{s}}$ is an independent process for each s. In that case, we could combine (13) and (10) into a single integral representation for $Z(\mathbf{x})$, but one which does not reduce to the form (9). An alternative approach would be to make $X_{\mathbf{s}}()$ a common white noise process for all $\mathbf{s}$, but that turns out to be harder to implement computationally than the present approach.

In (12) 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

$$C_M(\mathbf{s_1}, \mathbf{s_2}; \boldsymbol{\theta}) = M^{-1} \sum_{m=1}^{M} K(\mathbf{s}_1 - \mathbf{s}_m) K(\mathbf{s}_2 - \mathbf{s}_m) C_{\boldsymbol{\theta}(\mathbf{s}_m)}(\mathbf{s_1} - \mathbf{s_2}). \tag{14}$$

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. this paper, function $K(\mathbf{x} - \mathbf{s})$. A final complication is the "change of support" problem. 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. 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 × 36km. We have specified a covariance function between points, not grid boxes. A point covariance function is needed if we are to obtain predictive distributions at individual locations, as is the objective. However, the Models-3 data are on grid boxes, not at individual locations. The model fitting must reflect this discrepancy between the scale of model-based grid cell averages and monitoring data which are taken at individual locations. We derive the covariances of the block averages $Z(B_i)$, $i = 1, \ldots, N$, in terms of the pointwise covariance $C(\mathbf{u}, \mathbf{v})$.

$$\mathrm{cov}(Z(B_i), Z(B_j)) = \int_{B_i} \int_{B_j} C(\mathbf{u}, \mathbf{v}) d\mathbf{u} d\mathbf{v} / |B_i||B_j|, \tag{15}$$

where

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

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

$$\mathrm{cov}(Z(\mathbf{s}_i), Z(B_j)) = \int_{B_j} C(\mathbf{s}_i, \mathbf{v}) d\mathbf{v} / |B_j|. \tag{16}$$

The integrations in (16) are replaced by discrete sums for computational convenience. 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 $Z(B_1), Z(B_2), \ldots, Z(B_N)$.

2.3 Estimation

In this section we explain how to efficiently implement our algorithm for numerical models evaluation, using the approach proposed by Fuentes and Raftery (2001).

2.3.1 Algorithm

1. Posterior predictive values for CASTNet given Models-3 The quantity of interest is the predictive distribution for $\hat{Z}(\mathbf{x}_0)$ given the values of $\tilde{Z}$. We approximate the predictive distribution with the Rao-Blackwellized estimator, conditioning on the posterior simulated values for all the parameters, using for this simulation the following Gibbs algorithm.

2. Algorithm for Gibbs sampling We discuss now 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

   $$[Z|\boldsymbol{\beta}, \boldsymbol{\theta}] \text{ is Gaussian,}$$

   where the brackets [ ] are used here to denote densities.

   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

   $$[\hat{Z}(\mathbf{s})|Z(\mathbf{s}), \sigma_e^2] \text{ is normal with mean } Z(\mathbf{s}) \text{ and variance } \sigma_e^2.$$

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

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

Gibbs sampling: Stage 3. 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 evaluation, 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. For bias removal of the air quality models, we simulate values of the parameters a and b from the posterior distribution:

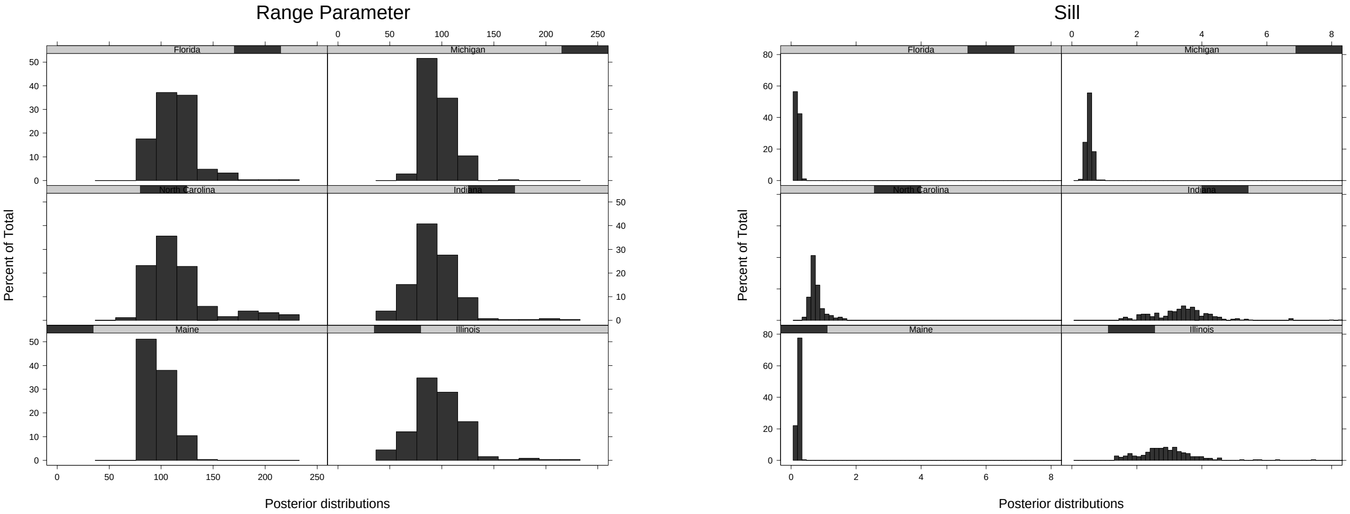
$$P(a, b|\hat{Z}, \tilde{Z}),$$

obtained in Stage 2 of the Gibbs sampling approach.

