Alma Mater Studiorum - Università di Bologna

DOTTORATO DI RICERCA IN

Metodologia Statistica per la Ricerca Scientifica

XXVI Ciclo

Settore Concorsuale di afferenza: 13/D1
Settore Scientifico disciplinare: SECS-S/01

Bayesian space-time data fusion for real-time forecasting and map uncertainty

Presentata da: Lucia Paci

Coordinatore Dottorato
Prof.ssa Angela Montanari

Relatore
Prof.ssa Daniela Cocchi
Prof. Alan E. Gelfand

Esame finale anno 2012/2013

To my Family

Abstract

Environmental computer models are deterministic models devoted to predict several environmental phenomena such as air pollution or meteorological events. Numerical model output is given in terms of averages over grid cells, usually at high spatial and temporal resolution. However, these outputs are often biased with unknown calibration and not equipped with any information about the associated uncertainty. Conversely, data collected at monitoring stations is more accurate since they essentially provide the true levels. Due the leading role played by numerical models, it now important to compare model output with observations. Statistical methods developed to combine numerical model output and station data are usually referred to as data fusion.

In this work, we first combine ozone monitoring data with ozone predictions from the Eta-CMAQ air quality model in order to forecast real-time current 8-hour average ozone level defined as the average of the previous four hours, current hour, and predictions for the next three hours. We propose a Bayesian downscaler model based on first differences with a flexible coefficient structure and an efficient computational strategy to fit model parameters. Model validation for the eastern United States shows consequential improvement of our fully inferential approach compared with the current real-time forecasting system. Furthermore, we consider the introduction of temperature data from a weather forecast model into the downscaler, showing improved real-time ozone predictions.

Finally, we introduce a hierarchical model to obtain spatially varying uncertainty associated with numerical model output. We show how we can learn about such uncertainty through suitable stochastic data fusion modeling using some external validation data. We illustrate our Bayesian model by providing the uncertainty map associated with a temperature output over the northeastern United States.

Acknowledgments

I am deeply grateful to my supervisor Prof. Daniela Cocchi for her precious advice and encouragement throughout this research work. I would like to express my sincere gratitude to my co-supervisor, Prof. Alan Gelfand for his novel ideas, insightful discussions and 'contagious' enthusiasm. Their guidance helped me in all the time of research and writing of this thesis.

I am very grateful to Dr. Dave Holland for involving me in this important EPA's project. I also would like to thank my thesis committee: Dr. Rosalba Ignaccolo, Prof. Giovanna Jona Lasinio and Prof. Fabrizio Ruggeri, for their useful comments and suggestions.

A special thanks goes to Dr. Francesca Bruno who has always encouraged me and believed in me. I also would like to thank my research group for the helpful discussions, my Ph.D. colleagues, most of all Arianna, to whom I shared this experience and Dukies' colleagues that helped me during my stay in Durham.

Last but not least, I wish to thank Davide for all his support and love.

Contents

List of Figures

List of Tables

Chapter 1

Combining monitoring data and numerical model output

Environmental computer models are playing an increasing role as tools to understand and predict complex systems. They are deterministic simulation models that mathematically approximate the underlying physical and chemical processes via nonlinear partial differential equations. These models are often implemented as computer codes and depend on a number of input parameters which determine the nature of the output. The resulting output is usually given in terms of averages over grid cells. Using a large number of grid cells, the numerical model estimates can cover large spatial domains and may also have very high temporal resolution for current, past, and future time periods.

Several environmental sciences use numerical models to predict spatio-temporal processes. Meteorological centers produce weather forecasts using numerical weather prediction models; oceanographers forecast storm surges and ocean wave fields using computer models that simulate hurricane intensity and trajectory; atmospheric scientists predict concentration for several pollutants using air quality models. Predictions from numerical models are also used for environmental regulatory purposes and improved decision making strategies.

However, numerical model output are often biased with unknown calibration. Moreover, they are not equipped by any information about the associated uncertainty since they have been derived under a deterministic paradigm. In this regard, the paper by Kennedy and O'Hagan (2001) discusses prediction and uncertainty analysis for systems approximated by mathematical models.

For large spatial regions, the spatial coverage of the available network of mon-

itoring stations can never match the coverage at which computer models produce their output. However, monitoring data represent the most accurate way to obtain information on the variable of interest since, up to measurement error, they provide the actual true levels.

Due to the social and economic consequences, it is becoming more and more important that output from numerical models are evaluated and also calibrated. To accomplish that, output from numerical models must be compared with observations. Statistical methods developed to combine numerical model output and station data are usually referred to as data fusion. As shown in the next chapters, data fusion modeling may be motivated by different goals. For instance, we would combine monitoring data and numerical model output to improve the forecasting of some environmental variables or we could be interested in quantifying the uncertainty associated with computer models output.

While numerical model predictions are given in terms of averages over grid cells, observations are collected at points in the spatial domain. The spatial misalignment between point- and grid-referenced data is an example of what, in spatial statistics, is called the *change of support problem* (COSP; see e.g., Cressie 1993; Gelfand et al. 2001; Gotway and Young 2002; Banerjee et al. 2004; Gelfand 2010), which concerns the inference of a spatial variable at a certain resolution using data with different spatial support. A brief review of this problem is given in the next section.

1.1 Change of support problem

According to Gotway and Young (2002), "One of the most challenging and fascinating areas in spatial statistics is the synthesis of spatial data collected at different spatial scales". In fact, the spatial scale is the key choice for the study of spatial processes since it affects the process dynamics; mechanisms operating at small spatial scale may be not relevant at large scales and, conversely, mechanisms operating at large spatial scale may not even be seen at fine scale. Such scaling concerns are particularly appreciated in studying human, animal and plant populations as well as environmental phenomena and investigated by researchers in many different fields.

Consider a variable that is observed either at points in space (i.e. point-referenced data) or over areal units (i.e. block data). Then, the change of support problem refers to making inference about the variable at a different spatial scale from the one at which it has been observed. COSP may result also when studying

the connections between spatial variables with different supports. As an example, we might have weather predictions at low-resolution from a global climate forecast model, and seek to predict at higher resolution. Or, we might obtain the spatial distribution of some variable at the county level, even though it was originally collected at the census tract level. Also, the change of support problem arises when the objective is the calibration of weather radar data using raingauge observations (e.g. Fuentes et al., 2008; Orasi et al., 2009; Bruno et al., 2013a).

A solution of the change of support problem is also required in many health science applications, such as spatial and environmental epidemiology. Most of this research focuses on the effect of air quality on health (Dominici et al., 2002; Zhu et al., 2003; Greco et al., 2005; Fuentes et al., 2006). In this context, exposure and response variables are often measured at different levels of spatial aggregation: disease data are often collected as counts over spatial units e.g., zip codes, counties, or census tracts, while environmental exposure is measured by monitoring networks producing point-referenced data. In other cases, data are available for both disease and environmental exposure on an aggregate scale, but on different grids; for instance when air quality data are provided by computer models.

Arbia (1989) uses the terminology *spatial data transformations* to refer to situations where the process of interest has a different spatial scale with respect to the spatial form of the observed data. These transformations are the basis of the so-called *modifiable areal unit problem* (MAUP). In this case, the purpose is to understand the distribution of the variable at a new level of spatial aggregation or perhaps relate it to another variable that is already available at this level. A special case of MAUP is the so called *ecological inference problem* (Robinson, 1950; Wakefield, 2003, 2008) which concerns the process of deducing individual behavior and relationships from aggregated data, leading to "ecological bias". The relationships observed between variables measured at the ecological (aggregate) level may not accurately reflect the relationship between the same variables measured at the individual level. This bias depends upon two components: the *aggregation bias* due to the grouping of individuals and the *specification bias* due to the fact that the distribution of confounding variables varies with grouping.

In COSP, following Gelfand (2010), we can envision four different situations (univariate setting):

1. We have observations at point-level $Y(s_i)$ at locations s_i, $i = 1, 2, \ldots, n$ and we would infer about the process at new locations s'_j, $j = 1, 2, \ldots, m$ (point-

Table 1.1: Examples of COSP.

Observed	Inference	Examples
Point-level	Point-level	Kriging
Point-level	Areal unit	Spatial smoothing, Block kriging
Area unit	Point-level	Ecological inference
Area unit	Area unit	MAUP, areal interpolation

to-point).

2. We have, as above, observations at point-level $Y(s_i)$ at locations s_i, $i = 1, 2, \ldots, n$ and we would infer about the process at a collection of areal units $Y(B_k)$ associated with B_k, $k = 1, 2, \ldots, K$ (point-to block).

3. We have, observations associated with areal units $Y(B_k)$ at areal unit B_k, $k = 1, 2, \ldots, K$ and we would infer about the process at a collection of sites s'_j, $j = 1, 2, \ldots, m$ (block-to-point).

4. We have, as above, observations associated with areal units $Y(B_k)$ at areal unit B_k, $k = 1, 2, \ldots, K$ and we would infer about the process at a collection of areal units $Y(B_{k'})$, associated with $B_{k'}$, $k' = 1, 2, \ldots, K'$ (block-to-block). Here, the $B_{k'}$'s can be nested or not within the B_k's.

The above scheme can be easily arranged to the regression setting where the COSP arises from the difference in spatial resolution between the response variable and the covariates. Some common COPS are given in Table 1.1, adapted from Gotway and Young (2002) where a comprehensive review of the literature on the change of support problem is discussed.

To study the COSP in details, we introduce the following notation. Let $\{Y(s) : s \in D \subset \mathcal{R}^d\}$ be the spatial process measured at location s in some region of interest D. Assume that $Y(s)$ has mean $E\big(Y(s)\big) = \mu(s; \boldsymbol{\beta})$ and covariance function $Cov\big(Y(s), Y(s')\big) = C(s, s'; \boldsymbol{\theta})$ for $s, s' \in D$. For block data we assume that the observations arise as block averages $Y(B)$ where

$$Y(B) = \frac{1}{|B|} \int_B Y(s) ds \tag{1.1}$$

and $|B|$ denotes the area of the block $B \subset D$. The integration (1.1) is an average of random variables, hence a stochastic integral. The assumption underlying the

spatial process is appropriate only for block data that may be viewed as an average over point data; for instance it might include pollutant level, rainfall, temperature, etc. but it would not be suitable for a population (i.e. there is no population at a given point).

The moments of $Y(B)$ can be derived from the moments of the underlying process as follows:

$$E\big(Y(B)\big)= \frac{1}{|B|}\int_B \mu(s;\boldsymbol{\beta})ds$$

$$Cov\big(Y(B),Y(B')\big)= \frac{1}{|B|}\frac{1}{|B'|}\int_B\int_{B'} C(s,s';\boldsymbol{\theta})ds\; ds'$$

Here, we focus on the point-to-block COSP which represent the most attractive situation according to our scope, as we clarify in the next chapters. The inferential problem concerns the prediction of $\mathbf{Y}_B^T = \big(Y(B_1),\ldots,Y(B_K)\big)$ from point-referenced data $\mathbf{Y}_s^T = \big(Y(s_1),\ldots,Y(s_n)\big)$ observed at a finite set of sites s_i, $i = 1,\ldots,n$. In the geostatistical framework, the solution to the point-to-block change of support problem is given by the *block kriging* (Cressie, 1993; Chiles and Delfiner, 1999) which enables predictions of $Y(B_k)$ given observations collected at points; some extensions of the block kriging have been proposed by Carroll et al. (1995) and Gotway and Young (2007).

Alternatively, Bayesian hierarchical models have been developed to address the COSP (e.g. Mugglin et al. 2000; Best et al. 2000; Gelfand et al. 2001; Wikle and Berliner 2005). Following Gelfand et al. (2001), we consider a stationary Gaussian process with mean $\mu_s(\boldsymbol{\beta})_i = \mu(s_i;\boldsymbol{\beta})$ and covariance matrix $\big(C_s(\boldsymbol{\theta})\big)_{ii'}= C(s_i - s_{i'};\boldsymbol{\theta})$ that is,

$$\mathbf{Y}_s \,|\, \boldsymbol{\beta},\boldsymbol{\theta} \sim N\big(\mu_s(\boldsymbol{\beta}),\; C_s(\boldsymbol{\theta})\big).$$

Under the Bayesian perspective, the prediction for blocks $B_1, B_2, \ldots, B_K$ are based upon the predictive distribution $f\big(\mathbf{Y}_B \,|\, \mathbf{Y}_s,\boldsymbol{\beta},\boldsymbol{\theta}\big)$ given by

$$N\Big(\mu_B(\boldsymbol{\beta}) + C_{s,B}^T(\boldsymbol{\theta})C_s^{-1}(\boldsymbol{\theta})\big(\mathbf{Y}_s - \mu_s(\boldsymbol{\beta})\big),\; C_B(\boldsymbol{\theta}) - C_{s,B}^T(\boldsymbol{\theta})C_s^{-1}(\boldsymbol{\theta})C_{s,B}(\boldsymbol{\theta})\Big) \quad (1.2)$$

where

$$\mu_B(\boldsymbol{\beta})_k = \frac{1}{|B_k|}\int_{B_k} \mu(s;\boldsymbol{\beta})ds$$

$$\big(C_B(\boldsymbol{\theta})\big)_{kk'}= \frac{1}{|B_k|}\frac{1}{|B_{k'}}\int_{B_k}\int_{B'_k} C(s-s';\boldsymbol{\theta})\; ds\; ds'$$

and the matrix $C_{s,B}(\boldsymbol{\theta})$ contains the point-to-block covariance given by

$$Cov\big(Y(B_k),Y(s_i)\big)= \frac{1}{|B_k|}\int_{B_k} C(s_i - s';\boldsymbol{\theta})ds' \;.$$

All the entries in (1.2) requires a stochastic integration; in practice integrals are computed by approximations such as Monte Carlo integration.

We recall that the COSP arises from the combination of monitoring data observed at point-level with numerical model output expressed as averages over grid cells. Such synthesis is referred in the literature to as *data fusion*.

1.2 Data fusion modeling

Data assimilation, or data fusion, refers to the statistical techniques used to combine numerical models with observations to give an improved estimate of the state of a system or process (Nychka and Anderson, 2010). Recently, many papers have been published on data fusion methods for combining observed data and computer model output (see Gelfand and Sahu (2010) and references therein). In this section we review the main fully model-based approaches proposed to address the data fusion problem.

In air quality context, Meiring et al. (1998) propose to predict hourly ozone concentrations on grid cell scale by using the observations at monitoring site in order to compare these predictions with those provided by a numerical model at the grid cell level. A different strategy has been proposed by Jun and Stein (2004) who ignore the difference in spatial resolution between model output and observations, rather suggesting to evaluate a numerical model by looking for differences between the model output and the observations in terms of variograms and correlograms.

Wikle and Berliner (2005) presented a hierarchical Bayesian model to combine data observed at different spatial scales. Their approach is based on the idea of conditioning a "true" unobserved spatially continuous process on a areal average of the process at some resolution; then, also the data are conditioned to this areal averaged true process. In a similar fashion, Fuentes and Raftery (2005) proposed a Bayesian model to fuse air pollution measurements and predictions from an air quality model. Working with block averaging as in (1.1), the model could be considered an instance of Bayesian Melding (Poole and Raftery, 2000) and builds upon earlier works of Cowles et al. (2002) and Cowles and Zimmerman (2003). Fuentes and Raftery (2005) assumed that there exists an underlying point-level spatial process driving both the monitoring data and the numerical model output. In particular, observations are linked to the latent spatial process via a measurement error model while the numerical model output is expressed in terms of stochastic integrals over grid cells of the underlying process, also accounting for

potential bias in the output. This approach has gained considerable popularity and used in several applications (Swall and Davis, 2006; Smith and Cowles, 2007; Fuentes and Foley, 2008; Liu et al., 2011). However, despite its popularity, the Bayesian melding fusion strategy suffers from some limitations. First, it becomes computationally infeasible for fusing a very large number of grid cells (and usually a sparse number of monitoring sites); in fact, a huge amount of stochastic integrals needs to be computed. Secondly, it does not consider the temporal dimension and the implementation of the dynamic extension becomes even more infeasible.

A Bayesian space-time data fusion model has been proposed by McMillan et al. (2010) for combining output from a numerical model and measurements of fine particulate matter from the U.S. monitoring network. As in Fuentes and Raftery (2005), both sources of data are driven by an underlying "true" process; however, rather than assuming the latent process at the point-level, McMillan et al. (2010) specified the "true" process at the grid cell level. In this fashion, the model offers a solution to the upscaling problem and the computation is simplified because it avoids the computationally demanding stochastic integrations required in Fuentes and Raftery (2005).

A different solution to the data fusion problem uses two-stage regression models; building upon the work of Guillas et al. (2008) and subsequently Zidek et al. (2012), this approach enables to calibrate the numerical model output by downscaling the predictions from grid cells to point level and comparing them with observations. Within this fashion, Berrocal et al. (2010b) introduce the downscaler model that we review in the next section.

1.3 Downscaler models: a review

Recently, an innovative solution to the COSP has been provided by the so called *downscaler* model introduced by Berrocal et al. (2010b). Rather than assuming the existence of a latent process driving both the observations and the numerical model output, Berrocal et al. (2010b) take the numerical model output as explanatory variable and relate observations and numerical model output using a regression model with spatially varying coefficients (Gelfand et al., 2003) in turn, modeled as Gaussian processes. The authors address the difference in spatial scale between the two sources of data for bias-correcting the predictions generated by the numerical model. The downscaler is simple, very flexible, computationally feasible and allows straightforward prediction to point level. Thus, it offers a fully model-based

solution to the problem of downscaling.

In the next sub-sections we review the spatial static version of different downscaler models, specified within the Bayesian framework. Each of these model can be also extended to handle spatio-temporal data, as we illustrate in Chapter 2.

1.3.1 Univariate downscaler

In this section we review the univariate spatial downscaler presented in Berrocal et al. (2010b). Let $Y(\mathbf{s})$ denote the observed concentration of a pollutant at a generic location $\mathbf{s}$ and $W(B)$ be the output from a numerical model at the grid cell B. The downscaler model addresses the difference in spatial resolution between monitoring data and numerical model output, by associating to each site $\mathbf{s}$ the grid cell B that contains $\mathbf{s}$. So, all the points $\mathbf{s}$ falling in the same grid cell are assigned to the same numerical model output value.

Then, the model links the observational data and the numerical model output as follows:

$$Y(\mathbf{s}) = \tilde{\beta}_0(\mathbf{s}) + \tilde{\beta}_1(\mathbf{s})W(B) + \epsilon(\mathbf{s}) \tag{1.3}$$

where

$$\begin{aligned} \tilde{\beta}_0(\mathbf{s}) &= \beta_0 + \beta_0(\mathbf{s}) \\ \tilde{\beta}_1(\mathbf{s}) &= \beta_1 + \beta_1(\mathbf{s}) \end{aligned} \tag{1.4}$$

and $\epsilon(\mathbf{s})$ is a white noise process with nugget variance τ^2. The spatially varying coefficients $\tilde{\beta}_0(\mathbf{s})$ and $\tilde{\beta}_1(\mathbf{s})$ can be interpreted as a random intercept process and a random slope process, respectively. Equivalently, $\beta_0(\mathbf{s})$ and $\beta_1(\mathbf{s})$ can be viewed as local spatial adjustments to the overall additive bias β_0 and global multiplicative bias β_1.