2.4 Statistical Assessment of Air Quality Models

The regional scale air quality models (Models-3) run by the U.S. EPA estimate hourly concentrations and fluxes of different air pollutants. The spatial domain, D, is a regular grid (81×87), the dimensions of each pixel in the grid are 36km $\times$ 36km. Models-3 provides hourly concentrations for each pixel. We study here sulfur dioxide and nitric acid. Figure 1 (a) shows the weekly averaged concentrations of SO_2 from Models-3 for the week starting July 11, 1995. Figure 2 (a) shows the weekly values of NHO_3 from Models-3. SO_2 is a primary pollutant, so it is emitted directly from its sources, and it tends to be more spatially heterogeneous than HNO_3, or other secondary pollutants, e.g. O_3. NHO_3, as a secondary pollutant, is the result of photochemical reactions in the atmosphere. The Clean Air Status and Trends Network (CASTNet) measures weekly averaged concentrations and fluxes at 50 sites of different pollutants, Figures 1 (a) and 2 (b) show the SO_2 and the HNO_3 values respectively for the week starting July 11, 1995. We use the methodology presented in Section 2.1 to evaluate Models-3 and to estimate the bias. 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 2.1). We modeled the covariance for Models-3 using equation (12), 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 4 locations in each pixel).

We used inverse gamma priors with infinite variance for all the Matérn covariance parameters, except for the sill parameter for which we used a uniform prior in the log scale. Figures 4 (a) and (b) show the posterior distributions of some covariance parameters for the underlying process Z at 6 selected sites. The sill parameter changes with location as illustrated by the variation in the distributions in Figure 4 (b). Thus, this indicates a lack of stationarity. The range parameter does not change much with location (Figure 4 (a)). The smoothing parameter does not change with location either, and is always close to 1/2 (exponential). We implemented the nonstationary model (10) with weight function $K(\mathbf{u}-\mathbf{s}) = \frac{1}{h^2}K_0\left(\frac{\mathbf{u}-\mathbf{s}}{h}\right)$, where $K_0(\mathbf{u})$



(a)

(b)

Figure 4: (a): Posterior distributions for the range parameter (km) of the Matérn covariance for the SO_2 concentrations of Z, for the week starting July 11, 1995, at 6 selected sites located (starting from the bottom left panel) in ME, IL, NC, IN, FL and MI respectively. (b): 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.

is the quadratic weight function

$$K_0(\mathbf{u}) = \frac{3}{4}(1 - {u_1}^2)_+ \frac{3}{4}(1 - {u_2}^2)_+, \tag{17}$$

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 (14), 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 (ppb), and for Models-3 it is .1 (ppb). The mode of the posterior distribution for the parameter that measures the multiplicative bias for Models-3 is .5 (ppb) with a standard error of .5 (ppb), and for the additive bias we have a polynomial of degree 4.

The graph on the left in Figure 5 presents a naive approach for evaluation of Models-3. This graph shows Models-3 versus CASTNet, without doing any spatial interpolation of Models-3. In this graph we simply

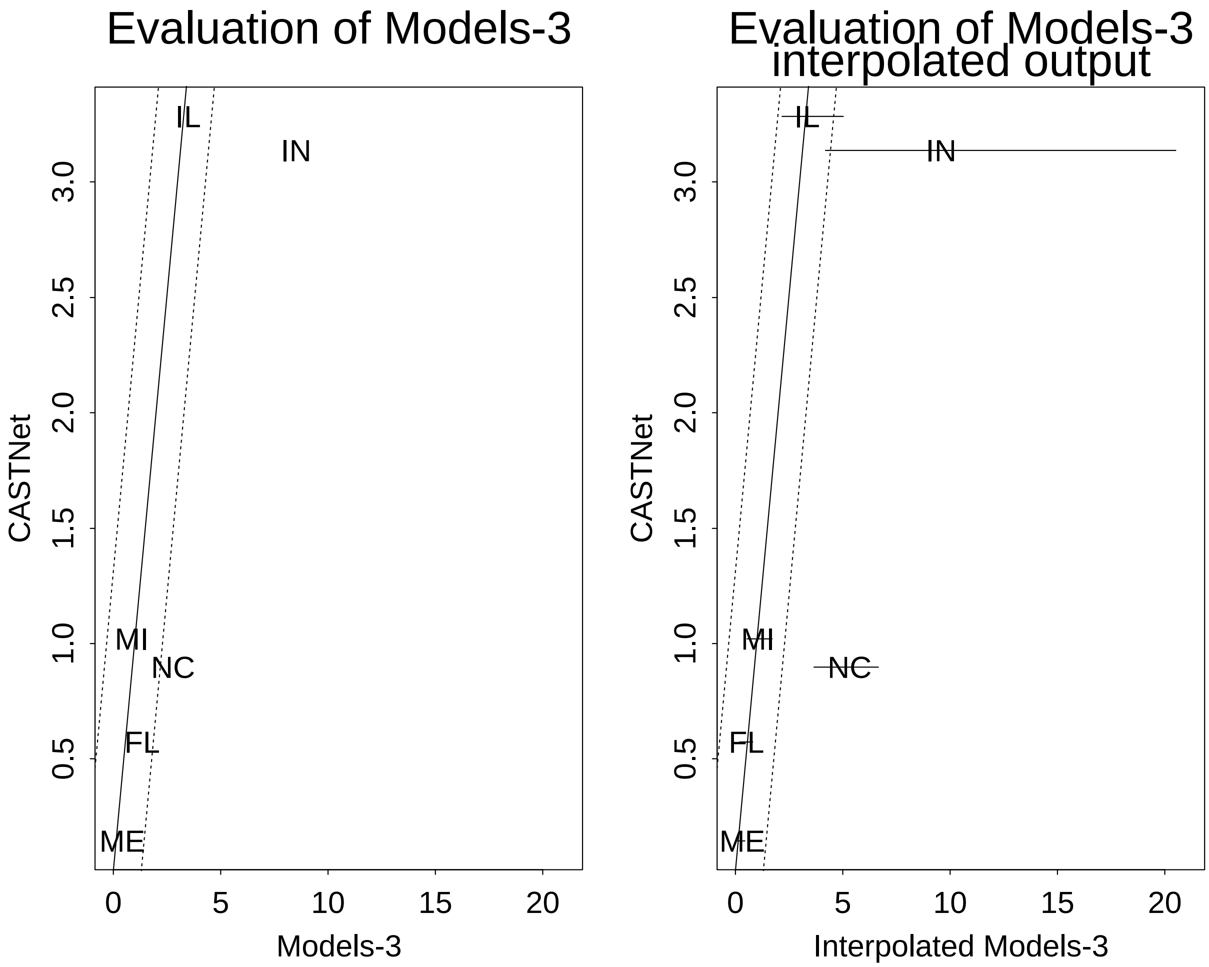


Figure 5: The graph on the left shows CASTNet measurements for the week starting July 11, 1995, versus the values of Models-3 for the pixels that are the closest to each CASTNet site, without considering the change of support. The graph on the right shows the CASTNet measurements versus the modes and 90% credible intervals of the predictive Bayesian distributions derived from Models-3 at the CASTNet locations.The dotted lines indicate a 90% confidence region for the CASTNet values.

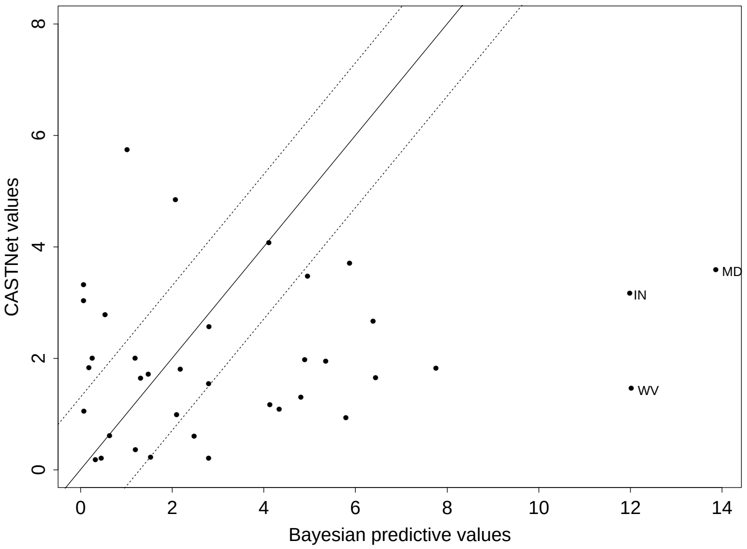


Figure 6: CASTNet values of SO_2 versus the mean of the predictive posterior distribution for Models-3 at each site.

have the values of Models-3 for the pixels that are the closest to each CASTNet site, without considering the change of support. In some areas the atmospheric pollutants vary significantly at scales smaller than the grid size of the model, therefore comparing the value of the grid cell with a point measurement in the ground would lead to erroneous conclusions. In Figure 5 the dotted lines indicate a 90% confidence region for CASTNet (CASTNet values $\pm 1.64 \cdot \sigma_e$). The graph on the right in Figure 5 shows CASTNet measurements versus the modes and 90% credible intervals of the predictive Bayesian distributions (eq. (7)) derived from Models-3 at the CASTNet locations for evaluation of Models-3. Some of the modes in the latter plot do not fall within the credible bands for CASTNet. The latter plot is much more informative about the fit of Models-3 to the real data, since we compare values that have the same spatial support instead of comparing grid cells with point measurements. The uncertainty in the estimated Models-3 values in this figure depends on location. The 90% credible intervals in Figure 5 show that at some locations, the bias in Models-3 is not significant. This Bayesian approach gives more reliable prediction errors, by taking into account the uncertainty in the covariance parameters, and the change of support.

Figure 6 we plot all SO_2 CASTNet values versus the means of the predictive distributions of Models-3

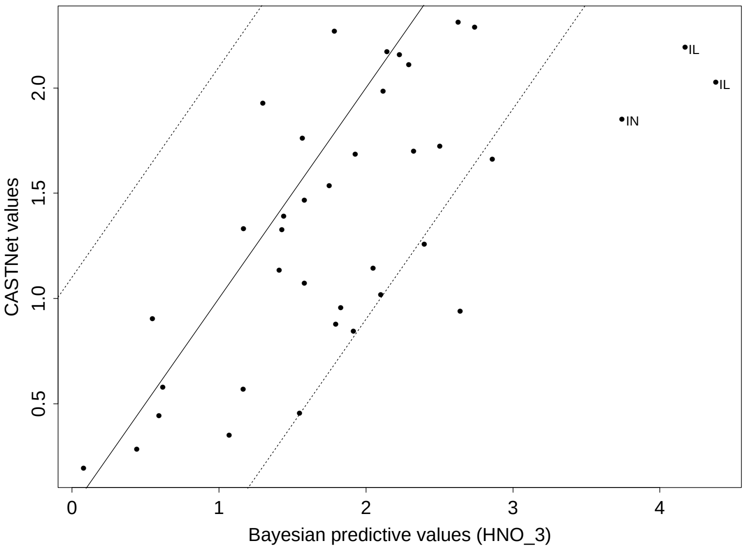


Figure 7: CASTNet values for HNO_3 versus the mean of the predictive posterior distribution for Models-3 at each site.