In order to introduce association between $\beta_0(\mathbf{s})$ and $\beta_1(\mathbf{s})$, the two spatially varying coefficients are in turn modeled as correlated mean-zero Gaussian spatial processes using the method of coregionalization[1] (Wackernagel, 2003; Gelfand et al., 2004). Therefore, they are modeled as a linear combination of two latent zero-mean unit-variance independent Gaussian processes $w_0(\mathbf{s})$ and $w_1(\mathbf{s})$ each equipped with an exponential covariance structure[2] having decay parameters, respectively, ϕ_0 and

[1] The term 'coregionalization' is intended to denote a model for measurements that co-vary over a region.

[2] Exponential correlation function: $cov(w_j(\mathbf{s}), w_j(\mathbf{s}')) = exp(-\phi_j \|\mathbf{s} - \mathbf{s}'\|)$, having $j = 0, 1$.

ϕ_1, such that

$$\begin{pmatrix} \beta_0(\mathbf{s}) \\ \beta_1(\mathbf{s}) \end{pmatrix} = \mathbf{A} \begin{pmatrix} w_0(\mathbf{s}) \\ w_1(\mathbf{s}) \end{pmatrix} \tag{1.5}$$

where the unknown $\mathbf{A}$ is the coregionalization matrix, usually assumed to be lower-triangular. Matrix $\mathbf{A}$ in (1.5) determines the correlation between the two spatially-varying coefficients $\beta_0(\mathbf{s})$ and $\beta_1(\mathbf{s})$ and thus on the covariance structure of $Y(\mathbf{s})$. The downscaler is specified under the Bayesian perspective and the prior distributions for all unknown model parameters complete the Bayesian hierarchy of the model.

Via a simulation study and a series of experiments carried out with ozone concentration data for 2001, Berrocal et al. (2010b) show that the downscaler outperforms Bayesian melding and ordinary kriging both in terms of computing speed and predictive performance: predictions obtained with the downscaler are better calibrated i.e. lower predictive mean square and absolute value errors and predictive intervals have empirical coverage closer to the nominal values.

1.3.2 Multivariate downscaler

Berrocal et al. (2010a) extend the downscaler model from the univariate setting to a bivariate setting in order to fuse ozone and fine particulate matter ($PM_{2.5}$) concentrations with the output of an air quality numerical model. The bivariate downscaler exploits the correlation between the observed levels of the pollutants and in the output provided by the numerical model. Moreover, the model enables to handle not only the spatial misalignment between monitoring data and model output, but also it allows to accommodate the spatial misalignment between the ozone and $PM_{2.5}$ monitoring data.

We illustrate the general multivariate downscaler in the static formulation of the model. Let $Y_i(\mathbf{s})$, $i = 1, \ldots, p$ be the observed data of the i-th variable at a site $\mathbf{s}$ and $W_i(B)$, $i = 1, \ldots, p$, be the numerical model output of the i-th variable over the grid cell B. Again, each site $\mathbf{s}$ is associated to the grid cell B in which $\mathbf{s}$ lies; then, the observational data and the numerical model output are linked via the model

$$Y_i(\mathbf{s}) = \tilde{\beta}_{i0}(\mathbf{s}) + \sum_{i=1}^{p} \tilde{\beta}_{ij}(\mathbf{s}) W_j(B) + \epsilon_i(\mathbf{s}) \tag{1.6}$$

where $\epsilon_i(\mathbf{s})$ are IID $N(0, \tau_i^2)$. As in the univariate case, each of the $p(p+1)$ terms

$\tilde{\beta}_{ij}$, $i = 1, \ldots, p$, $j = 0, \ldots, p$ is decomposed in the sum of an overall term and a random local adjustment, that is:

$$\tilde{\beta}_{ij}(\mathbf{s}) = \beta_{ij} + \beta_{ij}(\mathbf{s}) \tag{1.7}$$

where the $\beta_{ij}(\mathbf{s})$ are in turn modeled as correlated zero-mean Gaussian processes using the method of coregionalization. Therefore, the spatially varying coefficients $\beta_{ij}(\mathbf{s})$ are expressed as a linear combination of zero-mean unit-variance independent Gaussian processes $w_{ij}(\mathbf{s})$, each equipped with an exponential covariance structure such that

$$\begin{pmatrix} \beta_{10}(\mathbf{s}) \\ \ldots \\ \beta_{ij}(\mathbf{s}) \\ \ldots \\ \beta_{pp}(\mathbf{s}) \end{pmatrix} = \mathbf{A} \begin{pmatrix} w_{10}(\mathbf{s}) \\ \ldots \\ w_{ij}(\mathbf{s}) \\ \ldots \\ w_{pp}(\mathbf{s}) \end{pmatrix} \tag{1.8}$$

The coregionalization matrix $\mathbf{A}$ is a $p(p+1) \times p(p+1)$ matrix that is assumed, without loss of generality, to be lower-triangular. Matrix $\mathbf{A}$ determines the correlation between the spatially varying coefficients $\beta_{ij}(\mathbf{s})$ and also induces a correlation structure on the multivariate random vector $\mathbf{Y} = \{Y_i(\mathbf{s})_{i=1,\ldots,p}, \mathbf{s} \in S\}$. So, simplifications of model (1.6) - (1.8) can be considered by setting to zero some entries of the matrix $\mathbf{A}$. Again, the multivariate downscaler arises as a Bayesian hierarchical formulation and is completely specified by the prior distribution for all unknown parameters. Finally, the multivariate downscaler model can be extended to accommodate data collected also over time.

The empirical study in Berrocal et al. (2010a) shows that the bivariate downscaler leads predictions of both ozone and particulate matter levels more accurate than those obtained using the univariate downscaler which does not account for the association between the two pollutants.

1.3.3 Smoothed univariate downscaler

In Berrocal et al. (2012) two neighbor-based extensions of the univariate downscaler model have been proposed. The first extension introduces a Gaussian Markov random field (GMRF) to smooth the computer model output, while the second extension introduces spatially varying weights driven by a latent Gaussian process.

First, the authors smooth the numerical model output, $W(B)$, via the model

$$W(B) = \widetilde{W}_1(B) + \eta(B)$$

where $\eta(B) \sim N(0, \sigma^2)$ and $\widetilde{W}_1(B) = \mu + x(B)$ represents a smoothed version of $W(B)$. Here, $x(B)$ is a zero-mean Gaussian Markov random field equipped with a conditionally autoregressive (CAR) structure (Besag, 1974; Banerjee et al., 2004). Then, for $\mathbf{s} \in B$, model (1.3) is replaced by

$$Y(\mathbf{s}) = \tilde{\beta}_0(\mathbf{s}) + \beta_1 \widetilde{W}_1(B) + \epsilon(\mathbf{s})$$

where $\tilde{\beta}_0(\mathbf{s})$ and $\epsilon(\mathbf{s})$ are defined as above.

The second extension proposed by Berrocal et al. (2012) introduces smoothing via random spatially varying weights. In particular, the monitoring data are linked to a new point-level variable $\widetilde{W}_2(B)$ obtained, at each site $\mathbf{s}$, as weighted average of all the numerical model output. Hence, model (1.3) is modified as

$$Y(\mathbf{s}) = \tilde{\beta}_0(\mathbf{s}) + \beta_1 \widetilde{W}_2(\mathbf{s}) + \epsilon(\mathbf{s})$$

where $\tilde{\beta}_0(\mathbf{s})$ and $\epsilon(\mathbf{s})$ are defined as above and

$$\widetilde{W}_2(\mathbf{s}) = \sum_{k=1}^{g} \omega_k(\mathbf{s})\, W(B_k)$$

where g is the number of numerical model grid cells. The spatially varying weights $\omega_k(\mathbf{s})$ at each site $\mathbf{s}$, with $k = 1, \ldots, g$ are defined by

$$\omega_k(\mathbf{s}) = \frac{\mathcal{K}\big(\mathbf{s} - \mathbf{r}_k; \psi\big)\exp\big(\mathcal{Q}(\mathbf{r}_k)\big)}{\sum_{l=1}^{g} \mathcal{K}\big(\mathbf{s} - \mathbf{r}_l; \psi\big)\exp\big(\mathcal{Q}(\mathbf{r}_l)\big)}$$

where $\mathbf{r}_k$ are the centroids of the grid cells, $\mathcal{Q}(\mathbf{r}_l)$ is a zero-mean Gaussian process having exponential covariance function and $\mathcal{K}\big(\mathbf{s} - \mathbf{r}_l; \psi\big)$ is an exponential kernel with decay parameter ψ.

The authors apply their approach to predict daily ozone levels for the eastern United States during the summer 2001 using station data and the Community Multiscale Air Quality (CMAQ) model output. The results of both methods show, respectively, a 5% and a 15% predictive gain in overall predictive mean square error over the univariate downscaler model described in Section 1.3.1.

Currently, the space-time smoothed downscaler with spatially varying random weights is used by the U.S. Environmental Protection Agency (USEPA) to fuse daily ozone (8-hour max) and fine particulate air (24-hour average) monitoring data from the National Air Monitoring Stations/State and Local Air Monitoring Stations (NAMS/SLAMS) with 12 km gridded output from the CMAQ model. Daily predictions are available at the 2000 Census Tract centroid locations over the eastern U.S.; see Heaton et al. (2012) for details.

1.3.4 Downscaler with point masses

A further version of the downscaler model appears in Sahu et al. (2010), where observed point-referenced monitoring data and gridded output from the CMAQ numerical model are combined to provide accurate spatial interpolation and temporal aggregation of weekly wet chemical deposition in the eastern United States. The authors use precipitation information to model wet deposition since no deposition exists without precipitation. Moreover, their modeling for monitoring stations allows to accommodate point masses at 0 for both precipitation and wet deposition. First, they model precipitation and then, deposition given precipitation, considering the spatial misalignment. Both precipitation and wet deposition are driven by a point-referenced latent space-time atmospheric process. Similarly, the computer model output also supplies 0 values for wet deposition in some grid cells. To capture these point masses at 0, the authors introduce a latent process at the grid scale modeled through a conditionally autoregressive (CAR) specification. Then, the downscaling connects the point-level process to the grid-scale process using a measurement error model.

Here, we introduce some details of the static model developed by Sahu et al. (2010). Let $P(\mathbf{s})$ and $Z(\mathbf{s})$ denote the observed precipitation and deposition respectively at site $\mathbf{s}$. Both $P(\mathbf{s})$ and $Z(\mathbf{s})$ are driven by the latent process $V(\mathbf{s})$ as follows:

$$P(\mathbf{s}) = \begin{cases} \exp\{U(\mathbf{s})\} & \text{if } V(\mathbf{s}) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{1.9}$$

and

$$Z(\mathbf{s}) = \begin{cases} \exp\{Y(\mathbf{s})\} & \text{if } V(\mathbf{s}) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{1.10}$$

The random variables $U(\mathbf{s})$ and $Y(\mathbf{s})$ represent the log observed precipitation and deposition respectively when $V(\mathbf{s}) > 0$. Similarly, for the CMAQ output at grid cell B, $Q(B)$, we have:

$$Q(B) = \begin{cases} \exp\{X(B)\} & \text{if } \widetilde{V}(B) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{1.11}$$

leading to positive numerical model output when the areal level latent variable $\widetilde{V}(B)$ is positive. The likelihood at the first stage is derived from definitions (1.9)-(1.11) and is given by:

$$f(\mathbf{P}, \mathbf{Z}, \mathbf{Q} \mid \mathbf{U}, \mathbf{Y}, \mathbf{X}, \mathbf{V}, \widetilde{\mathbf{V}}) = f(\mathbf{P} \mid \mathbf{U}, \mathbf{V}) f(\mathbf{Z} \mid \mathbf{Y}, \mathbf{V}) f(\mathbf{Q} \mid \mathbf{X}, \widetilde{\mathbf{V}})$$

where $\mathbf{P}, \mathbf{Z}, \mathbf{Q}$ collect, respectively, all the precipitation values, deposition values and the CMAQ output while $\mathbf{U}, \mathbf{Y}, \mathbf{X}, \mathbf{V}, \widetilde{\mathbf{V}}$ denote the vectors corresponding to the random variables.

In the second stage the models for the latent variables are defined. First, the authors specified a spatial regression for the log-precipitation based on the latent process $V(\mathbf{s})$, that is

$$U(\mathbf{s}) = \alpha_0 + \alpha_1 V(\mathbf{s}) + \varepsilon(\mathbf{s})$$

where $\boldsymbol{\varepsilon} = \big(\varepsilon(\mathbf{s_1}), \ldots, \varepsilon(\mathbf{s_n})\big)$ is a Gaussian process equipped with an exponential correlation function.

Second, for each $\mathbf{s}$ in B, the model for the log-deposition is given by:

$$Y(\mathbf{s}) = \tilde{\beta}_0 + \tilde{\beta}_1 X(B) + \beta_2 U(\mathbf{s}) + \beta_3 V(\mathbf{s}) + \eta(\mathbf{s})$$

where $\tilde{\beta}_0$ and $\tilde{\beta}_1$ are independent and defined as in (1.4) and $\eta(\mathbf{s}) \sim N(0, \sigma_\eta^2)$.

Then, the CMAQ output $X(B)$ is modeled using the latent process $\widetilde{V}(B)$ as follows:

$$X(B) = \gamma_0 + \gamma_1 \widetilde{V}(B) + \delta(B)$$

where $\delta(B) \sim N(0, \sigma_\delta^2)$. The latent process $\widetilde{V}(B)$ is assumed to follow a CAR process in space.

Finally, the spatial misalignment between the observation and the numerical model output are addressed via a measurement error model, that is: $V(\mathbf{s}) \sim N(\widetilde{V}(B), \sigma_v^2)$. The Bayesian hierarchy is completed with non-informative prior distributions.

The space-time model is fitted on weekly wet chemical deposition data both for the sulfate and nitrate compounds covering the eastern United States. The comparison of the prediction of wet deposition obtained from the model above with the current system based on Inverse Distance Weighting (IDW) shows a remarkable reduction in mean-squared error calculated over validation sites.

1.4 Overview

In this chapter we have reviewed the main literature for combining monitoring data and numerical model output. We have discussed the change of support problem and detailed the downscaler approach. Our purpose was to introduce the reader to the next chapters where Bayesian data fusion models are proposed to achieve different goals.

Our first objective is to improve real-time forecasting of *current* 8-hour average ozone levels on the scale of the entire United States (U.S.). *Current* 8-hour ozone is defined as the average of the previous four hours, current hour, and *predictions* for the next three hours. In Chapter 2, we combine first differences of ozone monitoring data and air quality numerical model output via a regression model having space-time varying coefficients in order to forecast current 8-hour average ozone exposure in real-time. We propose an hybrid strategy blending offline model fitting with online predictions and interpolation to obtain ozone forecast maps within the real-time environment. We illustrate our strategy by modeling and forecasting ozone level for a large subregion of U.S. showing that our approach outperforms the current forecasting system.

In Chapter 3, a further version of our earlier real-time downscaler will be discussed. The model regresses ozone monitoring data on real-time temperature data arising as output from a weather computer model. Again, we exploit first differences to expedite computation. Model validation for the eastern U.S. shows how we can improve the forecasting of current 8-hour average ozone by downscaling temperature data.

Finally, in Chapter 4, we propose a Bayesian hierarchical model following the approach proposed by Ghosh et al. (2012) to attach uncertainty to deterministic spatial maps. As we already noted, numerical models output are not equipped with any measure of uncertainty since they are derived under the deterministic paradigm. We develop a Bayesian data fusion model to assess the uncertainty associated with forecast maps from a numerical model using external observed point-level data.

Chapter 2

Spatio-temporal modeling for real-time ozone forecasting

The evaluation and control of air pollution levels are fundamental environmental issues for environmental decision-makers. Tropospheric, or ground level ozone is one key air pollutant as defined and regulated in the United States (U.S.). A practical challenge facing the U.S. Environmental Protection Agency (USEPA) is to provide real-time forecasting of current 8-hour average ozone defined as the average of the previous four hours, current hour, and predictions for the next three hours. Such real-time forecasting is now provided as spatial forecast maps over the entire conterminous U. S. by the EPA-AIRNow web site (http://www.airnow.gov). The capability to provide real-time air quality information is important to protect public health. For many individuals, children, outdoor workers, and those who suffer from respiratory or cardiac problems knowing the quality of the air they breathe can affect their lives and their daily activities.

Here, we illustrate the spatio-temporal data fusion model for real-time ozone forecasting proposed by Paci et al. (2013). The contribution of this work is to show how we can substantially improve upon current real-time ozone forecasting systems. We introduce a Bayesian downscaler fusion model based on first differences of real-time ozone monitoring data and numerical model output. The model has a flexible coefficient structure with an efficient computational strategy to fit model parameters. This strategy can be viewed as hybrid in that it blends offline model fitting with online predictions followed by fast spatial interpolation to produce the desired real-time forecast maps. Moreover, the strategy provides uncertainty assessment associated with these predictions. Model validation for the eastern U.S.

shows consequential improvement of our fully inferential approach compared with the existing implementations.

The chapter is organized as follows. In Section 2.1 we describe the main features of the tropospheric ozone pollution. Section 2.2 provides the details of the AIRNow system and the current ozone forecasting. In Section 2.3 we present our strategy to produce real-time 8-hour average ozone forecasts. We also discuss model fitting, with computational details deferred to Appendix A. The prediction method is developed in Section 2.4. Model validation for the eastern U.S. is given in Section 2.5. Section 2.6 gives the detail on the feasibility of our method for real-time use. Finally, Section 2.7 presents a brief summary of the chapter.

2.1 Ground level ozone

Ozone (Cocchi and Trivisano, 2013) is a reactive oxidant that occurs in two parts of the Earth's atmosphere: the stratosphere (the layer between 20-30 km above the Earth's surface) and the troposphere (ground level to 15 km). Stratospheric ozone, also known as "the ozone layer", is formed naturally and shields life from the sun's harmful ultraviolet rays. Conversely, near the earth's surface, ground-level ozone can be harmful to human health and vegetation.

Ground level ozone is not emitted directly into the air but is produced as secondary pollutant by chemical reactions between oxides of nitrogen (NOx) and volatile organic compounds (VOC). Ozone, in high concentrations, is a toxic gas that can damage pulmonary tissues. People with lung disease, children, older adults, and people who are active outdoors may be particularly sensitive to ozone. Ozone also affects sensitive vegetation and ecosystems, including forests, parks, wildlife refuges and wilderness areas. Emissions from industrial facilities and electric utilities, motor vehicle exhaust, gasoline vapors, and chemical solvents are some of the major sources of NOx and VOC.

Meteorological factors such as solar radiation, wind speed, temperature, and pressure influence directly the photochemical reactions that produce ozone. In particular, solar radiation enters into the main reactions determining ozone and wind speed promotes transport and accumulation of primary pollutants. The temperature affects directly the kinetics of reactions determining ozone and produces the mixing height, which influences the accumulation of the other chemical pollutants. Major episodes of high concentrations of ozone were most likely in the presence of weak, slow-moving, persistent high-pressure systems. Due to the strong depen-

dence on meteorological conditions, ozone levels are highly seasonal. The ozone annual behavior is characterized by higher values in summer and minimum values in winter. Also, a diurnal cycle is present since a peak in concentration occurs in the early afternoon. In urban areas, ozone levels decrease during the night while in rural areas, concentrations are stationary due to the absence of NOx sources.