at those locations. The dotted lines in Figure 6 indicate a 90% confidence region for the SO_2 CASTNet measurements, and the solid line has slope 1 and intercept 0. There are 3 sites where Models-3 overestimates the SO_2 values considerably, a site in Indiana, a site in Maryland, and a site in West Virginia. These three sites are close to power plants. In Figure 7 we have all HNO_3 CASTNet values versus the means of the predictive distributions of Models-3 at those locations. The dotted lines in Figure 7 indicate a 90% confidence region for the HNO_3 CASTNet measurements, and the solid line has slope 1 and intercept 0. For the nitric acid, the CASTNet values have a smaller measurement error .7 (ppb) versus .8 (ppb) for SO_2. However, the relative standard error (coefficient of variation) is larger for HNO_3 than for SO_2. The sites where Models-3 seems to perform worse are located in Illinois and Indiana, these are again sites close to power plants. Figures 6 and 7 is just an illustration of the powerful application of the technique presented in this paper, though it does arises some important questions that are currently being discussed with the Models-3 group regarding the formulation and improvement of the physical models. Some of the main sources of uncertainty that affect the performance of the models are the following; the photo-chemistry model parameterizations (which is the treatment of photo-chemistry 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 fact that for the SO_2 the models perform worse in areas closer to power plants suggests that the dispersion modeling of pollution plumes in that areas 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. New dispersion models are currently being added to Models-3. For more information about dispersion modeling of pollution plumes, see e.g. Beychok (1995). For the HNO_3, Models-3 seem to also perform worse in areas with high emissions, this probably indicates that the photo-chemical parameterizations at scales smaller than the grid size of the model need to be improved.

3 A geostatistical aproach

3.1 The SARMAP study

In 1990 several air quality research groups and govenment organizations went together in the SARMAP project to gather data at a fairly dense network in California's San Joaquin Valley, in order to evaluate the output from a model. SARMAP is a multi-acronym, standing for the SJVAQS/AUSPEX Regional Model Adaptation Project, where SJVAQS stands for the San Joaquin Valley Air Quality Study and AUSPEX for Atmospheric Utility Signatures, Predictions and Experiments. The SJVAQS focused on determining the causes of the exceedances in the San Joaquin Valley of the U.S. ozone air quality standard (120 ppb maximum one hour average[4]). AUSPEX was intended to develop a comprehensive model addressing ozone, aerosol, visibility, and acid deposition issues and to obtain a high-quality data base for model evaluation and application.

Data were available from two summer months in 1990 with hourly samples at 131 stations. Meteorological variables as well as ozone precursors were also measured in the SARMAP study. In the application by Meiring et al. (1998) a subset of these data was used to assess some runs of the AUSPEX model. In this case the models runs were only made available for a few short extreme episodes during the two months.

[4] A revised maximum eight hour average standard has not been implemented due to legal complications

A preliminary data analysis indicated that the ozone process is neither temporally stationary nor spatially homogeneous (Guttorp et al., 1994). There is a diurnal cycle with peak ozone concentrations in the afternoons, and minimal concentrations during the night. This is explained by the fact that ozone formation is a photochemical reaction, with a variety of ozone sinks (and virtually no production) during the night. In order to estimate the space-time covariance of the data, we first take out a daily mean from square-root transformed measurements at each station. The residuals from the mean exhibit temporal autocorrelation, so an AR 2-model was fit to each station separately. To the residuals from the mean and time series models, we want to fit a spatial covariance model. Because of the nature of ozone data, we expect this spatial covariance to vary with the hour of the day (being stronger in the afternoons at times of peak ozone production and weaker during the night-time depletion), as well as showing a meteorologically and orographically induced heterogeneity. In other words, the space-time covariance structure is non-separable, spatially heterogeneous and anisotropic. Assumptions of space-time separability, spatial homogeneity (or stationarity) and isotropy are commonly made in many geostatistical applications, but are not appropriate here. In the next subsection we present an approach to nonstationary covariance estimation that is different from the Fuentes approach in Section 2.

3.2 Bayesian estimation of the spatial deformation model

We assume that temporally independent samples $Z_{it} = Z(x_i, t)$ are available at each of N geographic locations and at the same T points in time: $i = 1, 2, \dots, N$ and $t = 1, 2, \dots, T$. We consider the following model for the underlying process:

$$Z(x,t) = \mu(x,t) + \nu(x)^{1/2} E(x,t) + \varepsilon(x,t) \tag{18}$$

where x denotes location and t time. $\mu(x,t)$ represents the spatial-temporal mean field. $E(x,t)$ is a mean zero, variance one, Gaussian spatial process with a correlation function that depends smoothly on the geographic coordinates. We assume that $E(x,t)$ is mean square continuous in space, namely, $\mathrm{E}\left(E(x+h,t) - E(x,t)\right)^2 \to 0$, as $h \to 0$. $\nu(x)$ is a smooth function representing the variance (in time) of the process observed at location x. $\varepsilon(x,t)$ is a white-noise process (i.e. it has mean zero, constant variance and zero correlation), independent of $E(x,t)$.

Each of the components of this model requires careful thought and modeling. In this subsection we consider a flat prior for the space-time mean field, $\mu(x,t)$, to remove it from the likelihood. The process $\varepsilon(x,t)$ is a white-noise process representing measurement error and small scale spatial variability. The variance of the space-time process at points in space, $\nu(x)$, is itself modeled as a random field with spatial structure related to the structure of the process $E(x,t)$, where t simply indexes temporally independent replicates of the spatial process. This is the focus of our analysis here. Damian et al. (2000) contains technical details for the case of constant variance field, while Damian et al. (2003) describes the modification needed for spatially varying variances.

Many spatial processes may be modeled as Gaussian, either directly, or after a suitable transformation. We thus model $E(x,t)$ as normal, with mean zero, variance 1 ($\nu(x)$, the variance of $Z(x,t)$, is considered separately) and a Sampson-Guttorp correlation function (Guttorp and Sampson, 1994)

$$\mathrm{Corr}\left(E(x), E(y)\right) = \rho_\theta\left(\|f(x) - f(y)\|\right) \tag{19}$$

The process $E(x,t)$ is assumed to be mean square continuous. This assumption is equivalent to the continuity of $\rho_\theta(\cdot)$ at the origin, that is, $\rho_\theta(d) \to 1$ as $d \to 0$. The correlation function of $Z(x,t)$ does not necessarily converge to 1. This fact is accounted for by the process $\varepsilon(x,t)$. We assume that $\rho_\theta(\cdot)$ belongs to a known parametric family with unknown parameter(s) θ.

The covariance and correlation functions of the observed process $Z(x,t)$ are easily obtained from the model (18)

$$\mathrm{Cov}\left(Z(x,t), Z(y,t)\right) = \begin{cases} \left(\nu(x)\nu(y)\right)^{1/2} \rho_\theta\left(\|f(x) - f(y)\|\right) & x \neq y \\ \nu(y) + \sigma_\varepsilon^2 & x = y \end{cases} \tag{20}$$

In summary, the elements of the model to be estimated are:

- θ, the parameter(s) of the correlation function
- f, the spatial deformation, represented as a pair of thin-plate splines
- $\nu(x)$, the spatial variance, represented as a random field
- σ_ε^2, the nugget effect

- $\mu(x,t)$, the mean.

We model the function f as a pair of thin-plate splines because this provides a flexible family of deformations that provides a basis for characterizing spatially varying anisotropy in the spatial covariance structure. In addition, the spline representation provides a convenient parameterization for specifying priors that penalize deformations that are not smooth. Simple affine transformations play a special role in this model as they correspond to stationary, but anisotropic models. The parameterization allows us to specify priors (penalties) for both the degree of anisotropy and the smoothness ("bending energy") of the non-affine component of the deformation, which corresponds to the spatial scale at which we represent nonstationarity.

3.3 The geostatistical model assessment approach

In analyzing the SARMAP data (Meiring et al., 1998) the procedure needed to assess the model had the following steps:

1. Square root transform data

2. Take out station means

3. Prewhiten station by station using an AR(2)-model

4. Fit a nonstationary covariance model hour by hour to the residuals

5. Estimate means at grid squares

6. Predict untransformed grid square values from mean model and postcolored residuals

   Model-based hourly grid square calculations can now be compared to the data-based estimates, which have associated standard errors. Spatial and/or temporal discrepancies indicate areas where the model does not correspond well to data. Note that it is important, in order to avoid confounding, not to include the atmospheric variables in the data analysis, since the model also uses these variables.

   The analysis in Meiring et al. (1998) indicated difficulties with the model involving the location of afternoon peaks, and lack of sufficient sinks at nighttime. The latter, often considered unimportant by modelers, may be particularly troublesome if the model is used to evaluate alternative scenarios using proposed control strategies, since this will force the model to operate outside the parameters for which it was tuned. Hence, regiions in space and/or time where a model consistently fails may be an indication of structural problems.

4 The Statistical Analysis of Computer Code Output (SACCO)

A complementary approach to comparing model output to data is to perform a statistical analysis of the model itself. However complex the numerical model we can consider it simply as a function mapping a set of inputs to a set of outputs. By inputs here we do not only mean the model forcings but also any uncertain internal parameters. SACCO methods are concerned with the statistical analysis of this function. In particular we want to know what is the uncertainty on the model outputs given uncertainty on the inputs. As a consequence of this we can also find which values of the inputs give the best fit to data (calibration), look at the sensitivity of the outputs to inputs and make model predictions in an efficient manner.

The basis of these methods is to use what is called an *emulator*. An emulator is a statistical approximation to the output of the numerical model. We can either analytically compute the statistical properties of the emulator or it is simple enough to allow us to use Monte Carlo methods that are computationally impractical with the full numerical model. In a number of papers OHagan (Haylock and O'Hagan, 1996; O'Hagan and Haylock, 1997; O'Hagan *et al* 1998; Kennedy and O'Hagan, 2000; Kennedy and O'Hagan, 2001; Kennedy *et al*, 2002; Oakley and O'Hagan, 2002;) has developed Bayesian SACCO methods. Similar linear Bayes methods have been explored by Goldstein and co-workers (Craig et al, 1996, 1997, 2001). In both approaches the emulator is modelled as a random function. It could be argued that since our numerical model is deterministic it should not be modelled as a random process. However until we run the model we are ignorant as to its value and the expensive of obtaining output means that over most of the models domain we will not know the true value.

We use a Gaussian process to model our initial beliefs. To some extent the use of a Gaussian process is arbitrary. We could use any other method of modelling a random process. However Gaussian processes are adaptable, can fit any function and can be shown to be equivalent to some of the other candidate methods, for example neural networks. For a good description of Gaussian processes in this context see Kennedy and OHagan (2001). Thus if our random function is given by $\eta(\mathbf{x})$ ($\eta(\mathbf{x}) \in \mathbf{R}$) where $\mathbf{x}$ is the vector of inputs defined on a subset of R^m. Denote the mean function of η by $m(\mathbf{x})$ and the covariance function by $v(\mathbf{x_1}, \mathbf{x_2})$.

Our priors on the mean and covariance functions are

$$m(\mathbf{x}) = \mathbf{h}(\mathbf{x})^{\mathbf{T}}\beta \tag{21}$$

and

$$v(\mathbf{x_1}, \mathbf{x_2}) = \sigma^{\mathbf{2}}\mathbf{c}(\mathbf{x_1}, \mathbf{x_2}) \tag{22}$$

where $\mathbf{h}(\mathbf{x})$ is a vector of q regression functions and β is a vector of q parameters.

The form of $h(\mathbf{x})$ is arbitrary. Low order polynomials, or even simply a constant, have proved effective. In specifying the correlation function stationarity is assumed, i.e. $c(\mathbf{x_1}, \mathbf{x_2}) = \mathbf{c}(\|\mathbf{x_1} - \mathbf{x_2}\|)$ where $\|\mathbf{x_1} - \mathbf{x_2}\|$ is a suitable norm. The form of c is not critical. Most examples so far have used a Gaussian correlation function

$$c(\mathbf{x_1}, \mathbf{x_2}) = \mathbf{exp}(-(\mathbf{x_1} - \mathbf{x_2})^{\mathbf{T}}\mathbf{B}(\mathbf{x_1} - \mathbf{x_2})) \tag{23}$$

Our prior knowledge can now be expressed in terms of the parameters β, σ^2 and $\mathbf{B}$. For a full Bayesian analysis we would now specify a probability distribution for each of these parameters. However including the smoothing parameters, $\mathbf{B}$, in such an analysis makes the problem intractable. These are therefore dealt with in a non-Bayesian way.