In the last few decades, the phenomenon of ozone pollution has been analyzed extensively. Researchers proposed, besides chemistry transport and meteorological deterministic models, statistical models for the analysis of ozone data. Statistical analysis of ozone data is motivated by the need to summarize large amounts of data collected in time and space, to account for confounders and to evaluate uncertainties due to measurement errors. Space-time modeling of ground level ozone has received much recent attention in the literature; Cox and Chu (1993) used a generalized linear model to estimate site specific trends in daily maximum ozone levels. Guttorp et al. (1994) developed models for the space-time correlation structure that enable to spatially interpolate ozone data in a moderately homogeneous region. Bruno et al. (2009), instead of assuming the traditional spatio-temporal stationarity and the separability of spatial and temporal components, proposed a model with nonseparable structures arising from nonstationarity due to time. Nychka et al. (2002) described a multiresolution (wavelet) approach to produce nonstationary spatial covariance functions for daily average surface ozone level. Wavelets are also used to model high-frequency ozone concentration as, for instance, in Katul et al. (2006). Ignaccolo et al. (2008) developed a two-stage procedure to classify ozone monitoring stations using functional cluster analysis where Partitioning Around Medoids algorithm is embedded.

Hierarchical Bayesian approaches for spatial prediction of air pollution have also been developed; see, e.g. Wikle (2003); Huerta et al. (2004); McMillan et al. (2005) and references therein. Sahu et al. (2007) proposed a very flexible model which detects long-term trends, handles the problem of misalignment between ozone and meteorological data, and allows the calculation of summaries coherent with regulatory standards both at the local and the global scale. Dou et al. (2010) introduced complex Bayesian space-time models for hourly ozone concentration fields. Sahu and Bakar (2012) compared the dynamic linear model (see Stroud et al., 2001) with a hierarchical version of the auto-regressive model for daily maximum 8-hour average ozone concentration data. Bruno et al. (2013b) proposed hierarchical spatio-temporal model to account for differences between ozone background monitoring stations and traffic sites.

2.2 AIRNow system and ozone forecasting

Accurate assessment of exposure to ambient ozone concentrations is important for informing the public and pollution monitoring agencies about ozone levels that may lead to adverse health effects. The United States Environmental Protection Agency developed the AIRNow web site to provide the public, air regulatory agencies and health scientists with easy access to real-time national air pollution information. Current and next day forecasts of ozone and fine particulate matter are produced at over 300 cities across the United States on a daily basis. For ozone, forecasts at these monitoring sites are then interpolated across the continent, at a chosen spatial scale, to provide forecast maps for current 8-hour average ozone levels and next day patterns of 8-hour maximum ozone concentration. We focus here on current 8-hour average patterns which are updated hourly throughout the day on the AIRNow web site in the form of point estimates with no uncertainties provided. Here, current 8-hour ozone is defined as the average of the previous four hours, current hour, and predictions for the next three hours. Current patterns are updated hourly throughout the day on the EPA-AIRNow web site.

Measurements at monitoring stations present the most direct and accurate way to obtain air quality information. However, monitoring sites are often sparsely and irregularly spaced over large areas and affected by missingness. These data are the sole data source used to develop the AIRNow forecasts. However, a second source of real-time spatial information is available that could be used to improved forecasting. A numerical atmospheric model known as the Eta-Community Multi-Scale Air Quality (CMAQ) model (Yu et al., 2010) is used by EPA to simultaneously estimate multiple air pollutants (http://www.epa.gov/asmdnerl/CMAQ). Using emission inventories, meteorological information and chemical modeling components, Eta-CMAQ provides predictions of average pollution concentrations at 12 km grid cell resolution for successive time periods including 48 hours into the future. At this resolution, we have hourly numerical model information for approximately $54,000$ grid cells spanning the conterminous U.S.. However, these predictions are expected to be biased with unknown calibration.

Thus, it is important to develop computationally efficient models to combine air monitoring data and numerical model output to improve air pollution forecasting. Sahu et al. (2009a) proposed a Bayesian spatio-temporal model applied to hourly ozone concentrations. They used data over a running window of seven days to predict 8-hour average ozone level for the current hour. To allow real-time

hourly forecasting, they developed a spatial regression model that avoids iterative algorithms such as Markov Chain Monte Carlo (MCMC) methods. In Sahu et al. (2009b), a dynamic model is developed for forecasting next day 8-hour maximum ozone patterns. However, the dynamic model is computationally intensive and not feasible for use in real-time forecast applications. Kang et al. (2008) consider Kalman-filter approaches to improve next day forecasts of ozone concentrations at individual U.S. monitoring sites for the summer of 2005.

We develop a new space-time data assimilation strategy to enable use of both data sources to provide the forecasts of current 8-hour average ozone level in *real-time*. Data from the real-time ozone monitoring network and the output from the Eta-CMAQ computer model are combined, using first differences along with a regression model having spatio-temporally varying coefficients. We propose a combination of offline fitting, post-model fitting prediction, and fast online interpolation using an available kriging package to enable feasible real-time forecasting.

2.2.1 Data description

To evaluate the accuracy of the forecasts, we use historical data from a large conterminous subregion of U.S. to show that our overall approach validates well and provides significant improvement in the accuracy of forecasting relative to that of AIRNow. The first source of data we use consists of current 8-hour average ozone concentrations in parts per billion units (ppb) collected at 717 real-time monitoring stations operating in the eastern U.S. during a two-week period over August 1-14, 2011; see Figure 2.1. The region used in our application covers roughly half the conterminous U.S. and the monitoring sites farthest apart are about 2860 km from each other. We set aside data from 70 monitoring sites for validation purposes; these sites were chosen at random (again, see Figure 2.1).

The second source of data is the numerical output of the Eta-CMAQ model. This model uses meteorological information, emission inventories, and land usage to estimate average pollution levels for gridded cells (at 12 km^2 resolution) over successive time periods without any missing values. In practice, real-time hourly output from the Eta-CMAQ model is available up to 48 hours in the future. Figure 2.2 shows the Eta-CMAQ predictions of the current 8-hour average ozone level in the eastern U.S. at 12PM on August 8th, 2011. There are 21,109 Eta-CMAQ grid cells spanning our study region.

Figure 2.1: Ozone monitoring sites in the eastern U.S.. Dots and crosses represent data and validation sites, respectively.

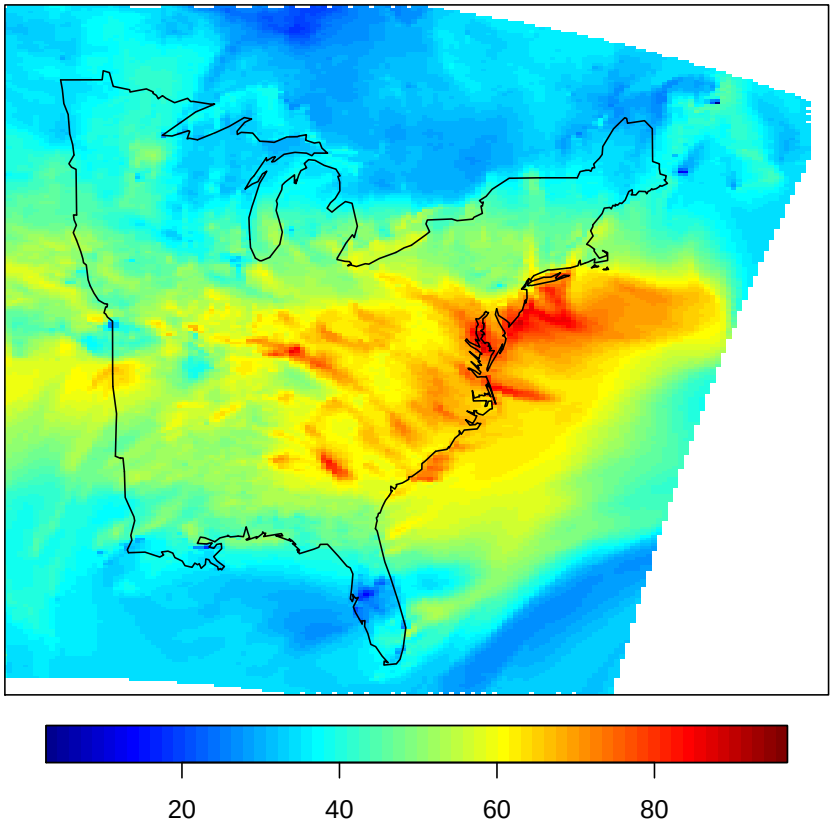

Figure 2.2: Eta-CMAQ predictions of the current 8-hour average ozone level in the eastern U.S. on August 8th, 2011.

2.3 Modeling

The spatial downscaler introduced by Berrocal et al. (2010b) has been illustrated in Section 1.3.1. In this section, we briefly review the univariate downscaler for spatio-temporal data and we propose the model for current 8-hour average ozone concentration. Then, we present our strategy to obtain real-time and accurate predictions within the real-time environment.

2.3.1 Downscaler for 8-hour average ozone level

Recall the downscaler in (1.3) - (1.4). The model can be extended to accommodate data collected over time as follows. Let $Y_t(\mathbf{s})$ denote the ozone concentration at a generic location $\mathbf{s}$ for the hour t and $W_t(B)$ be the hourly Eta-CMAQ output over grid cell B. Again, the downscaler addresses the difference in spatial resolution between monitoring data and numerical model output, by associating to each site $\mathbf{s}$ the grid cell B that contains $\mathbf{s}$. Then, the model links the observational data and the Eta-CMAQ output via a regression model with spatio-temporally varying coefficients, that is:

$$Y_t(\mathbf{s}) = \tilde{\beta}_{0,t}(\mathbf{s}) + \tilde{\beta}_{1,t}(\mathbf{s})W_t(B) + \epsilon_t(\mathbf{s}) \tag{2.1}$$

where

$$\begin{aligned} \tilde{\beta}_{0,t}(\mathbf{s}) &= \beta_0 + \beta_{0,t}(\mathbf{s}) \\ \tilde{\beta}_{1,t}(\mathbf{s}) &= \beta_1 + \beta_{1,t}(\mathbf{s}) \end{aligned} \tag{2.2}$$

and $\epsilon_t(\mathbf{s})$ is a white noise process with τ^2 as the nugget variance[1]. Coefficients $\tilde{\beta}_{0,t}(\mathbf{s})$ and $\tilde{\beta}_{1,t}(\mathbf{s})$ can be interpreted as a spatio-temporal intercept process and a spatio-temporal slope process, respectively. Equivalently, $\beta_{0,t}(\mathbf{s})$ and $\beta_{1,t}(\mathbf{s})$ in (2.2) can be viewed as local spatio-temporal adjustment to the overall intercept β_0 and global slope β_1.

Now, consider the current 8-hour average ozone level $Z_t(\mathbf{s})$ defined, from above, as the average of the previous four hours, the current hour and the next three hours in the future, that is

$$Z_t(\mathbf{s}) = \frac{1}{8}\sum_{k=-4}^{+3} Y_{t+k}(\mathbf{s}) \tag{2.3}$$

[1]In principle, other explanatory variables, such as real-time temperature or elevation, could be added to the downscaler model. Moreover, these variables can be at areal or point scale. We defer the discussion to Chapter 3.

According to the definition in (2.3), under the model in (2.1)-(2.2), the downscaler model for $Z_t(\mathbf{s})$ is given by:

$$Z_t(\mathbf{s}) = \frac{1}{8}\sum_{k=-4}^{+3}\tilde{\beta}_{0,t+k}(\mathbf{s}) + \frac{1}{8}\sum_{k=-4}^{+3}\tilde{\beta}_{1,t+k}(\mathbf{s})W_{t+k}(B) + \frac{1}{8}\sum_{k=-4}^{+3}\epsilon_{t+k}(\mathbf{s}) \tag{2.4}$$

Hourly modeling to obtain real-time prediction of the 8-hour averages, $Z_t(\mathbf{s})$, in (2.4) is infeasible. Furthermore, model fitting based upon modeling the $Z_t(\mathbf{s})$ will also be not feasible within a real-time environment. The induced dependence structure in the $Z_t(\mathbf{s})$ process will become very messy and intractable for fast model fitting; consider, for example, the induced association between $Z_t(\mathbf{s})$ and $Z_{t-1}(\mathbf{s}')$. However, if we work with differences we can simplify the specifications and can still capture the ozone diurnal variation, the influence of the Eta-CMAQ output, and the space-time random variation. Moreover, less uncertainty is associated to the predictions when modeling monitoring data differences compared with modeling the hourly ozone concentrations and converting to the $Z_t(\mathbf{s})$'s, as we show below.

2.3.2 Downscaler for monitoring data differences

Denote the monitoring data differences $\Delta_t^Z(\mathbf{s})$ by

$$\Delta_t^Z(\mathbf{s}) = 8(Z_t(\mathbf{s}) - Z_{t-1}(\mathbf{s})). \tag{2.5}$$

First differences are a commonly-used tool in time series analysis settings and motivate the introduction of $\Delta_t^Z(\mathbf{s})$. The spatial time series of first differences in (2.5) is more stable than the original series and enables us to highlight the short-term pattern which strongly characterizes the ozone levels. Moreover, we can reduce our attention from eight to only two elements when we compute monitoring data differences. That is, we have

$$\Delta_t^Z(\mathbf{s}) = Y_{t+3}(\mathbf{s}) - Y_{t-5}(\mathbf{s}). \tag{2.6}$$

Suppose we insert (2.1) into (2.6). For the resulting $\Delta_t^Z(\mathbf{s})$, the overall intercept β_0 will disappear and we obtain

$$\begin{aligned}\Delta_t^Z(\mathbf{s}) =& \beta_{0,t+3}(\mathbf{s}) + \Big(\beta_1 + \beta_{1,t+3}(\mathbf{s})\Big)W_{t+3}(B) + \epsilon_{t+3}(\mathbf{s}) \\ &- \beta_{0,t-5}(\mathbf{s}) - \Big(\beta_1 + \beta_{1,t-5}(\mathbf{s})\Big)W_{t-5}(B) - \epsilon_{t-5}(\mathbf{s}).\end{aligned} \tag{2.7}$$

Expression (2.7) is still too cumbersome to work with. To expedite computation for model fitting, we will simplify (2.7) so that we regress $\Delta_t^Z(\mathbf{s})$ on the change

in Eta-CMAQ. Let $X_t(\mathbf{s})$ denote the current 8-hour average Eta-CMAQ output for each site $\mathbf{s}$ belonging to the grid cell B. Analogous to (2.5), we define the Eta-CMAQ data differences

$$\Delta_t^x(\mathbf{s}) = 8(X_t(\mathbf{s}) - X_{t-1}(\mathbf{s})).$$

In fact, for $\mathbf{s} \in B$,

$$\Delta_t^x(\mathbf{s}) = W_{t+3}(B) - W_{t-5}(B). \tag{2.8}$$

Figure 2.3 shows the monitoring data differences for four randomly chosen sites and the Eta-CMAQ data differences for the corresponding grid cells, for one-week period. The plots show good agreement between $\Delta_t^Z(\mathbf{s})$ and $\Delta_t^x(\mathbf{s})$ suggesting that the Eta-CMAQ data differences will be useful predictors of the monitoring data differences.

So, we make two simplifying assumptions in (2.7) to connect $\Delta_t^Z(\mathbf{s})$ to $\Delta_t^x(\mathbf{s})$. First, we assume that the slope random effects are not time dependent. This reduces (2.7) to

$$\Delta_t^Z(\mathbf{s}) = \beta_{0,t}^*(\mathbf{s}) + \tilde{\beta}_1(\mathbf{s})\Delta_t^x(\mathbf{s}) + \Delta_t^\epsilon(\mathbf{s}) \tag{2.9}$$

where $\Delta_t^\epsilon(\mathbf{s}) = \epsilon_{t+3}(\mathbf{s}) - \epsilon_{t-5}(\mathbf{s})$ and

$$\begin{aligned} \beta_{0,t}^*(\mathbf{s}) &= \beta_{0,t+3}(\mathbf{s}) - \beta_{0,t-5}(\mathbf{s}) \\ \tilde{\beta}_1(\mathbf{s}) &= \beta_1 + \beta_1(\mathbf{s}). \end{aligned} \tag{2.10}$$

Second, we assume the intercept random effects have a multiplicative form in space and time. We write $\beta_{0,t}^*(\mathbf{s}) = \beta_0(\mathbf{s})\beta_{0,t}$. With say M locations and T time points, we reduce from MT to $(M + T)$ latent variables, with evident computational savings. As we clarify below, this will not imply space-time separability for the dependence structure of the Δ's. Altogether, we introduce three independent zero-mean Gaussian processes, $\beta_{0,t}, \beta_0(\mathbf{s})$, and $\beta_1(\mathbf{s})$. Of course, it would be possible to introduce association between intercept and slope using, say the method of coregionalization (Wackernagel, 2003; Gelfand et al., 2004) briefly described in Chapter 1. However, we do not pursue this further here.

The independence between $\beta_0(\mathbf{s})$ and $\beta_{0,t}$ implies that $\beta_{0,t}^*(\mathbf{s})$ emerges as a zero-mean (nonGaussian) process with a separable covariance structure which we write as

$$Cov[\beta_{0,t}^*(\mathbf{s}), \beta_{0,t'}^*(\mathbf{s}')] = \sigma^2\rho^{(s)}(\mathbf{s} - \mathbf{s}'; \phi_0)\rho^{(t)}(t - t'; \varphi) \tag{2.11}$$

where $\rho^{(s)}$ is a valid two-dimensional spatial correlation function and $\rho^{(t)}$ is a valid one-dimensional temporal correlation. Furthermore, the local spatial adjustment

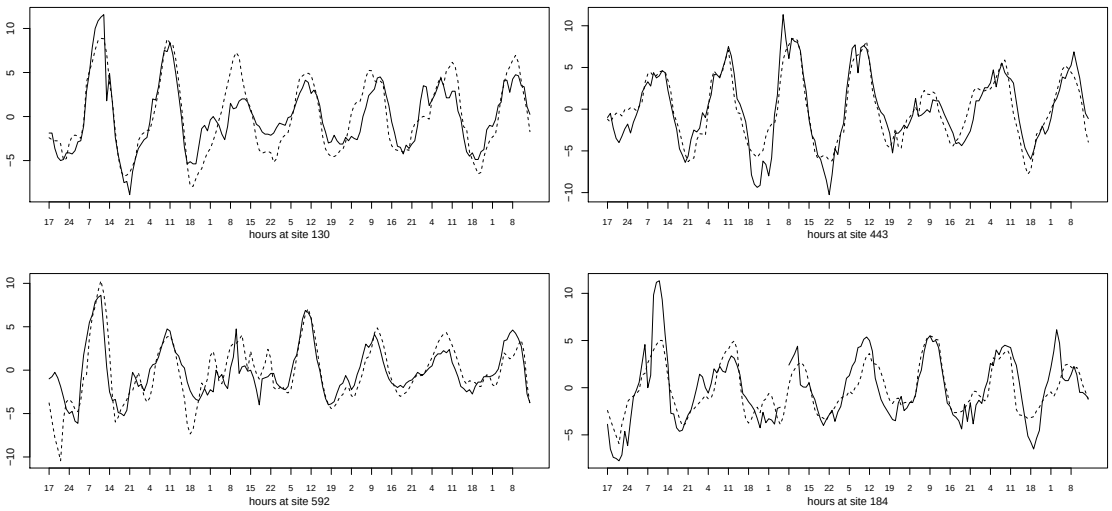

Figure 2.3: Monitoring data differences $\Delta_t^Z(\mathbf{s})$ (solid line) and Eta-CMAQ data differences $\Delta_t^x(\mathbf{s})$ (dashed line) from 4 randomly chosen sites for one-week period.