We now run the model at a number, n, of values of the inputs $\mathbf{x_1}, \ldots, \mathbf{x_n}$ to produce data, d. The selection of these design points is an important problem and is discussed further below. From the prior we have that

$$d|\beta, \sigma^2 \sim N(\mathbf{H}\beta, \sigma^{\mathbf{2}}\mathbf{A}) \tag{24}$$

where

d are the model results corresponding to the inputs $\mathbf{x_1}, \ldots, \mathbf{x_n}$

$$\mathbf{H^T} = (\mathbf{h}(\mathbf{x_1}), \ldots, \mathbf{h}(\mathbf{x_n}))$$

and

$$\mathbf{A} = \begin{pmatrix} 1 & c(\mathbf{x_1}, \mathbf{x_2}) & \cdots & c(\mathbf{x_1}, \mathbf{x_n}) \\ c(\mathbf{x_2}, \mathbf{x_2}) & 1 & & \vdots \\ \vdots & & \ddots & \\ c(\mathbf{x_1}, \mathbf{x_n}) & \cdots & & 1 \end{pmatrix}$$

Now we can update the distribution of $\eta(.)$ using the properties of conditional Normal distributions (Krzanowski, 1988)

$$\eta(.)|\beta, \sigma^2, d \sim N(m^*(.), \sigma^2 c^*(.,.)) \tag{25}$$

where

$$m^*(x) = h(x)^T)\beta + t(x)^T \mathbf{A^{-1}}(\mathbf{d} - \mathbf{H}\beta)$$

$$c^*(x, x\prime) = c(x, x\prime) - t(x)^T \mathbf{A}^{-}\mathbf{1t(x\prime)}$$

$$t(x)^T = (c(x, x_1), \ldots, c(x, x_n))$$

and

$$d^T = \eta(x_1), \ldots, \eta(x_n) \tag{26}$$

We now need to integrate out β and σ^2 Combining and using Bayes theorem we get

$$\beta|\sigma^2, d \sim N(\hat{\beta}, \sigma^2(\mathbf{H^T A^{-1} H})^{\mathbf{-1}}) \tag{27}$$

and

$$\sigma^2|d \sim \{n - q - 2\}\hat{\sigma}^2 \chi^2_{n-q} \tag{28}$$

where

$$\hat{\beta} = (\mathbf{H^T A^{-1} H})^{\mathbf{-1}} \mathbf{H^T A^{-1} d} \tag{29}$$

$$\hat{\sigma}^2 = \frac{d^T(\mathbf{A^{-1}} - \mathbf{A^{-1}H}(\mathbf{H^T A^{-1} H})^{\mathbf{-1}}\mathbf{H^T A^{-1}})\mathbf{d}}{\mathbf{n - q - 2}} \tag{30}$$

Take the product of (22), (25) and (26) to get the distribution of $\eta(x), \beta, \sigma^2|d$ and then integrate out β and σ^2. this gives us the marginal distribution of $\eta(x)|d$ as

$$\frac{\eta(x) - m^{**}(x)}{\hat{\sigma}\{c^{**}(x,x)\}^{\frac{1}{2}}} \sim t_{n-q} \tag{31}$$

where

$$m^{**}(x) = h(x)^T\hat{\beta} + t(x)^T\mathbf{A^{-1}}(\mathbf{d} - \mathbf{H}\hat{\beta}) \tag{32}$$

and

$$c^{**}(x,x) = c^*(x,x) + (h(x)^T - t(x)^T\mathbf{A^{-1}H})(\mathbf{H^TA^{-1}H})^{\mathbf{-1}}(\mathbf{h(x)^T} - \mathbf{t(x)^TA^{-1}H})^{\mathbf{T}} \tag{33}$$

$m^{**}(x)$ gives us a quick approximation, the emulator, to the model. Thus we do not need to run the full, expensive model in future. We also have the variance for this estimate of the model output. At the design points, where we have run the model, m^{**} equals the model output and at points in between it smoothly interpolates. At first look it is surprising that an emulator for a complex non-linear model can be so simple. As an example in figure 8 we show the emulation of a simple, but highly non-linear model the function y=7+x+cos(2x). Using only a linear regression function (i.e. $h^T(x) = (1x)$) and five model evaluations a very realistic approximation (shown dashed) is made to the function (the solid line).

The uncertainty distribution

We now consider the distribution of the outputs in terms of some statistical distribution of the inputs.

Consider the inputs x as random variables with a distribution G. Then $Y = \eta(x)$ is also a random variable. We call the probability distribution of Y induced by the distribution of x the uncertainty distribution. Because the model is so complex it is non-trivial to obtain the distribution of Y given the distribution of x. Before considering the entire distribution we look at the moments of Y. The mean of the uncertainty distribution, K, is given by