$\beta_1(\mathbf{s})$ in (2.10) is, again, a zero-mean Gaussian process with covariance structure assumed to be of the form

$$Cov[\beta_1(\mathbf{s}), \beta_1(\mathbf{s}')] = \xi^2 \rho^{(s)}(\mathbf{s} - \mathbf{s}'; \phi_1) \tag{2.12}$$

We acknowledge the simplification associated with the separable specification for $\beta_{0,t}^*(\mathbf{s})$ but note that the resulting process for the Δ^Zs does not have a separable covariance function. Indeed, we have

$$\begin{aligned} Cov[\Delta_t^Z(\mathbf{s}), \Delta_{t'}^Z(\mathbf{s}')] =& \sigma^2 \rho^{(s)}(\mathbf{s} - \mathbf{s}'; \phi_0)\rho^{(t)}(t - t'; \varphi) + \\ &+ \Delta_t^x(\mathbf{s})\Delta_{t'}^x(\mathbf{s}')\xi^2 \rho^{(s)}(\mathbf{s} - \mathbf{s}'; \phi_1) \end{aligned} \tag{2.13}$$

which is nonseparable and, in fact, nonstationary. We take $\rho^{(s)}$ in (2.11) and (2.12) to be exponential correlation functions, i.e. $\rho^{(s)}(\mathbf{s}-\mathbf{s}'; \phi) = \exp(-\phi \|\mathbf{s} - \mathbf{s}'\|)$ while $\rho^{(t)}$ is the correlation function of an $AR(1)$ model, i.e. $\rho^{(t)}(t - t'; \varphi) = \varphi^{|t-t'|}/(1 - \varphi^2)$.

Figure 2.4 gives a graphical representation for the differencing leading to the proposed model. In the figure we can also see the future Δ's necessary for the current 8-hour average forecasting (prediction of these Δ's is discussed in Section 5). So, first differences enable useful simplification of the downscaler: a space-time process for the intercepts and a purely spatial process for the slopes. With a smaller number of parameters and a straightforward dependence structure, we

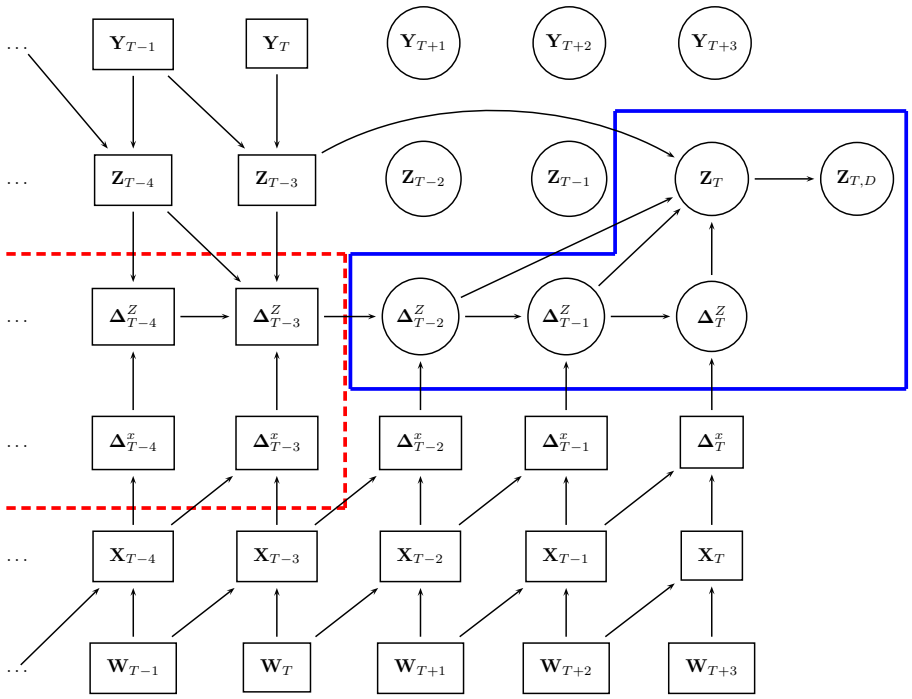

Figure 2.4: Graphical representation of model (2.9)-(2.10) at the current hour T (implicitly, at observed locations). □: observed variables. ◯: unobserved variables. We model the variables inside the dashed box and we predict the quantities inside the solid box. $\mathbf{Z}_{T,D}$ represents the interpolated surface.

reduce the computing time needed for fitting the model and facilitate forecasting the current 8-hour average ozone concentration.

Lastly, we might consider an additive form in space and time for $\beta^*_{0,t}(\mathbf{s})$. The implied simplification in (2.10) is that the spatial effect cancels out and $\beta^*_{0,t}(\mathbf{s})$ becomes purely temporal. So, this corresponds to setting $\beta_0(\mathbf{s}) = 1$ in our above modeling and becomes a reduced model which we can compare with our full specification.

2.3.3 Model fitting

It is well-known that it is not possible to consistently estimate the decay and variance parameter in a spatial model with a covariance function belonging to the Matérn family (Zhang, 2004) as the exponential covariance functions. Moreover, the spatial interpolation is sensitive to the product $\sigma^2\ \phi$ but not to either one individually (Stein, 1999). For these reasons, along with our ongoing objective of rapid computation for model fitting, we choose optimal values of ϕ and φ offline, using a validation mean square error criterion (see Section 2.5) and then infer about the variances conditional on these values.

Denote the remaining unknown parameters by $\boldsymbol{\theta} = (\beta_1, \tau^2, \sigma^2, \xi^2)$. For the parameter β_1 we assume a normal prior distribution $N(0, g^2)$ with g^2 taken to be large. For the variance parameters σ^2, ξ^2 and τ^2 we specify independent proper inverse gamma prior distributions $IG(a, b)$; in our implementation we take $a = 2$ and $b = 1$, i.e., a rather vague prior distribution with mean 1 and infinite variance.

2.3.4 Posterior details

For an observed set of locations $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n$ and hours $t = 1, \ldots, (T-3)$, given $\{\beta_{0,t}\}$, $\{\beta_0(\mathbf{s}_i)\}$, $\{\beta_1(\mathbf{s}_i)\}$ and $\boldsymbol{\theta}$, the $\Delta^Z_t(\mathbf{s}_i)$ are conditionally independent. Hence, the likelihood is

$$L(\boldsymbol{\theta}, \mathbf{B}_0^{(t)}, \mathbf{B}_0^{(s)}, \mathbf{B}_1^{(s)}; \boldsymbol{\Delta}^Z)$$
$$\propto (\tau^2)^{\frac{(T-3)n}{2}} \exp\left\{-\frac{1}{2\tau^2}\sum_{t=1}^{T-3}\sum_{i=1}^{n}\left(\Delta^Z_t(\mathbf{s}_i) - \beta_{0,t}\beta_0(\mathbf{s}_i) - \beta_1\Delta^x_t(\mathbf{s}_i) - \beta_1(\mathbf{s}_i)\Delta^x_t(\mathbf{s}_i)\right)^2\right\}$$

where $\mathbf{\Delta}^Z$ denotes all the data, $\mathbf{B}_0^{(t)} = \left(\beta_{0,1}, \ldots, \beta_{0,T-3}\right)'$, $\mathbf{B}_0^{(s)} = \left(\beta_0(\mathbf{s}_1), \ldots, \beta_0(\mathbf{s}_n)\right)'$ and $\mathbf{B}_1^{(s)} = \left(\beta_1(\mathbf{s}_1), \ldots, \beta_1(\mathbf{s}_n)\right)'$. The joint posterior distribution is given by

$$\begin{aligned}\pi(\boldsymbol{\theta}, \mathbf{B}_0^{(t)}, \mathbf{B}_0^{(s)}, \mathbf{B}_1^{(s)} \,|\mathbf{\Delta}^Z) \propto L(\boldsymbol{\theta}, \mathbf{B}_0^{(t)}, \mathbf{B}_0^{(s)}, \mathbf{B}_1^{(s)}; \mathbf{\Delta}^Z) \times \\ \times \pi(\beta_1) \times \pi(\tau^2) \times \pi(\sigma^2) \times \pi(\xi^2) \times \\ \times \pi(\mathbf{B}_0^{(t)}) \times \pi(\mathbf{B}_0^{(s)}) \times \pi(\mathbf{B}_1^{(s)})\end{aligned}$$

where $\pi(\beta_1)$, $\pi(\tau^2)$, $\pi(\sigma^2)$ and $\pi(\xi^2)$ denote the prior distributions described above. This model is fitted using a Gibbs sampler. The full conditional distributions are developed in Appendix A.

2.4 Prediction details

Once the model is fitted, we turn to the primary goal of forecasting 8-hour average ozone concentration at the current hour T. According to the definition of $Z_T(\mathbf{s})$ in (2.3), we will always need to predict three hours into the future in order to forecast current 8-hour average concentration. Equivalently, monitoring data differences are available up to $\Delta^Z_{T-3}(\mathbf{s})$. So, we need to predict $\Delta^Z_{T-2}(\mathbf{s})$, $\Delta^Z_{T-1}(\mathbf{s})$ and $\Delta^Z_T(\mathbf{s})$ in order to forecast $Z_T(\mathbf{s})$, that is,

$$\begin{aligned}Z_T(\mathbf{s}) &= Z_{T-1}(\mathbf{s}) + \Delta^Z_T(\mathbf{s})/8 \\ &= Z_{T-2}(\mathbf{s}) + \Delta^Z_{T-1}(\mathbf{s})/8 + \Delta^Z_T(\mathbf{s})/8 \\ &= Z_{T-3}(\mathbf{s}) + \Delta^Z_{T-2}(\mathbf{s})/8 + \Delta^Z_{T-1}(\mathbf{s})/8 + \Delta^Z_T(\mathbf{s})/8.\end{aligned} \tag{2.14}$$

Returning to the graphical representation of the model in Figure 2.4, we see how the available information is used to obtain the forecasts we require. As noted in Section 2.2, the Eta-CMAQ forecasts are available 48 hours into the future, so we have the necessary ingredient to make these predictions using the model in (2.9)-(2.10). Predictions at new site $\mathbf{s}'$ and hours of interest $T + l$, $(l = -2, -1, 0)$ are based upon the predictive distribution of $\Delta^Z_{T+l}(\mathbf{s}')$. Under our model (2.9)-(2.10), $\Delta^Z_{T+l}(\mathbf{s}')$ is conditionally independent of the data $\mathbf{\Delta}^Z$ up to time T, given $\boldsymbol{\theta}$, $\beta_{0,T+l}$, $\beta_0(\mathbf{s}')$ and $\beta_1(\mathbf{s}')$ and its distribution is

$$\Delta^Z_{T+l}(\mathbf{s}') \sim N\left(\beta^*_{0,T+l}(\mathbf{s}') + \tilde{\beta}_1(\mathbf{s}')\Delta^x_{T+l}(\mathbf{s}'), \tau^2\right). \tag{2.15}$$

Again, the distribution in (2.15) highlights the contribution of the Eta-CMAQ output $\Delta^x_{T+l}(\mathbf{s}')$ which, as we have noted, is available for these three future hours.

The posterior predictive distribution of $\Delta^Z_{T+l}(\mathbf{s}')$ is given by

$$\begin{aligned}\pi\Big(\Delta^Z_{T+l}(\mathbf{s}') \mid \boldsymbol{\Delta}^Z\Big) = \int & \pi\Big(\Delta^Z_{T+l}(\mathbf{s}') \mid \beta_{0,T+l}, \beta_0(\mathbf{s}'), \beta_1, \beta_1(\mathbf{s}'), \tau^2\Big) \times \\ & \pi\Big(\beta_{0,T+l} \mid \mathbf{B}_0^{(t)}, \varphi\Big) \times \\ & \pi\Big(\beta_0(\mathbf{s}') \mid \mathbf{B}_0^{(s)}, \sigma^2, \phi\Big) \times \\ & \pi\Big(\beta_1(\mathbf{s}') \mid \mathbf{B}_1^{(s)}, \xi^2, \phi\Big) \times \\ & \pi\Big(\boldsymbol{\theta}, \mathbf{B}_0^{(t)}, \mathbf{B}_0^{(s)}, \mathbf{B}_1^{(s)} \mid \boldsymbol{\Delta}^Z\Big) \\ & d\beta_{0,T+l}\, d\beta_0(\mathbf{s}')\, d\beta_1(\mathbf{s}')\, d\mathbf{B}_0^{(t)}\, d\mathbf{B}_0^{(s)}\, d\mathbf{B}_1^{(s)}\, d\boldsymbol{\theta}.\end{aligned} \tag{2.16}$$

The predictive distribution in (2.16) is sampled by composition. In particular, we need to generate draws for $\beta_{0,T+l}$, $\beta_0(\mathbf{s}')$ and $\beta_1(\mathbf{s}')$, conditional on the posterior samples at the observed locations and hours, in order to obtain draws for $\Delta^Z_{T+l}(\mathbf{s}')$. Given the $AR(1)$ model for $\mathbf{B}_0^{(t)}$, we have

$$\beta_{0,T+l} \mid \mathbf{B}_0^{(t)}, \varphi \sim N\Big(\varphi\beta_{0,(T+l-1)}, 1\Big)$$

For the spatially varying intercept, the joint distribution of $\mathbf{B}_0^{(s)}$ and $\beta_0(\mathbf{s}')$ is a multivariate normal from which the conditional distribution is the univariate normal

$$\beta_0(\mathbf{s}') \mid \mathbf{B}_0^{(s)}, \sigma^2, \phi \sim N\Big(\sum_{i=1}^{n} b_i(\mathbf{s}')\beta_0(\mathbf{s}_i), \sigma^2 C(\mathbf{s}')\Big)$$

where

$$b_i(\mathbf{s}') = \sum_{j=1}^{n} \rho^{(s)}(\mathbf{s}' - \mathbf{s}_j; \phi)\big(H^{-1}(\phi)\big)_{ij}$$

and

$$C(\mathbf{s}') = 1 - \sum_{j=1}^{n}\sum_{i=1}^{n} \rho^{(s)}(\mathbf{s}' - \mathbf{s}_i; \phi)\big(H^{-1}(\phi)\big)_{ij}\rho^{(s)}(\mathbf{s}_j - \mathbf{s}'; \phi)$$

Similarly, we generate the random variable $\beta_1(\mathbf{s}')$ conditional on the posterior samples at the observed locations. For this, we have

$$\beta_1(\mathbf{s}') \mid \mathbf{B}_1^{(s)}, \xi^2, \phi \sim N\big(\sum_{i=1}^{n} b_i(\mathbf{s}')\beta_1(\mathbf{s}_i), \xi^2 C(\mathbf{s}')\big)$$

where $b_i(\mathbf{s}')$ and $C(\mathbf{s}')$ are defined as above. The conditional means and variances are computationally expensive to compute. However, by fixing the decay parameters ϕ and φ, the quantities b_i and $C(\mathbf{s}')$ need only be calculated once and stored; no updating is required in the MCMC, facilitating real-time forecasting.

2.4.1 Forecast map

Recall that our goal is to provide, in a real-time environment, hourly spatial interpolation maps of 8-hour average ozone concentration. To obtain these maps, we need spatial predictions at each Eta-CMAQ grid cell centroid, such as what the EPA AIRnow system supplies, roughly $54,000$ cells. Given the limited time available to produce plausible predictions at such a large number of grid cell points, formal Bayesian kriging (as in say, Banerjee et al. (2004)) will not offer a feasible approach. So, at this last stage, we introduce approximation. Again, this last stage is only for the map making. There will be sufficient time for the foregoing model fitting.

The strategy is to use equation (2.14) to obtain predictions at the n monitoring sites. Then, we interpolate these predictions to the Eta-CMAQ grid cell centroids by ordinary kriging, using a fast, available package. In this regard, we can adopt one of the following approaches. The first method is to apply the kriging interpolation both to $Z_{T-3}(\mathbf{s}_i)$ and the posterior predictive samples of $\Delta^Z_{T-2}(\mathbf{s}_i)$, $\Delta^Z_{T-1}(\mathbf{s}_i)$ and $\Delta^Z_T(\mathbf{s}_i)$, with $i = 1, \ldots, n$. Then, the posterior predictive distribution of $Z_T(\mathbf{s})$ at the Eta-CMAQ centroids can be provided by the sum in (2.14). This approach, however, will be slow and it will introduce large uncertainty to the predictions. Thus, we first sum the last available observation $Z_{T-3}(\mathbf{s}_i)$ and the posterior predictive samples of $\Delta^Z_{T-2}(\mathbf{s}_i)$, $\Delta^Z_{T-1}(\mathbf{s}_i)$ and $\Delta^Z_T(\mathbf{s}_i)$. Then, we obtain the posterior predictive distribution of $Z_T(\mathbf{s})$ at the Eta-CMAQ centroids by kriging. We get the predicted surface of 8-hour average ozone concentration as an average of the posterior predictive distribution of the kriged $Z_T(\mathbf{s})$. A posterior standard deviation map gives a measure of the uncertainty associated with our forecasts.

2.5 Analyses

We illustrate our strategy by modeling and forecasting ozone level for a large conterminous subregion of U.S. (Figure 1). In particular, we model data for a running window of 24 hours, starting at any given hour. We have investigated longer windows, such as 48-hour and 72-hours. However, the higher computational burden associated to more distant past data is not justified in terms of any improvement in the predictions.

About 5% of values are missing in the monitoring data set. We decided to handle the missingness by removing monitoring sites with at least one missing value

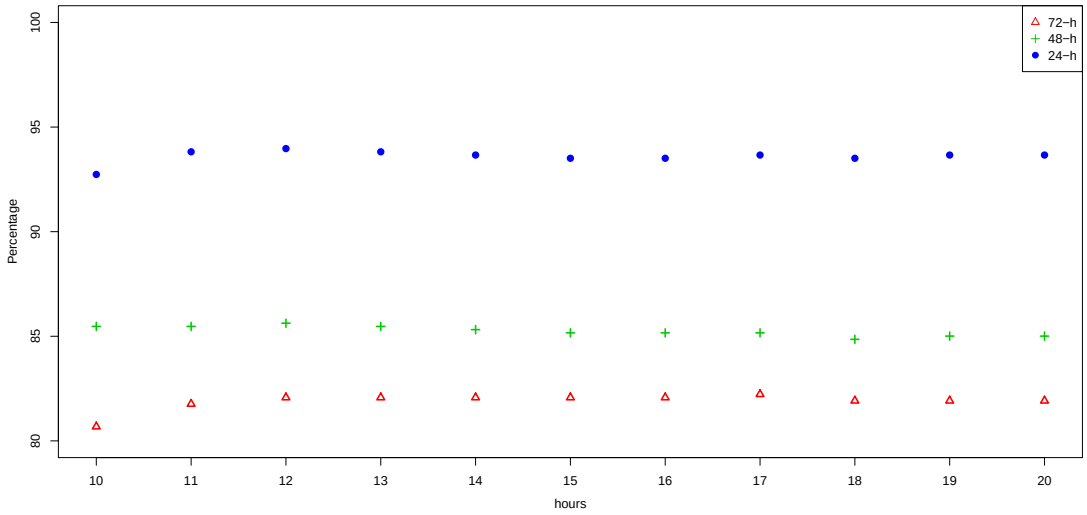

Figure 2.5: Percentage of monitoring sites available to fit the model.

in each selected 24-hour window. This choice reflects the structure of the missing values in the data set. As the window changes in time, so do the locations of the missing data. However, in general, missing values occur at monitoring sites for several consecutive hours. This discourages attempts to use 'cheap' imputation; alternatively, a fully model-based imputation would be too computationally expensive. Figure 2.5 shows the percentage of monitoring sites available to fit the model with respect to 24-hour, 48-hour and 72-hour windows. The 24-hour window enable us to save more than 93% percentage of monitoring sites. So, in addition to being computationally faster, the 24 hour window gains roughly 7% more sites than, say the 48 hour window.

First, we select the decay parameters using the validation criterion described below, recalling that we have set aside data from 70 monitoring stations (Figure 2.1). For convenience, we set $\phi_0 = \phi_1 = \phi$, imagining that the spatial range for the slope process might agree with that of the intercept process (this simplification is not critical and is really just illustrative). For ϕ and φ, let $\hat{\Delta}_t^Z(\mathbf{s}'_j)$ denote the predicted value at validation site $\mathbf{s}'_j$ for each $j = 1, \ldots, m = 70$ and hours $t = 1, \ldots, (T-3) = 24$.

We employ the Validation Mean Square Error (VMSE)

$$VMSE = \frac{1}{n_v} \sum_{j=1}^{m} \sum_{t=1}^{(T-3)} \left(\Delta_t^Z(\mathbf{s}'_j) - \hat{\Delta}_t^Z(\mathbf{s}'_j) \right)^2 I\left(\Delta_t^Z(\mathbf{s}'_j) \right) \tag{2.17}$$

where $n_v = \sum_{j=1}^{m} \sum_{t=1}^{(T-3)} I\left(\Delta_t^Z(\mathbf{s}'_j) \right)$ is the total number of available observations at the 70 validation sites for the 24 hours. We searched for the optimal value of

ϕ among the values, 1.5, 0.5 and 0.25 corresponding to spatial ranges of approximately 185, 560 and 1125 kilometers. For the temporal decay parameter φ, we searched for the optimal value in a grid formed by values of 0.75, 0.85 and 0.95.

For each selected 24-hour window, we compute the VMSE in (2.17) and we choose the combination of ϕ and φ which leads to the smallest VMSE. For instance, Table 2.1 shows the VMSE computed on the predictions at the validation sites for a given 24-hour window for each combination of ϕ and φ. In this case, we choose the values 0.25 and 0.95 as estimates of the parameters ϕ and φ, respectively. We experimented with many other values of ϕ and φ learning that the VMSE is not very sensitive to choices close to these optimal values. In fact, even a finer grid of values of ϕ and φ yields to results which are essentially equivalent to those presented in Table 2.1.

Table 2.1: VMSE for each combination of ϕ and φ when we model data starting at 10AM on August 7th.

		φ		
		0.75	0.85	0.95
ϕ	1.50	1.77075	1.77068	1.77062
	0.50	1.71869	1.71867	1.71866
	0.25	1.71792	1.71790	1.71789

We fit the model in (2.9)-(2.10) on 24-hour running windows starting at each hour from 8AM to 6PM of August 7th in order to forecast current 8-hour average ozone concentration from 10AM to 8PM on August 8th; this particular temporal window is characterized by a high level of variability in ozone concentrations. For each selected window, we predict monitoring data differences at the n available monitoring sites for the three future hours (corresponding to $\Delta^Z_{T-2}(\mathbf{s}_i)$, $\Delta^Z_{T-1}(\mathbf{s}_i)$ and $\Delta^Z_T(\mathbf{s}_i)$) and we forecast the current 8-hour average ozone concentrations ($Z_T(\mathbf{s}_i)$, for $i = 1, \ldots, n$). Then, these forecasts are interpolated to the Eta-CMAQ centroids, as we described in Section 2.4.1. For example, starting at 8AM of August 7th, we model 24 hourly monitoring data differences from 8AM on August 7th to 7AM on August 8th using data from all available monitoring sites. Predictions of monitoring data differences are computed at the monitoring sites for 8AM, 9AM and 10AM on August 8th and forecasts of the current 8-hour average ozone concentrations at the Eta-CMAQ centroids, associated with 10AM on August 8th, are obtained.

Table 2.2: Posterior parameter estimates under full model when we model data starting at 10AM on August 7th; 95% credible interval in the brackets.

β_1	τ^2	σ^2	ξ^2
0.634 (0.512 - 0.776)	1.470 (1.433 - 1.506)	0.499 (0.364 - 0.686)	0.453 (0.369 - 0.564)

2.5.1 Results

An example of parameter estimates is shown in Table 2.2 along with Figures 2.6 and 2.7 for the modeling of the data from 10AM of August 7th to 9AM of August 8th. The significant overall slope β_1 shows the expected positive association between Eta-CMAQ data differences and monitoring data differences. Mean spatial effects $\beta_0(\mathbf{s})$ and $\beta_1(\mathbf{s})$ are shown in Figure 2.6. Figure 2.7 shows the 95% credible intervals of the temporal effect. We see the anticipated higher variability for the three hours into the future. The diurnal pattern which characterizes the ozone levels is well reproduced on the first differences scale. Overall, the multiplicative form for $\beta^*_{0,t}(\mathbf{s})$ yields spatio-temporal intercepts that provide an hourly scaling of $\beta_0(\mathbf{s})$. Notably, we observed similar parameters estimates for all other starting hours. We illustrate the current 8-hour average map prediction at 12PM on August 8th in Figure 2.8 (left panel). The right panel shows the standard deviation map. For instance, the figure reveals that the highest ozone concentrations characterize the States of Delaware, Maryland, Virginia and North Carolina while the blue area over Florida supports that Florida's air quality can be considered fairly good, as we expected. Therefore, accurate, instantaneous and high resolution maps as in Figure 2.8 represent a useful tool to provide the public and the experts with air pollution levels that may lead to adverse health effects.

As a concluding exercise, we compare the *out-of-sample* predictive performance of the model (2.9)-(2.10) and the simpler version obtained by fixing the pure spatial component in the intercept $\beta_0(\mathbf{s}) = 1$. We evaluate Bayesian predictions by computing the mean squared error (MSE), mean absolute error (MAE), empirical coverage and average length of the 95% credible interval on $70 \times 11 = 770$ out-of-sample forecasts. Table 2.3 reports results for these summary statistics for the two models, revealing little difference except for somewhat shorter predictive intervals for the *reduced* model. This may be an artifact of the validation sites or may reflect possible overfitting for the *full* model. The empirical coverages agree and are a bit below nominal, suggesting that the intervals are bit short. This is likely be due to the simplifications we make in the model for the differences. Figure 2.9 provide

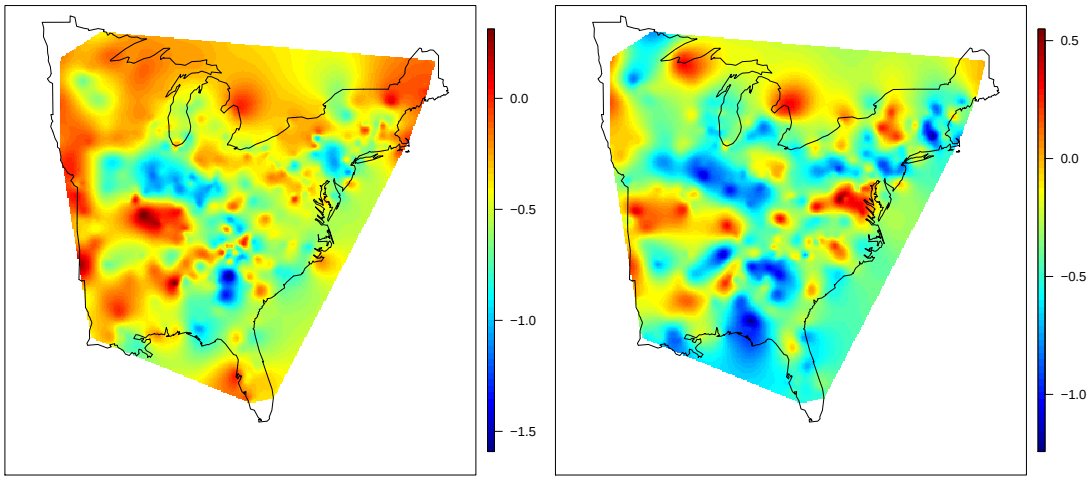

Figure 2.6: Mean spatial effects $\beta_0(\mathbf{s})$ (left panel) and $\beta_1(\mathbf{s})$ (right panel) when we model data starting at 10AM on August 7th.

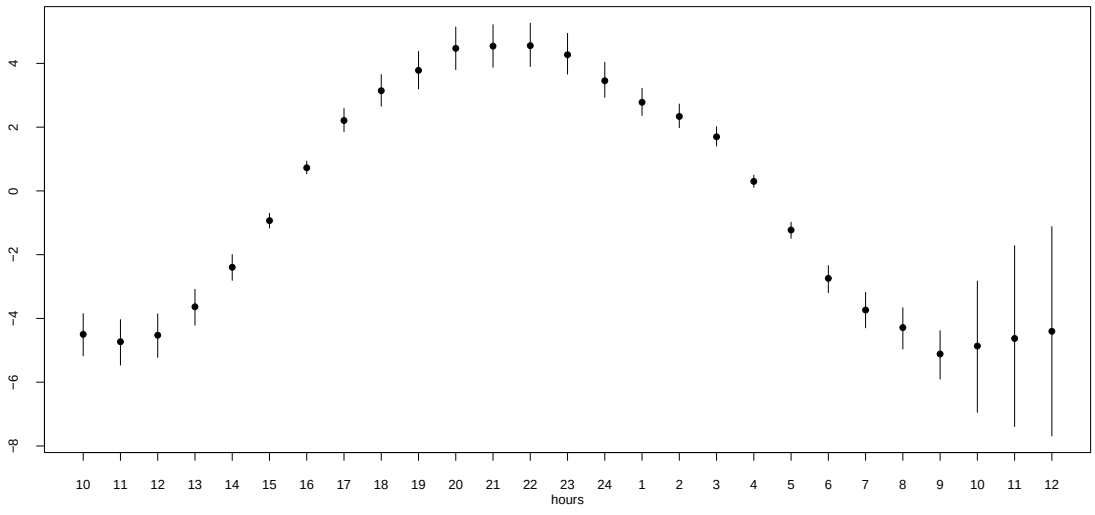

Figure 2.7: 95% credible interval of the temporal component $\beta_{0,t}$ when we model data starting at 10AM on August 7th.

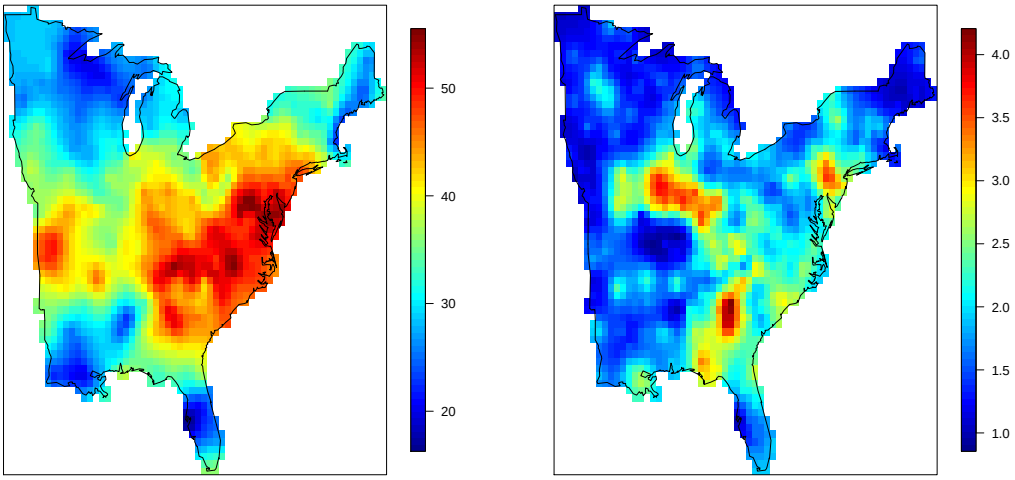

Figure 2.8: Current 8-hour average ozone forecast map (left panel) and standard deviation map (right panel) at 12PM on August 8th.

Table 2.3: Mean square error (MSE), mean absolute error (MAE), empirical coverage and average length of 95% predictive intervals (PI) for full model with $\beta_0(\mathbf{s}) \neq 1$ and reduced model with $\beta_0(\mathbf{s}) = 1$.

	MSE	MAE	Empirical coverage of 95% PI	Average length of 95% PI
$\beta_0(\mathbf{s}) \neq 1$	24.97	3.80	85.7%	15.70
$\beta_0(\mathbf{s}) = 1$	24.66	3.79	85.5%	13.67

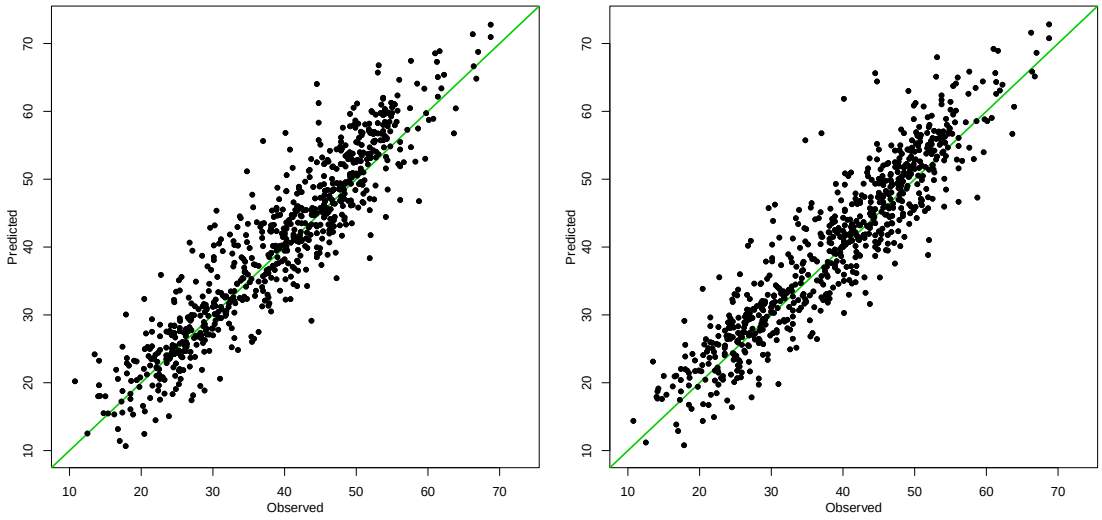

Figure 2.9: Validation plots for out-of-sample predictions using the full model (left panel) and the reduced model (right panel). The 45 degree reference line is superimposed.

detailed validation plots for the out-of-sample predictions obtained from both the full model and the reduced one; again, little differences appeared between the two models.

We can offer comparison with AIRNow predictions for the same time period. We have to consider this comparison with care for the following reasons. AIRNow makes its forecasts at each monitoring station, treating the stations as independent, building a historical regression at each station, and makes a simple local forecast. Then, AIRNow uses a kriging routine to predict to the continental scale. It does not use any computer model output. In particular, for any specified hour, AIRNow uses the subset of monitoring stations that reported for that hour, before kriging; the set of sites employed varies by the hour. So, we can consider two comparisons. Starting with our 717 monitoring stations, holding out 70 of them, leaves us with 647 fitting sites. We make hourly predictions for a subset of these sites as clarified above. So does AIRNow but for a different subset. So, hour by hour, if we consider the intersection of these two subsets, and for the intersection, take our predictions and those of AIRNow, we are able to make a fair, pre-interpolation comparison of forecasts. These results are shown in the first two columns of Table 2.4 and reveal a roughly 30% improvement in prediction at fitted sites. Interestingly, if we then interpolate hour by hour to the 70 hold-out sites, using a commonly employed kriging R-package *'fields'* (http://www.image.ucar.edu/Software/Fields) we obtain

the results in the last two columns of Table 2.4. We see that the kriging procedure introduces smoothing such that it reduces the benefit of our modeling approach in terms of interpolated predictive performance. Still we do improve and, in addition, we do have a measure of uncertainty through the predictive variance. Indeed, the results from Table 2.3 show that MSE and MAE for the Bayesian forecast validation at the holdout sites are indistinguishable from the pre-interpolation forecast validation results in Table 2.4 clarifying the improvement we would expect to see were we able to implement fully model based Bayesian kriging in real-time.

Finally, fitting the faster model of Sahu et al. (2009a) to our 8-hour average data inputs, we obtained[2] MSE = 50.64 and MAE = 5.61, somewhat larger than what we obtained for our models in Table 2.4.

Table 2.4: Mean square error (MSE) and mean absolute error (MAE) for full model, reduced model and AIRNow forecasts.

	Pre-Interpolation		Post-Interpolation	
	MSE	MAE	MSE	MAE
$\beta_0(\mathbf{s}) \neq 1$	25.46	3.91	42.35	4.96
$\beta_0(\mathbf{s}) = 1$	24.43	3.87	41.95	4.95
AIRNow	36.39	4.73	45.72	5.35

2.6 Real-time computing

Regarding the feasibility of our method for real-time use, in terms of offline fitting time and time per hourly update, we note the following. The fitting time is evaluated per iteration of Markov Chain Monte Carlo (MCMC) on an Intel(R) Core(TM)2 Duo CPU E8600 (3.33 GHz, 8 GB RAM). The MCMC is well-behaved and the convergence is rapid. The computing time necessary to fit model (3.2) with $\beta_0(\mathbf{s}) = 1$ is about 1.1 seconds per iteration. The hourly update involves the forecasts of current 8-hour ozone concentrations. Typically, only three seconds are required to obtain each posterior predictive sample of $Z_T(\mathbf{s})$ at the Eta-CMAQ centroids, according to the strategy described in Section 2.4. The code is written

[2]These summary statistics are based on $50 \times 11 = 550$ forecasts. The corresponding statistics computed over our forecasts for the same hour-site combinations are: MSE = 42.31 and MAE= 4.91 for the full model and MSE= 42.08 and MAE= 5.01 for the reduced model.

in R and we can assert that, for the region we have investigated, our approach does work in real-time.

Moving to national scale where there are about 1400 monitoring sites, the code will be written with C^{++}, which would run possibly an order of magnitude faster compared to R code. We also have been using a single machine; a national undertaking would be expected to employ a better hardware environment, at the least, to run on a faster, multi-processor machine. Alternatively, it may prove more attractive to consider regional models and follow, for each region, the same path as we have developed above. In this way we can capture local effects and directly expedite computation by using parallelization.

2.7 Summary

In this chapter, we have addressed a specific applied challenge, real-time forecasting of current 8-hour average ozone levels on the scale of the conterminous U.S.. We have formulated a downscaler model that works with differences to expedite computation and have shown that it outperforms the current forecasting system. One added advantage of our proposed real-time forecasting model is the potential archival and access to current ozone spatial information. This would allow immediate access to up-to-date ozone patterns and solves the problem of waiting several years for retrospective numerical atmospheric model output to become available to develop predictive ozone surfaces.

Future work will focus on introducing real-time temperature data, as described in Chapter 3. We will also consider improved *next day* ozone forecasts. We are also interested in current and next day particulate matter forecasting where new challenges arise because particulate matter is not necessarily collected on a continuous, daily basis at the monitoring sites.

Finally, future efforts will find us looking at the possibility of considering the Partial Differential Equations (SPDE) approach proposed by Lindgren et al. (2011) in order to make use of the Integrated Nested Laplace Approximation (INLA) algorithm (Rue et al., 2009) as an alternative to MCMC methods adopted in this chapter.