$$K = \int \eta(x)dG(x) \tag{34}$$

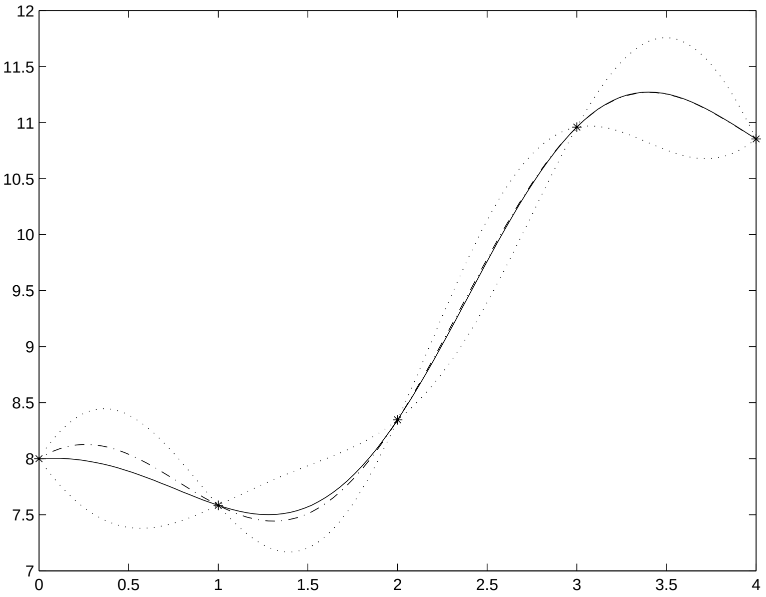


Figure 8: The function y=7+x+cos(2x) with an emulator(-.) and its 95% confidence limits (..). The emulation used only the values of the function at the *'s and a linear prior

Following O'Hagan (1992) it can be shown that

$$\frac{K-\hat{K}}{\hat{\sigma}\sqrt{X}}|d \sim t_{n-q} \tag{35}$$

where

$$\hat{K} = \int m^{**}(x)dG(x) = R\hat{\beta} + \mathbf{TA^{-1}}(\mathbf{y} - \mathbf{H}\hat{\beta})$$
$$R = \int h(x)^T dG(x)$$
$$T = \int t(x)^T dG(x)$$

Thus we have the posterior distribution of the mean of the uncertainty distribution. It is not possible to calculate the full posterior for the variance but it is possible to compute the first two moments.

If we want the full uncertainty distribution rather than simply its moments we can use Monte Carlo techniques. These are discussed in O'Hagan, Kennedy and Oakley (1998) and Oakley and O'Hagan (2002). An interesting aspect of these methods is the use of so called sampling design points. Since the full simulator is expensive to run we wish to minimise the number of runs we need. Oakley and O'Hagan (2002) supplement the actual simulator runs with runs of the emulator which are then treated as if they were simulator runs. Thus we have a two stage sampling process. Before any sampling takes place we run the simulator according to some design. These points are fixed and the emulator is forced to have the same value as the simulator at these points. In the first stage of the sampling we sample from the emulator at the sampling design points. We now fix these points as if they were from the simulator itself and sample from the uncertainty distribution. the use of sampling design points improves the efficiency of the sampling of the uncertainty distribution.

There are two outstanding issues with these methods. The first is the specification of smoothing parameters (**B**). The second is the dimensionality of the problem.

As we noted above the smoothing parameters are not dealt with in a fully Bayesian way. The methods we have described assume that they are known. There are two ways they can be estimated. The first (Oakley and O'Hagan; 2002) is to maximise the posterior probability. The other is to use cross-validation: exclude each simulator run in turn and use the reduced subset to predict its value. The value of the smoothing parameters that minimise the squared difference is then used for the rest of the calculations.

The other major issue is that of dimensionality. The methods described earlier in this section can deal with many inputs and a single output. We can extend the method to multiple outputs by adding an index to the inputs. So if we are interested in ten outputs, say, we would add an extra input variable taking the values one to ten indicating which output we are dealing with. A more important the problem is the dimension of the simulator itself. Numerical models of environmental process are often of very large dimension. The size of the matrices we have to deal with increases with the dimensionality of the problem. Thus the computational burden rises rapidly (although it will still normally be small compared to running the simulator itself). In addition as the dimension of the input space increases we will have to make more simulator runs to ensure that the emulator is interpolating rather than extrapolating the behaviour of the non-linear simulator. These issues are active research topics.

A topic which is likely to be of increasing importance as these methods are applied to increasingly complex (and higher dimension) problems is the use of hierarchies of simulators. These methods are described in Craig et al (1996), (1997), (2001), O'Hagan *et al* (1998) and Kennedy and O'Hagan(2000).

Simpler versions of the simulator which are faster to run, and possibly have lower dimension, are used to gain information about the form of the emulator. Used successfully this method can reduce the number of runs of the full simulator to a handful.

4.0.1 Bayes Linear Methods

Craig et al (2001) present an interesting alternative to the O'Hagan method. They are particularly interested in combining simulator output, data and expert opinion to make predictions. Their example, an oil reservoir simulator, is of high dimension (order 100). They conclude that a Bayesian solution to their problem (similar to the methods described above) is computationally impracticable for anything but low dimensional problems. They therefore propose a Bayes linear solution. Bayes linear methods (Goldstein, 1999) produce inferences only on the first two moments of the distribution rather than the full po. The relevant equations are

$$E_{z_P}(y_F) = E(y_F) + cov(y_F, z_P)var(z_P)^{-1}(z_P - E(z_P)) \tag{36}$$

$$var_{z_P}(y_F) = var(y_F) - cov(y_F, z_P)var(z_P)^{-1}cov(z_P, y_F) \tag{37}$$

where y_F is the future system output and z_P are the past system inputs. These include both measured data and simulator output. As above the simulator is emulated using a Gaussian process but they pay much more attention to producing a good prior model. Because of the high dimensionality of their problem they introduce the concept of active inputs. These are those inputs that are most important in explaining variation in each output and are used to reduce the dimensionality of the function $\mathbf{h}(\mathbf{x})$. To cope with the extra variability from the other inputs an extra error term is added to the emulator so it is not forced to pass through the simulator output values. Their method produces good predictions for a problem of large dimension. However since the inferences are limited to the first two moments it is difficult to make probability statements without, for example, assuming a Normal distribution.

5 Discussion

Evaluation of the performance of a numerical model is mostly constrained by the amount and quality of observational data available for comparison with modeling results, and by the ease with which the models can provide runs that are appropriate to compare to the data.

The simulation models are capable of providing estimates of a larger set of conditions than for which there is observational data. In many cases the uncertainty associated with such estimates is very difficult to assess, being related to how well the model can describe the actual physical processes beyond the conditions for which the model has been optimized.

Furthermore, most models do not provide estimates of directly measurable quantities. For instance, grid models even if a model provides an estimate of the concentration at a specific location it represents an average over some volume of air, for example, grid average. In this paper we present statistical methodologies aimed to deduce the statistical significance of differences seen in model performance in the face of all these large uncertainties and variation.

The approaches presented here can also be used for a second order model assessment, by comparing the spatial covariance of the data on the ground with the spatial covariance of the posterior distribution of the model output. This was not possible in the SARMAP experiment, since there were insufficient number of model runs, and also not possible in the Models-3 setup since the spatial data density was too low.

References

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

Clarke, J. F. (1964). A simple diffusion model for calculating point concentrations from multiple sources. *Journal of the Air Pollution Control Association,* **14**, 357-352.

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

Craig, P.S., Goldstein, M., Seheult, A.H. and Smith, J.A. (1996) Bayes linear strategies for history matching of hydrocarbon reservoirs. In *Bayesian Statistics 5*, eds. J.M. Bernardo *et al.*, Oxford University Press, pp.69–95.

Craig, P.S., Goldstein, M., Seheult, A.H. and Smith, J.A. (1997) Pressure matching for hydrocarbon reservoirs: a case study in the use of Cayes linear strategies for large computer experiments. In *Case studies in Bayesian Statistics: Volume III*, eds. C. Gatsonis *et al.*, Springer-Verlag, pp. 36–93.

Craig, P.S., Goldstein, M., Rougier, J.C. and Seheult, A.H. (2001) Bayesian forecasting for complex systems using computer simulators. *J. Am. Stat. Assoc.*, **96**, 717–729.

Damian, D., Sampson, P. D., and Guttorp, P. (2000). Bayesian estimation of semi-parametric non-stationary spatial covariance structure. *Environmetrics* **12**,161-176.

Damian, D., Sampson, P. D., and Guttorp, P. (2003). Variance modeling for nonstationary spatial processes with temporal replication. To appear, *Journal of Geophysical Research-Atmospheres.*

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.

Fuentes, M. (2001). A 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. Tech. report at North Carolina State University, Institute of Statistics Mimeo Series #2534.

Fuentes, M. and Raftery, A. (2001). Model evaluation and spatial interpolation by combining observations with outputs from numerical models via Bayesian Melding. Technical report no. 403, Department of Statistics, University of Washington.

Goldstein, M. (1999). Bayes linear analysis. In *Encyclopedia of Statistical Sciences: Update Volume 3*, eds. S. Kotz, C. B. Read and D.L. Banks, Wiley, 29–34.

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.

Hanna, S. R. (1971). A simple method of calculating dispersion from urban area sources. *Journal of the Air Pollution Control Association,* **21**, 774-777.

Haylock, R.G. and O'Hagan A (1996) On inference for outputs of computationally expensive algorithms with uncertain tyon the inputs. In *Bayesian Statistics 5*, eds. J.M. Bernardo *et al.*, Oxford University Press, pp. 629–637.

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.

Kennedy, M.C. and O'Hagan, A (2000) Predicting the output from a complex computer code when fast approximations are available *Biometrika*, **87**, 1–13.

Kennedy, M.C. and O'Hagan, A (2001), Bayesian calibration of computer models (with discussion) *J. R. Stat. Soc. Ser. B*, **63**, 425–464

Kennedy, M., O'Hagan, A. and Higgins, N. (2002). Bayesian Analysis of Computer Code Outputs. In *Quantitative Methods for Current Environmental Issues*, C W Anderson, V Barnett, P C Chatwin, A H El-Shaarawi (editors), 227–243. Springer-Verlag.

Krzanowski, W.J. (1988) *Principles of Multivariate Analysis, a Users Perspective.* Oxford University Press.

Martin, D. O. (1971). An urban diffusion model for estimating long term average vlues of air quality. *Journal of the Air Pollution Control Association,* **21**, 16-19.

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.

Meiring, W., Guttorp, P., and Sampson, P. S. (1998). Space-time estimation of grid-cell hourly ozone levels for assessment of a deterministic model. *Environmental and Ecological Statistics*, **5**, 197-222.

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.

Oakley, J. and O'Hagan, A. (2002). Bayesian inference for the uncertainty distribution of computer model outputs. *Biometrika* **89**, 769–784.

O'Hagan (1992) Some Bayesian numerical analysis (with discussion). In *Bayesian Statistics 4*, eds J.M. Bernardo, J.O. Berger, A.P. Dawid and J.F.M Smith, Oxford University Press,345-363.

O'Hagan, A., Kennedy, M.C., and Oakley, J.E. (1998) Uncertainty analysis and other inference tools for complex computer codes (with discussion). In *Bayesian Statistics 6*, eds J.M. Bernardo, J.O. Berger, A.P. Dawid and J.F.M Smith, Oxford University Press, 503–524.

O'Hagan, A. and Haylock, R. G. (1997). Bayesian uncertainty analysis and radiological protection. In *Statistics for the Environment 3, Pollution Assessment and Control*, V. Barnett and K. F. Turkman (eds.). Wiley: Chichester. pp 109–128.

Oreskes, N. Sharader-Frechette, K., and Belitz, K. (1994). Verifcation, validation and confirmation of numerical models in the earth sciences, *Science,* **263**, 641-646.

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. (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.

Schmidt, A. M. and O'Hagan, A. (2000). Bayesian inference for nonstationary spatial covariance structure via spatial deformations. Research Report No. 498/00 Department of Probability and Statistics, University of Sheffield. Submitted and accepted by the Journal of the Royal Statistical Society, Series B.

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

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