Chapter 3

Ozone real-time forecasting downscaling temperature

Since one of the main objective of this work is to improve the assessment of ozone exposure within the real-time environment, we introduced the real-time downscaler in Chapter 2, fusing the ozone station data and the output from the air quality Eta-CMAQ model. An unexpected difficulty is due to the fact that the air quality computer model was not longer run by EPA. Since the end of 2012 the Eta-CMAQ model has been dismissed and its output was not available anymore[1]. The new circumstances require to replace the Eta-CMAQ output in the downscaler by a new ozone predictor that enables us to yield accurate real-time ozone forecasting. Thus, we now look for another covariate strongly correlated with ozone concentrations as well as available in real-time for current and future time periods.

Variations in weather conditions play an important role in producing ozone concentrations, as we noted in Section 2.1. In particular, ozone is strongly correlated with temperature (Cox and Chu, 1996; Bloomfield et al., 1996; Jacob and Winner, 2009) since the temperature influences directly the kinetics of reactions determining ozone (Cocchi and Trivisano, 2013). In general, increasing temperature is usually associated with increasing ground-level ozone levels. The ozone-temperature relationship has been largely investigated in the literature. Massart and Kvalheim (1998) studied the importance of several meteorological variables (such as wind speed, wind direction, air stability, temperature and light intensity) for forecasting the next day's ozone level for the region of Grenland (southern Norway); their

[1]In practice, Eta-CMAQ output was no longer available for empirical studies since the end of 2011.

main result was that the temperature produced the best ozone predictions. In Thompson et al. (2001) several statistical methods for meteorological adjustment of ground level ozone are discussed; temperature is included as a covariate in most models reviewed by the authors. Hence, temperature can be used as a surrogate for the meteorological factors influencing ozone formation (Camalier et al., 2007; Bloomer et al., 2009).

To identify the most suitable source of temperature data in our context, we recall that an important feature of the new covariate is its availability in real-time for current and future hours. Again, predictions from numerical models can usually have very high temporal resolution for current, past and future time periods. Moreover, we look for a data source which provides temperature at high spatial resolution covering the conterminous U.S., since our goal is to produce real-time ozone forecasting at the national scale.

The U.S. National Oceanic and Atmospheric Administration (NOAA) developed several weather forecast models providing predictions of many meteorological variables (such as temperature, relative humidity, precipitation, sea-level pressure) at different temporal and spatial resolution. The NOAA's National Climatic Data Center (NCDC) provides near-real-time easy access to these weather model forecast data in addition to historical model data at the web site: http://www.ncdc.noaa.gov/. Among the NOAA's numerical models, weather short-term predictions for the conterminous U.S. are produced by the Rapid Update Cycle (RUC) model (until May, 2012) and the Rapid Refresh (RAP) model (from May, 2012).

Recall that in the univariate downscaler the response and its predictor are the same variable expressed at different spatial resolution, i.e. a pollutant at point level and its prediction from a numerical model output at grid cell spatial resolution. In this chapter we modify our earlier downscaler for real-time ozone forecasting such that the covariate differs from the response variable. We combine ozone data from real-time monitoring network with temperature output from a weather computer model via a regression model having space-time varying coefficients along with first differences. Model validation for the eastern U.S. shows improved predictions of current 8-hour average ozone levels relative to those obtained using the air quality model output as a predictor and presented in Chapter 2.

The chapter is organized as follows. In Section 3.1 we describe the weather numerical models we use for ozone prediction. Modeling developments are presented in Section 3.2. Section 3.3 provides the analyses and results considering both RUC and RAP output. In Section 3.4 we briefly discuss short-term ozone predictions

obtained by Bruno and Paci (2013) for the Emilia-Romagna region. Concluding remarks are summarized in Section 3.5.

3.1 RUC and RAP models

The RUC model (Benjamin et al., 2004) is a regional short-term weather forecast model of the Continental United States (CONUS) developed by the National Centers for Environmental Prediction (NCEP) to serve users needing frequently updated short-term weather forecasts. When it was first implemented in 1994, the model was run every three hours making forecasts out to 12 hours. By 2002, the RUC was run every hour, on the hour, producing 12-hour forecasts at 13 km spatial resolution. The output from the RUC model is available, for free, at the website: http://ruc.noaa.gov/.

Starting on May 1, 2012, the NCEP replaced the RUC model by the RAP numerical weather model (http://rapidrefresh.noaa.gov/). The RAP model is the next-generation version of the 1-hour cycle system; multiple data sources go into the RAP forecasts such as commercial aircraft weather data, balloon data, radar data, surface observations, and satellite data. The RAP model shows improvement over RUC forecasts for wind, relative humidity, temperature, and heights at almost all levels and forecast durations, as claimed by the NCEP. RAP forecasts are generated every hour with forecast lengths going out 18 hours at 13 km spatial resolution.

We consider surface temperature forecasts (2 meter above the ground, in °C) from the weather numerical model for the conterminous U.S.. As an illustration, Figure 3.1 shows the temperature predictions from RUC model at 10AM on August 7th, 2011 in the eastern U.S..

Since there is no overlap of RUC and RAP output in a time period, we first analyze ozone and RUC data during August 2011; for this period we will also have ozone estimates from Eta-CMAQ model. Then, we will consider ozone and RAP output corresponding to August 2012 when Eta-CMAQ output was not available anymore.

3.2 Downscaler using temperature

Current 8-hour average ozone level $Z_t(\mathbf{s})$ is defined, according to equation (2.3), as the average of the previous four hours, the current hour and the next three hours in

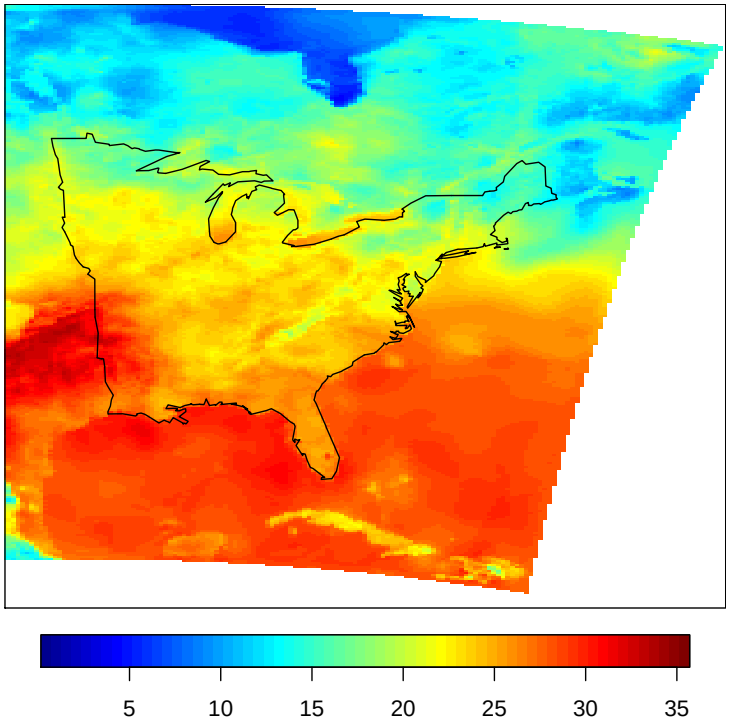

Figure 3.1: Temperature forecasts from RUC model in the eastern U.S. at 10AM on August 7th, 2011.

the future. In Chapter 2, we developed a space-time data assimilation strategy to enable use of both monitoring data and Eta-CMAQ output to provide the forecasts of current 8-hour average ozone level in real-time. We also showed the benefit of using the monitoring data differences $\Delta_t^Z(\mathbf{s})$ denoted by $\Delta_t^Z(\mathbf{s}) = 8\big(Z_t(\mathbf{s}) - Z_{t-1}(\mathbf{s})\big)$, i.e. (2.5) or equivalently, by $\Delta_t^Z(\mathbf{s}) = Y_{t+3}(\mathbf{s}) - Y_{t-5}(\mathbf{s})$, i.e. (2.6).

The model we proposed in Chapter 2 (and, in general, any downscaler) combines the observations of the pollutant with its predictions from the air quality numerical model. However, we already mentioned the dismission of the Eta-CMAQ model and the need to replace it by a new ozone predictor available in real-time. In this work, we generalized the downscaler model to include discrepancies between the response variable and the covariate. We introduce a data fusion model based on first differences of ozone real-time monitoring data and temperature estimates from a weather numerical model. In particular, we replace the Eta-CMAQ output in model (2.9) by the RUC (RAP) output, so we will regress the monitoring data differences on the change in RUC (RAP) output.

Similarly to (2.8), for each site $\mathbf{s}$ belonging to grid cell B, we define the RUC (RAP) data differences by

$$R_t^*(\mathbf{s}) = R_{t+3}(B) - R_{t-5}(B) \tag{3.1}$$

where $R_t(B)$ denotes the temperature forecasts from RUC (RAP) model at hour t over grid cell B. Thus, the definition of variable $R_t^*(\mathbf{s})$ guarantees the temporal alignment between the weather numerical model output and ozone monitoring data differences.

Figure 3.2 shows the ozone monitoring data differences $\Delta_t^Z(\mathbf{s})$ for six randomly chosen sites and the RUC data differences $R_t^*(\mathbf{s})$ for the corresponding grid cells, during the period August 6-9, 2011. The plots reveal good agreement between ozone monitoring data differences and RUC data differences; in fact, for the same period, the overall correlation is 0.78. Figure 3.3 shows, instead, the ozone monitoring data differences for six randomly chosen sites and the RAP data differences for the corresponding grid cells, during for the period August 1-2, 2012. Again, ozone monitoring data differences and RAP data differences show similar behavior and, for this period, the overall correlation between is 0.84. Therefore, the explanatory analyses suggest that $R_t^*(\mathbf{s})$ differences can be good predictors of the ozone monitoring data differences.

Similarly to the developments in Section 2.3, we address the spatial misalignment between ozone monitoring data differences and RUC (RAP) differences by

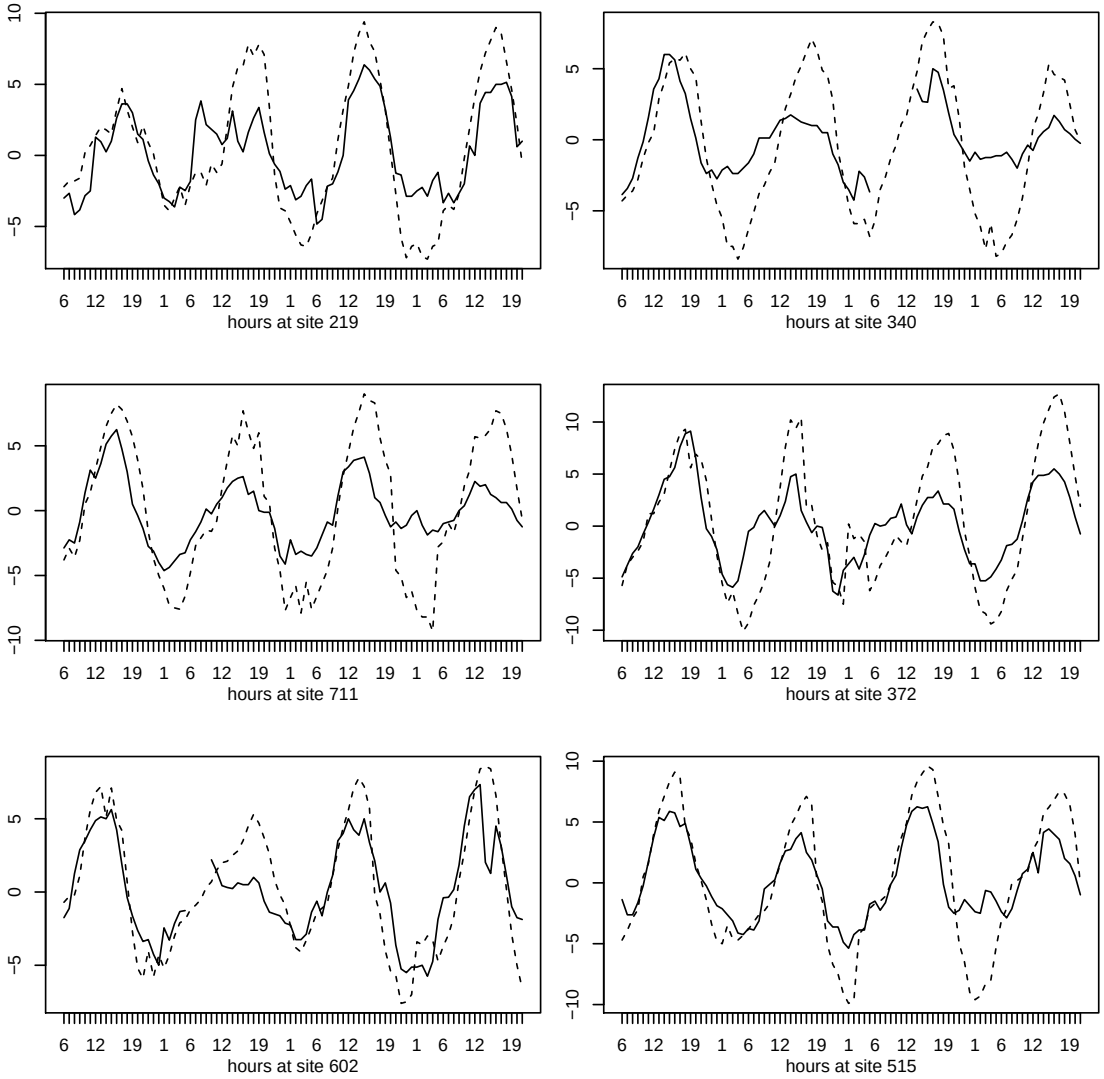

Figure 3.2: Monitoring data differences (solid line) and corresponding RUC data differences (dotted line) from 6 randomly chosen sites during August 6-9, 2011.

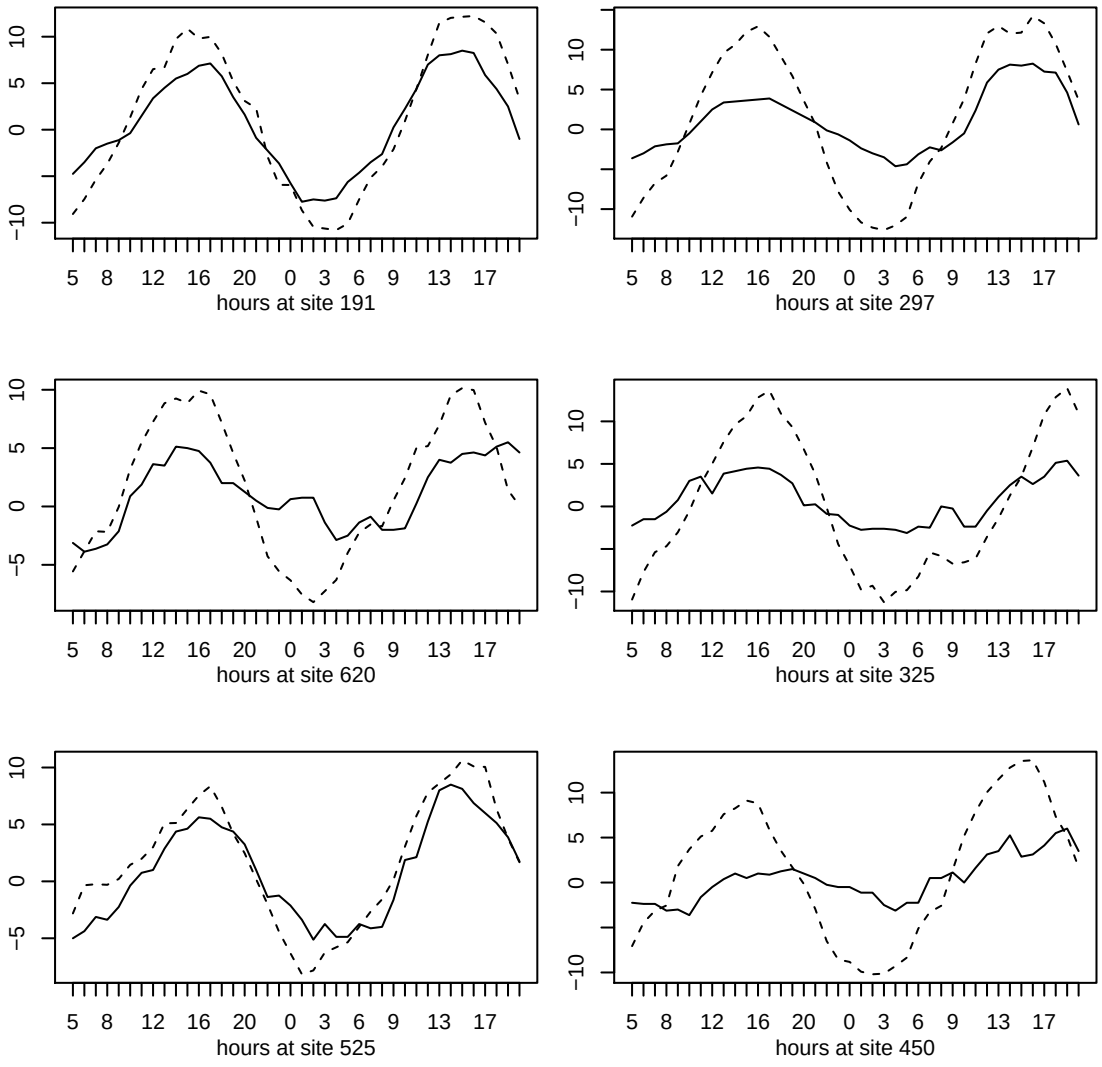

Figure 3.3: Monitoring data differences (solid line) and corresponding RAP data differences (dotted line) from 6 randomly chosen sites during August 1-2, 2012.

employing the downscaling approach. For each $\mathbf{s} \in B$, model (2.9) is modified as follows:

$$\Delta_t^Z(\mathbf{s}) = \beta_{0,t}^*(\mathbf{s}) + \tilde{\beta}_1(\mathbf{s})R_t^*(\mathbf{s}) + \Delta_t^\epsilon(\mathbf{s}) \tag{3.2}$$

where $\beta_{0,t}^*(\mathbf{s})$, $\tilde{\beta}_1(\mathbf{s})$ and $\Delta_t^\epsilon(\mathbf{s})$ are defined as in Chapter 2. Again, we assume that the slope random effects are not time dependent and the intercept random effects have a multiplicative form in space and time, that is $\beta_{0,t}^*(\mathbf{s}) = \beta_0(\mathbf{s})\beta_{0,t}$. The spatio-temporal covariance function of the data process is now adapted from (2.13) as

$$Cov[\Delta_t^Z(\mathbf{s}), \Delta_{t'}^Z(\mathbf{s}')] = \sigma^2\rho^{(s)}(\mathbf{s}-\mathbf{s}';\phi_0)\rho^{(t)}(t-t';\varphi) + R_t^*(\mathbf{s})R_{t'}^*(\mathbf{s}')\xi^2\rho^{(s)}(\mathbf{s}-\mathbf{s}';\phi_1)$$

which is still nonseparable and nonstationary.

Fitting details associated to model (3.2) are equivalent to those described in Section 2.3.3. Again, we use non informative prior distributions for the unknown parameters and we fit the model using a Gibbs sampler.

Predictions of the current 8-hour average ozone level at a new site $\mathbf{s}'$ and hours $T+l$, $(l = -2, -1, 0)$ are obtained via the conditional posterior predictive distribution

$$\Delta_{T+l}^Z(\mathbf{s}') \sim N\left(\beta_{0,T+l}^*(\mathbf{s}') + \tilde{\beta}_1(\mathbf{s}')R_{T+l}^*(\mathbf{s}'), \tau^2\right).$$

The predictive distribution is sampled by composition as described in Section 2.4. Such forecasting is still feasible since the RUC (RAP) output is available up to 12 (18) hours in the future. Finally, the forecast map is produced according to the strategy described in Section 2.4.1; again, we use equation (2.14) to obtain predictions at the n monitoring sites. Then, we interpolate these predictions to the Eta-CMAQ grid cell centroids by ordinary kriging, using a fast, available package. We get the predicted surface of current 8-hour average ozone concentration as an average of the posterior predictive distribution of the kriged $Z_T(\mathbf{s})$. The posterior standard deviation map gives a measure of the uncertainty associated with our forecasts.

In the next section, we present the results obtained by fitting model (3.2) where we fix the pure spatial component in the intercept at $\beta_0(\mathbf{s}) = 1$ (reduced model). In fact, Table 2.3 revealed little differences between the full model and the reduced one. However, the computing time necessary to fit the reduced model is smaller (roughly the half) than the fitting time needed for the full model. Hence, the simpler version of our model appears more suitable to use within the real-time environment and so we investigate the predictive performance of this model.

3.3 Predictive performance

Recall that there is no overlap of RUC and RAP output in a time period. Hence, we first assess the predictive performance of our new strategy using ozone historical data from the large subregion described in Section 2.2.1 with $n = 717$ real-time ozone stations during August 2011. Moreover, ozone estimates from Eta-CMAQ are available as well as AIRNow predictions. For the same period, temperature data arises as output from the RUC model. The weather forecast data are averages over grid cells, i.e. 17,773 grid cells spanning the study region. So, in Subsection 3.3.1 we can offer a comparison with predictions obtained in the previous chapter.

In Subsection 3.3.2, we illustrate our strategy by modeling current 8-hour ozone level collected from $n = 696$ real-time monitoring stations operating in the eastern U.S. during August 2012. In this case, temperature forecasts are provided by the RAP model at the 17,773 grid cells spanning the study region. For this period, Eta-CMAQ output is not available but we can still present a comparison between ozone forecasts obtained from our model and those provided by the AIRNow system.

We use the procedure described in Section 2.5 to handle the missingness in data sets. Thus, we remove monitoring sites with at least one missing value in each selected 24-hour window.

3.3.1 Results using RUC output

Equivalently to what is done in Section 2.5, we fit model (3.2) on 24-hour running windows starting at each hour from 8AM to 6PM on August 7th, 2011 in order to forecast current 8-hour average ozone level at 70 validation sites from 10AM to 8PM on August 8th, 2011.

Again, we evaluate Bayesian predictions by computing the mean squared error (MSE), mean absolute error (MAE), empirical coverage and average length of the 95% credible interval on $70 \times 11 = 770$ out-of-sample forecasts. Table 3.1 reports results for these summary statistics for model (2.9) using Eta-CMAQ data differences and model (3.2) using RUC data differences. We note that the proposed strategy yields to increased accuracy in ozone predictions relative to the results in Section 2.5. This result might be surprising if we recall that the Eta-CMAQ model is essentially devoted to estimate air pollution concentrations and its output has a finer spatial resolution with respect to temperature data. However, the output from the weather computer model is less smooth than the air quality model output and this feature enables us to compute more accurate ozone forecasts.

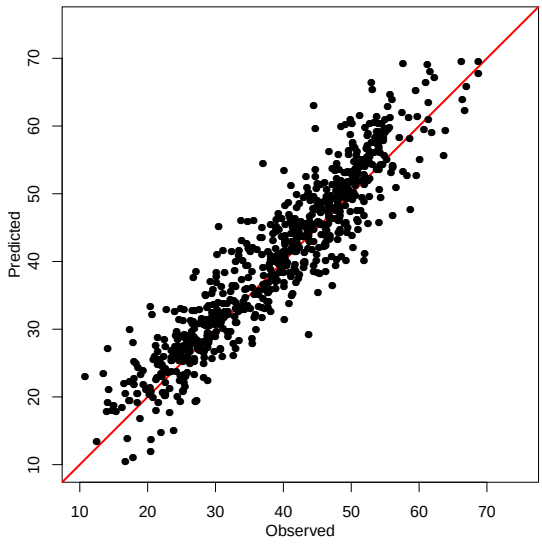

Figure 3.4: Validation plot for out-of-sample predictions when model (3.2) is fitted using RUC data differences. The 45 degree reference line is superimposed.

Figure 3.4 provides a validation plot for the out-of-sample predictions obtained from model (3.2) fitted using RUC data differences. This plot can be compared with Figure 2.9 (right panel) revealing that forecasts from model (3.2) are closer to the observations than those obtained from model (2.9) using Eta-CMAQ data differences.

Table 3.1: Mean square error (MSE), mean absolute error (MAE), empirical coverage and average length of 95% predictive intervals (PI) for model (2.9) and model (3.2) with RUC data differences.

	MSE	MAE	Empirical coverage of 95% PI	Average length of 95% PI
Eta-CMAQ	24.66	3.79	85.5%	13.67
RUC	21.69	3.60	84.8%	13.53

Figures 3.5 and 3.6 show the forecast maps of current 8-hour average ozone level at 12PM on August 8th, 2011 (left panel) and the standard deviation maps (right panel) resulting from the reduced version of model (2.9) using Eta-CMAQ output and model (3.2) using RUC output, respectively. Less uncertainty is clearly associated to the ozone predictions obtained from model (3.2) using the RUC data differences.

Table 3.2: Mean square error (MSE) and mean absolute error (MAE) for model (2.9), model (3.2) using RUC data differences and AIRNow forecasts.

	Pre-Interpolation		Post-Interpolation	
	MSE	MAE	MSE	MAE
Eta-CMAQ	24.43	3.87	41.95	4.95
RUC	17.63	3.26	37.90	4.80
AIRNow	36.39	4.73	45.72	5.35

We can offer a comparison among the predictions obtained from model (2.9) with Eta-CMAQ data differences, model (3.2) fitted using the RUC data differences and the forecasts provided by the AIRNow system. Table 3.2 shows this comparison in terms of MSE and MAE computed on pre-interpolation predictions (at monitoring sites) and post-interpolation forecasts (at the 70 hold-out sites). The table reveals a reduction in MSE and MAE that results in using the proposed strategy with RUC data differences rather than the other two approaches. We achieve roughly 30% and 50% improvement in prediction at fitted sites upon the AIRNow system and model (2.9) with Eta-CMAQ data differences, respectively. The improvement in terms of interpolated predictive performance is sligthly reduced because of the smoothing introduced by the the kriging procedure, but the benefit of our modeling approach developed using the RUC data differences can still be appreciated.

3.3.2 Results using RAP output

In this subsection, we present the results obtained fitting model (3.2) on 24-hour running windows starting at each hour from 12AM to 4PM of August 1st, 2012 and forecasting current 8-hour average ozone level from 2AM to 6PM on August 2nd, 2012. In this case, model fitting and ozone forecasting are evaluated for 17 consecutive windows.

Figures 3.7 shows the current 8-hour average ozone forecast map prediction at 12PM on August 2nd, 2012 (left panel) and the standard deviation map (right panel) resulting from model (3.2). Figure 3.8 provides the scatter plot of the predicted current 8-hour average ozone levels versus the observed values, showing that model (3.2) produces accurate current 8-hour average ozone forecasts.

The MSE, MAE, empirical coverage and average length of the 95% credible

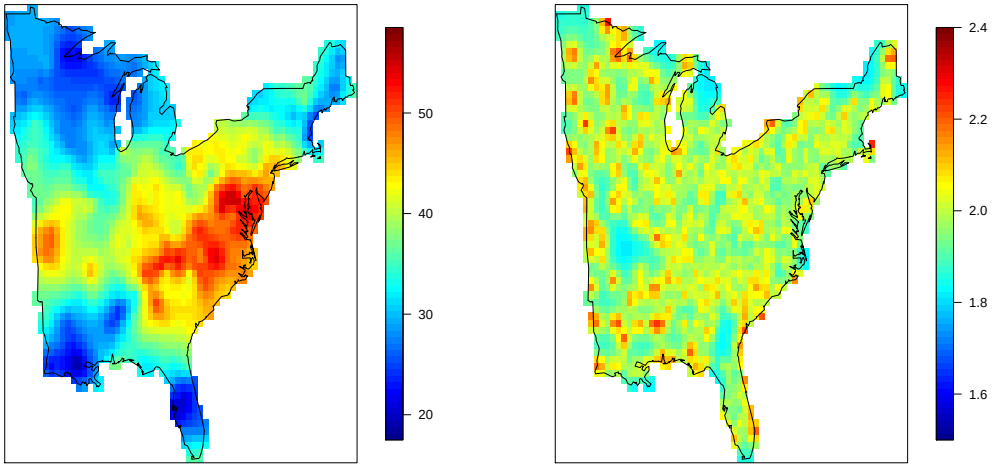

Figure 3.5: Current 8-hour average ozone forecast map (left panel) and standard deviation map (right panel) at 12PM on August 8th, 2011 obtained from model (2.9) using Eta-CMAQ data differences.

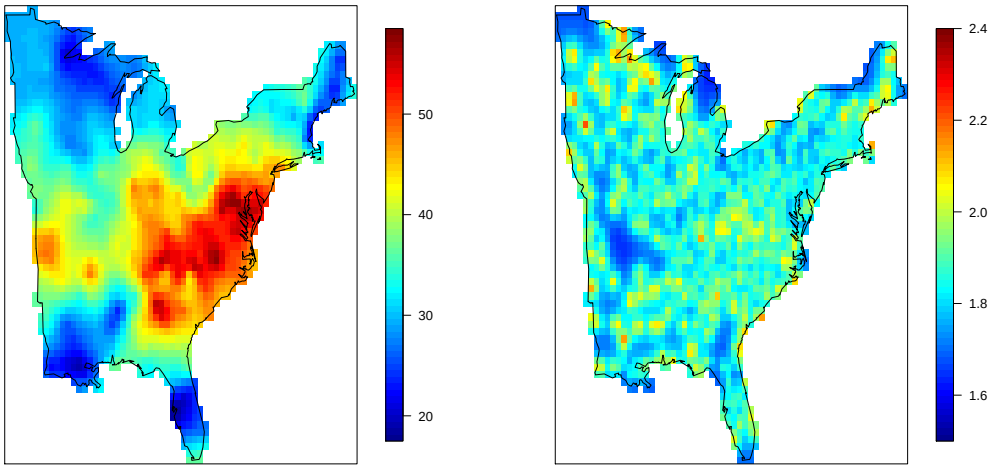

Figure 3.6: Current 8-hour average ozone forecast map (left panel) and standard deviation map (right panel) at 12PM on August 8th, 2011 obtained from model (3.2) using RUC data differences.

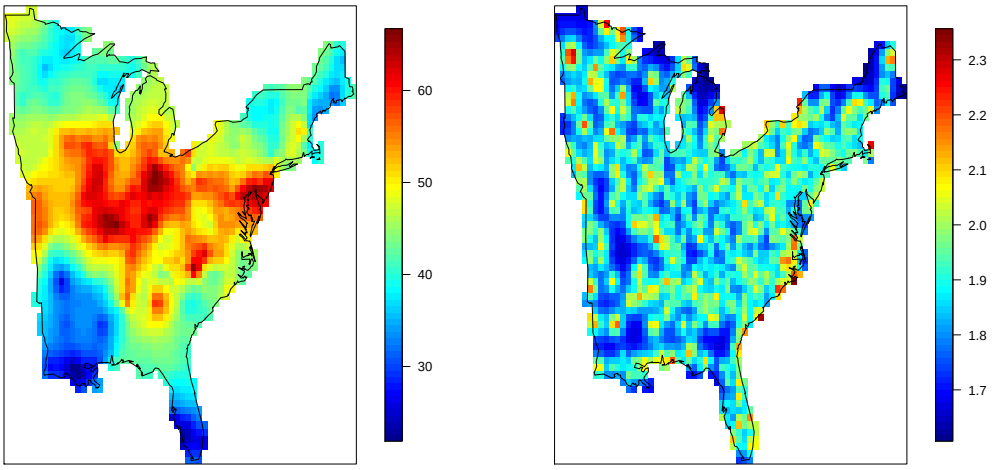

Figure 3.7: Current 8-hour average ozone prediction map (left panel) and standard deviation map (right panel) at 12PM on August 2nd, 2012 obtained from model (3.2) using RAP data differences.

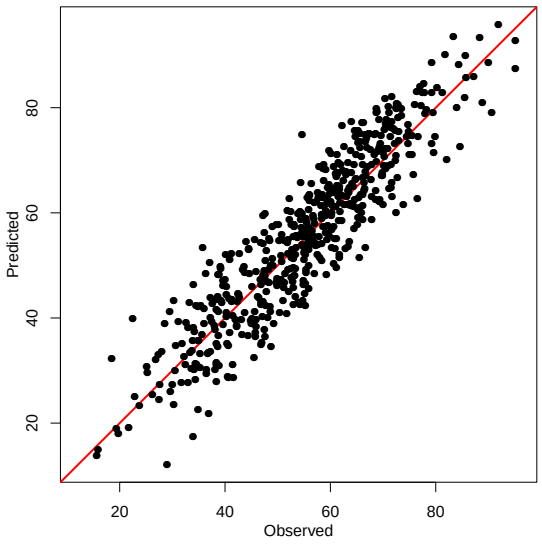

Figure 3.8: Validation plots for out-of-sample predictions when model (3.2) is fitted using RAP data differences. The 45 degree reference line is superimposed.

interval on $70 \times 17 = 1190$ out-of-sample forecasts are 36.36, 4.82, 80% and 16.10, respectively. Table 3.3 offers a comparison of our predictions with those provided by AIRNow system at the fitting sites (pre-interpolation) and at the 70 hold-out sites (post-interpolation). Again, we appreciate the improvement of our modeling approach using RAP data differences in terms of both pre-interpolated (roughly 33%) and post-interpolated (about 18%) predictive performance relative to the current system.

Table 3.3: Mean square error (MSE) and mean absolute error (MAE) for model (3.2) using RAP data differences and AIRNow forecasts.

	Pre-Interpolation		Post-Interpolation	
	MSE	MAE	MSE	MAE
RAP	29.17	4.24	74.69	6.80
AIRNow	43.82	5.31	90.84	7.53

3.4 Short-term ozone predictions in Emilia-Romagna

Model validation for the eastern U.S. in the previous subsections shows improved ozone predictions when we replace ozone estimates from Eta-CMAQ in the downscaler by temperature forecasts produced by weather numerical RUC (RAP) model.

Recently, Bruno and Paci (2013) arrived at similar results studying hourly ozone concentrations in the Emilia-Romagna region (in Italy). The authors proposed a hierarchical spatio-temporal model to exploit different sources of information in order to provide short-term air pollution forecasting in the region. They employed the downscaling approach to combine hourly ozone monitoring data with two alternative numerical model output: ozone estimates from Chimere chemistry-transport model and temperature forecasts from weather forecast Cosmo model. The two systems are currently in use at the regional protection agency of Emilia-Romagna (ARPA-ER) to provide the public with air quality information and weather forecasts, respectively. Also, the orography of the region has been taken into account since the ozone level changes according to the elevation.

Bruno and Paci (2013) showed how the model fitted using temperature predictions from the weather numerical model outperforms the one fitted using the air quality model output. They also noted that the inclusion of the elevation of the

sites in the model improved the ozone forecasting in their study region.

3.5 Summary

In this chapter we have proposed an extension of the real-time downscaler of Chapter 2, here based on real-time temperature data provided as output of a weather numerical model. We have shown how we can substantially improve the current 8-hour ozone forecasting upon the earlier model based on the air quality computer model output. Moreover, less uncertainty is associated with our new predictions.

Our real-time downscaler with temperature is feasible for real-time use and one added advantage of the strategy is its easy and cheap implementation allowed by the free access to the RUC (RAP) output at the NOAA's web site. In fact, the hybrid strategy here proposed is currently being implemented by EPA to provide the public and experts with real-time current 8-hour average ozone predictions. The pseudo algorithm describing each step of the implemented procedure is deferred to Appendix B.

Future work will provide improved real-time regional forecasts at finer resolution than the national ones, say for urban areas of interest, obtained concurrently with the national forecasts.

Chapter 4

Data fusion modeling for map uncertainty

Numerical models are deterministic models developed by several environmental agencies to simulate and predict complex systems, as we illustrated in Chapter 1. Computer model outputs are usually provided as averages over grid cells and, using a large number of grid cells, they can cover large spatial domains and may also have very high temporal resolution. However, numerical model estimates can be biased with unknown calibration. Furthermore, they do not provide any measure of uncertainty associated to their output, since they are derived under a deterministic system. For instance, in Chapter 2 we discussed one of them, the Eta-CMAQ model which has been designed by the EPA to provide air quality information over the conterminous U.S., while in Chapter 3 we described the RUC (RAP) model developed by the NCEP to produce short-term weather forecasts over the CONUS.

In this chapter, we move our attention from calibration and prediction improvement of computer model output to uncertainty quantification. In particular, Section 4.1 presents an overview of statistical methods proposed in the literature for quantifying uncertainty in numerical model output. We also highlight our contribution on this topic. Modeling developments are presented in Section 4.2. In Section 4.3 we first clarify what we mean by uncertainty and then we propose two alternative approaches to model it. Fitting details are discussed in Section 4.4, with computation details deferred to Appendix C. Section 4.5 offers some simulation results, while in Section 4.6 we apply our model to attach uncertainty to RUC model output. A brief summary of this chapter is given in Section 4.7.

4.1 Views of uncertainty in numerical models

Large sources of uncertainty in constructing and employing numerical models do exist. In many applications, these sources can be classified into four basic types: input uncertainty, function uncertainty, model discrepancy and observational error (Cumming and Goldstein, 2010). All of these uncertainty sources can be taken into account by the Bayesian approach and so a wide range of methods have been developed using Bayesian statistics to deal with the uncertainty analysis for complex computer models (Sacks et al., 1989; Craig et al., 1998; O'Hagan, 2006).

Numerical models are often implemented as computer codes and depend on a number of inputs and initial conditions which determine the nature of the output. These inputs represent unknown parameters and the uncertainty about them propagates through the numerical model, inducing uncertainty in the output. The problem concerning how input uncertainty propagates through to the model solution, is usually referred in engineering and applied mathematics literature as to forward problem. Instead, a general statistical framework has been presented by Givens et al. (1993) and Raftery et al. (1995) for mapping from a set of input parameters to a set of model outputs, the so-called Bayesian synthesis. The approach consists of establishing a joint probability distribution on the model inputs and outputs and then restricting this to a subspace defined by the model in order to obtain the joint posterior distribution, from which inferences are drawn. Also, statistical methods have been proposed to handle the sensitivity analysis which is concerned with understanding how the model output is influenced by changes in the model inputs (e.g. Draper et al., 1999; Oakley and O'Hagan, 2004).

Parameter uncertainty is a form of epistemic uncertainty, deriving from our lack of knowledge about the real system. A second form of epistemic uncertainty is structural uncertainty which is introduced by scientific choices of model design and development. Although numerical models are deterministic, i.e. no random components are considered along model development, their predictions are subject to error because any model is a simplification of reality. So, even in case of no parameter uncertainty, model output cannot ever equal the "true" value of the process of interest and this discrepancy is the well-known *model inadequacy* (Kennedy and O'Hagan, 2001). Model discrepancy can be evaluated by comparing model output with observations. Customarily, researchers make use of observations from the process to deal with the calibration question and, in this case, they should take into account also the observational error.

Structural uncertainty can be also quantified by analyzing multi-model ensembles. In this case, the output consists of different versions of a numerical model, i.e. a model is run several times with different initial conditions and/or model physics. Statistical approaches for quantifying uncertainty with ensembles have recently received considerable attention (see e.g. Gneiting et al. 2005; Raftery et al. 2005; Berrocal et al. 2007; Sloughter et al. 2007, 2010; Kleiber et al. 2011; Sloughter et al. 2013). Raftery, Gneiting and co-authors developed a statistical approach for post-processing ensembles based on Bayesian model averaging (BMA), which is a standard method for combining predictive distributions from different sources. Bayesian hierarchical approaches are also proposed by Smith et al. (2009), Tebaldi and Smith (2010) and Di Narzo and Cocchi (2010) to tackle ensemble weather forecasting and uncertainty assessment.

Despite uncertainty quantification is a pressing research issue, not much has been said about statistical methods for attaching uncertainty to model output when we do not have information about how such deterministic predictions are created. Indeed, our proposal builds upon the notion of uncertainty introduced by Ghosh et al. (2012) when numerical models are unavailable, rather only deterministic outputs at some spatial resolution are provided. In other words, we do not know how the deterministic surfaces have been developed, instead they come from some "black box" which we know nothing about. Ghosh et al. (2012) proposed a general Bayesian approach to associate uncertainties with deterministic interpolated surfaces, using some external *validation* data collected independently over the same spatial domain as the deterministic map. Although numerical models produce deterministic surfaces, we highlighted above that the output will not ever be the "true" value of the process. In this framework, given the truth and the model output, the associated error is not stochastic. But, under suitable stochastic modeling, this error can be reinterpreted as a random unknown which we can infer about using a Bayesian specification within the data fusion setting. Making inference about the uncertainty might sound odd in usual statistical speaking, but, again, here we want to attach some uncertainty measure to deterministic output and so inference about such model-based uncertainty is needed.

Uncertainty maps associated with numerical model output provide useful information to guide environmental agencies in thinning and improving computer models. Furthermore, when we use the model output as predictor of some environmental variable (see for instance the downscaler in the previous chapters) we might be interested to evaluate how these uncertainties propagate from the model

output to the response forecasting. In this contest, spatial and spatio-temporal errors-in-variables models has been proposted, among others, by Van de Kassteele and Stein (2006) and Cameletti et al. (2011).

The contribution of this chapter is to develop a Bayesian hierarchical model to provide spatially smoothed uncertainty associated with numerical model output, regardless of how it was created. We can learn about such uncertainty through stochastic data fusion modeling using some external validating data. We also take into account the change of support problem, which arises from the spatial misalignment between the numerical model output and the validation data. Statistical methods for blending observed data with model output has been deeply discussed in the previous chapters, showing the benefit of data fusion modeling to improve the forecasting. Conversely, our objective here is not the calibration of numerical model predictions rather we are interested in spatially smoothed uncertainties associated with the output. To attach such varying uncertainty across grid cells we offer a fully model-based approach that can be used to assign uncertainty to any deterministic surface. Here, we apply our Bayesian model to obtain the uncertainty map associated with temperature output provided by RUC weather model over the northeastern U.S.. The validation data set consists of temperature measurements collected at monitoring stations operating in the same study region.

4.2 Data fusion model

Let $R(A_i)$ denote the numerical model output (e.g. temperature predictions from RUC model) over grid cell A_i, $(i = 1, \ldots, I)$. As usual, we interpret $R(A_i)$ as an average value over cell A_i, i.e. $|A_i|^{-1} \int_{A_i} R(s)ds$, see (1.1). First, we specify a measurement error model (MEM)[1] for the numerical model output $R(A_i)$ relative to the truth, that is:

$$R(A_i) = \widetilde{R}(A_i) + \varepsilon_r(A_i) \tag{4.1}$$

where $\widetilde{R}(A_i)$ is the underlying process which represents the 'true' average value for A_i and we assume $\varepsilon_r(A_i) \sim N\big(0, \sigma_r^2(A_i)\big)$ independently $\forall i = 1, \ldots, I$. The true average value $\widetilde{R}(A_i)$ arises from a Gaussian Markov Random Field (GMRF) equipped with a conditionally autoregressive structure (CAR) (Besag, 1974; Banerjee et al.,

[1]The measurement error model is also known as error-in-variables model; see for instance Fuller (1987) and references therein.

2004) that is:

$$\widetilde{R}(A_i) \mid \left\{\widetilde{R}(A_{i'}) : i' \neq i\right\} \sim N\left(\sum_{i' \sim i} \frac{\widetilde{R}(A_{i'})}{w_i}, \frac{\tau^2}{w_i}\right) \tag{4.2}$$

where $i' \sim i$ identifies the cell $A_{i'}$ adjacent to cell A_i and w_i is the number of neighbors of cell A_i.

Let $V(\mathbf{s}_j)$ be the temperature at location $\mathbf{s}_j$, $(j = 1, \ldots, n)$ gathered from independent station data over the same region as the output, and $\widetilde{V}(\mathbf{s}_j)$ denotes the true value at $\mathbf{s}_j$. For the validation data, we assume a spatial model given by:

$$V(\mathbf{s}_j) = \widetilde{V}(\mathbf{s}_j) + \varepsilon_v(\mathbf{s}_j) \tag{4.3}$$

where $\boldsymbol{\varepsilon}'_v = (\varepsilon_v(\mathbf{s}_1), \ldots, \varepsilon_v(\mathbf{s}_n))$ is a zero-mean Gaussian process equipped with a spatial exponential correlation function, i.e. $\boldsymbol{\varepsilon}_v \sim N\big(\mathbf{0}, \sigma_v^2\, H(\phi)\big)$.

Finally, we address the change of support problem between the station data and the numerical model output by assuming a further measurement error model for $\widetilde{V}(\mathbf{s}_j)$. We avoid the integration problem associated with scaling from point to grid level by employing the downscaling approach which associates to each site $\mathbf{s}_j$ the grid cell A_i that contains $\mathbf{s}_j$. Then, for each $j = 1, \ldots, n$ belonging to grid cell A_i we have:

$$\widetilde{V}(\mathbf{s}_j) = \widetilde{R}(A_i) + \varepsilon_{\tilde{v}}(\mathbf{s}_j) \tag{4.4}$$

where $\varepsilon_{\tilde{v}}(\mathbf{s}_j)$ are independent $N(0, \sigma^2_{\tilde{v}})$.

Figure 4.1 shows a graphical representation of the model described above. In order to illustrate how we can learn about the uncertainty through such stochastic modeling, we clarify in the next section what we mean by uncertainty when dealing with a deterministic output without information about how it was created.

4.3 Defining and modeling uncertainty

Recall that our primary goal is to provide a measure of uncertainty associated with numerical model output over grid cells. To clarify what we mean by uncertainty, we might concentrate about the "true" error, say $R(A_i) - \widetilde{R}_{true}(A_i)$, where $\widetilde{R}_{true}(A_i)$ is the true average value for the numerical model output over cell A_i. When this error is small for a grid cell, it implies small uncertainty associated to the numerical model prediction. Conversely, if the error is large then we would imagine high uncertainty for such cell.

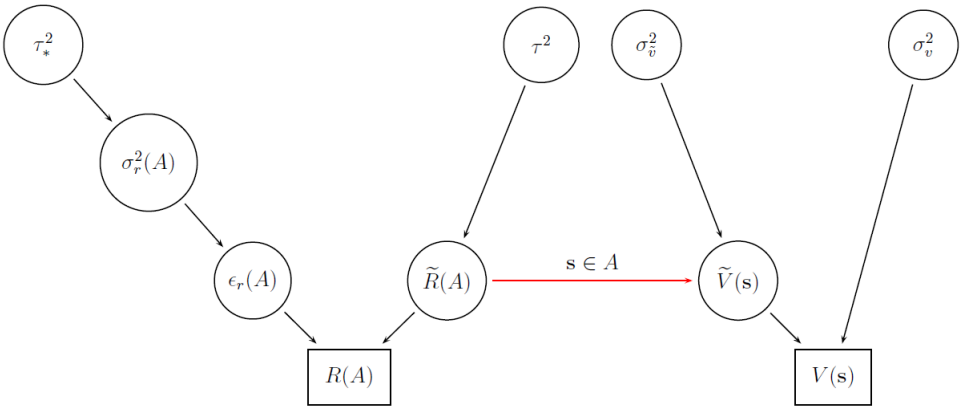

Figure 4.1: Graphical representation of model (4.1) - (4.4) under prior (4.7).

To inform about the true error, we might compute the *observed* residuals $R(A_i) - V(\mathbf{s}_j)$, i.e. compare the numerical model output with the validation data for each grid cell that contains a site. Then, high "disagreement" between $R(A_i)$ and $V(\mathbf{s}_j)$ for $s_j \in A_i$ suggests high uncertainty in A_i. Conversely, we expect small uncertainty at grid cells where the disagreement between the numerical model output and the observed temperature is low. However, two main issues arise when we look at the observed residuals: first, the comparison between the average $R(A_i)$ with the point-level measurement $V(\mathbf{s}_j)$ is unfair because of the different spatial support of the two data sources, i.e COSP thoroughly discussed in Chapter 1. Second, the observed residuals are available only for grid cells where sites lie, while our goal is to attach uncertainty to every grid cell. To accomplish that, we consider instead the so-called *realized* residuals (Zellner, 1975; Chaloner and Brant, 1988; Chaloner, 1994), that is $\epsilon_r(A_i) = R(A_i) - \widetilde{R}(A_i)$ from (4.1). To further clarify, the true error for $R(A_i)$ is not known and, as usual within the Bayesian framework, we model unknowns as random and look at their posterior distributions for inference. In fact, under the specification above, we take $\widetilde{R}(A_i)$ as the model for the truth and we look at the posterior distribution of the realized residuals, $\big[\epsilon_r(A_i)|Data\big]$. The posterior variance $var\left(\epsilon_r(A_i) \mid Data\right)$ provides the desired uncertainty, varying across grid cells.

We can obtain our local uncertainties by composition sampling, i.e. drawing posterior samples of $\epsilon_r(A_i)$ and then compute their variance. Alternatively, we can obtain local uncertainties as the posterior means $E\big(\sigma_r^2(A_i) \mid Data\big)$. Indeed, under

model (4.1) - (4.4), we have

$$\begin{aligned} var\left(\epsilon_r(A_i) \mid Data\right) =& E\left[var\left(\epsilon_r(A_i) \mid \widetilde{R}(A_i), \sigma_r^2(A_i), Data\right)\right] \\ &+ var\left[E\left(\epsilon_r(A_i) \mid \widetilde{R}(A_i), \sigma_r^2(A_i), Data\right)\right] \end{aligned} \tag{4.5}$$

The second term in (4.5) is clearly 0 and the first reduces to $E\big(\sigma_r^2(A_i) \mid Data\big)$. So, the Rao-Blackwellized estimates can be directly obtained by computing the mean of the posterior sampled draws of $\sigma_r^2(A_i)$.

Since we are interested in the posterior distribution of $\sigma_r^2(A_i)$, the specification of its prior distribution represents a crucial step. The prior modeling of the σ_r^2's is covered in the next two subsections.

4.3.1 Modeling via hierarchical approach

A naive way to model, through a prior, the variances of interest would be to assume that all $\sigma_r^2(A_i)$ are independently and identically distributed according to an inverse gamma $IG(a, b)$. Notice that the estimates can be sensitive to different choices of the scale parameter b, but we do not have any knowledge about the size of the uncertainties. Moreover, if a and b are fixed in advance, under the independence assumption, the information we might have about the set of variances $\left\{\sigma_r^2(A_{i'}), i' \neq i\right\}$ is not of help to estimate $\sigma_r^2(A_i)$. In other words, a form of borrowing strength across grid cells has to enter in our specification. So, we add a further level to the hierarchy of our Bayesian model so that all variances $\sigma_r^2(A_i)$ are samples from the same prior distribution, that is

$$\sigma_r^2(A_i) \sim IG(a, b_*) \tag{4.6}$$

with the scale parameter b_* to estimate, while a remains fixed. This extra layer has the effect of smoothing out the estimates of $\sigma_r^2(A_i)$, and substantially reducing the sensitivity as well.

4.3.2 Modeling via spatial smoothing

As described in the beginning of this section, we look at the posterior variance of the realized residuals $\epsilon_r(A_i)$ to obtain the desired local uncertainties, attaching high uncertainty to grid cells for which we suppose large differences between the model output and the true value. In addition, for a large realized residual at grid cell A_i, we expect a similar behavior in its neighborhood, i.e. we envision that changes in

variance occur smoothly over space. In other words, we figure out some spatial smoothness for the uncertainty associated with the numerical model output, based on the neighborhood structure of the grid cells. We formalize this belief assuming a CAR process for the logarithm[2] of the latent variances $\sigma_r^2(A_i)$ that is,

$$\log\big(\sigma_r^2(A_i)\big)\big|\ \big\{\log\big(\sigma_r^2(A_{i'})\big) : i' \neq i\big\} \sim N\left(\sum_{i'\sim i} \frac{\log\big(\sigma_r^2(A_{i'})\big)}{w_i},\ \frac{\tau_*^2}{w_i}\right) \tag{4.7}$$

where, following the notation in Section 4.2, $i' \sim i$ identifies the cell $A_{i'}$ adjacent to cell A_i and w_i is the number of neighbors of A_i. The logCAR prior model in (4.7) is analogous to the spatial stochastic volatility approach developed by Yan (2007) and revised by Reich and Hodges (2008) to capture spatial clustering in heteroscedasticity. Model (4.7) enables us to explicitly impose a spatially varying structure on the variances, allowing for the borrowing of strength across grid cells and inducing local spatial smoothing to uncertainty estimates towards their neighboring grid cells.

4.3.3 Comparing uncertainty assignments

The comparison of alternative models is traditionally performed with attention to uncertainty reduction, which is not really our objective. To further clarify, we consider the "true" error introduced at the beginning of this section, i.e. $R(A_i) - \widetilde{R}_{true}(A_i)$, where $\widetilde{R}_{true}(A_i)$ is the "true" average value for A_i. Again, with the specification above we take $\widetilde{R}(A_i)$ as the model for the truth and we look at the posterior variances of the realized residuals to obtain our local uncertainties. A general balanced criterion needs to account for the trade-off between uncertainty and bias in $\widetilde{R}(A_i)$ that is

$$\begin{aligned} R(A_i) - \widetilde{R}_{true}(A_i) = &\Big(R(A_i) - \widetilde{R}(A_i)\Big) \\ &+ \Big(\widetilde{R}(A_i) - \widetilde{R}_{true}(A_i)\Big). \end{aligned} \tag{4.8}$$

Therefore, we compare models considering both the posterior variance arising from the first term in (4.8) and the squared bias associated with the second term. As pointed out by Ghosh et al. (2012), to inform about bias with available data, we can only compare $\widetilde{R}(A_i)$ with validation data $V(\mathbf{s}_j)$, for each $\mathbf{s}_j \in A_i$. Then, the balanced loss idea yields to the criterion

$$\frac{1}{I}\sum_{i=1}^{I} var\big[\epsilon_r(A_i)\,|\,data\big] + \frac{c}{n}\sum_{j=1}^{n} E\big[\big(\widetilde{R}(A_i) - V(\mathbf{s}_j)\big)^2 |\,data\big] \tag{4.9}$$

[2]The logarithm ensures the positivity of the variances.

where c indicates the relative regret for the two losses (Gelfand and Ghosh, 1998). This induces to choose the model leading the smallest value of (4.9).

4.4 Fitting details

Since it is not possible to consistently estimate the decay and variance parameter in a spatial model with a covariance function belonging to the Matérn family (Zhang, 2004) as the exponential covariance function we employed, we fix the decay parameter and we put a prior distribution on $\sigma^2_{\tilde{v}}$. On the variance parameters σ^2_v and $\sigma^2_{\tilde{v}}$ we place conjugate inverse gamma priors $IG(a_\sigma, b_\sigma)$ where we take a_σ and b_σ so that

$$\frac{b_\sigma}{a_\sigma - 1} = \frac{MSE}{2} \quad \text{and} \quad \frac{b^2_\sigma}{(a_\sigma - 1)^2(a_\sigma - 2)} = 10^2$$

where MSE is the mean square error arising from a simple linear regression of $V(s_j)$ on $R(A_i)$ for each $s_j \in A_i$. The prior distributions for τ^2 and τ^2_* are specified as independent proper inverse gamma distributions $IG(a_\tau, b_\tau)$. Recently, Sørbye and Rue (2013) proposed a general approach for choosing the prior distribution for the precision parameter of intrinsic GMRF, according to the specific type of GMRF used. The authors suggested to select this prior by mapping the precision parameter to the marginal standard deviation of the model, under linear constraints. In their applications, they showed that there were no significant differences in the estimated spatial effects using the default and the scaled priors for the precision parameter of a CAR process and their results were not sensitive to tuning of the prior. Due to the insensitivity to different choices of a_τ and b_τ, in our implementation we take $a_\tau = 2$ and $b_\tau = 1$, implying that these variance components have prior mean 1 and infinite variance.

Finally, a prior distribution for the parameter b_* in (4.6) is needed. We assume that this parameter is sampled from a uninformative gamma prior $Ga(c, d)$, with $c = d = 0.01$.

Define $\mathbf{R} = \big(R(A_1), \ldots, R(A_I)\big)'$ and $\mathbf{V} = \big(V(\mathbf{s}_1), \ldots, V(\mathbf{s}_n)\big)'$; then the full distributional specification of model (4.1) - (4.4) using the logCAR prior model is given by:

$$\left[\mathbf{R} \,|\, \widetilde{\mathbf{R}}, \boldsymbol{\sigma}^2_r\right] \left[\mathbf{V} \,|\, \widetilde{\mathbf{V}}, \sigma^2_v\right] \left[\widetilde{\mathbf{V}} \,|\, \widetilde{\mathbf{R}}, \sigma^2_{\tilde{v}}, \phi\right] \left[\widetilde{\mathbf{R}} \,|\, \tau^2\right] \left[\boldsymbol{\sigma}^2_r \,|\, \tau^2_*\right] \tag{4.10}$$

where $\widetilde{\mathbf{R}} = \big(\widetilde{R}(A_1), \ldots, \widetilde{R}(A_I)\big)'$, $\widetilde{\mathbf{V}} = \big(\widetilde{V}(\mathbf{s}_1), \ldots, \widetilde{V}(\mathbf{s}_n)\big)'$ and $\boldsymbol{\sigma}^2_r = \big(\sigma^2_r(A_1), \ldots, \sigma^2_r(A_I)\big)'$. Along with the prior distributions for all the unknown

parameters, the Bayesian hierarchical model is completely specified. The model is fitted using Markov Chain Monte Carlo (MCMC) algorithm; details are deferred to Appendix C.

4.5 Simulation study

In this section, we perform simulation examples to illustrate the performance of the two approaches described in Sections 4.3.1 and 4.3.2. Since our attached uncertainty is model-based, it is not trivial to evaluate the modeling performance, as we discussed in the previous section on comparing uncertainty assignments. Via the simulation study we gain knowledge about the truth and so about the true errors for assessing map uncertainty.

The simulation design is built from several sampling/fitting combinations allowing the investigation of different features. Simulation experiments are performed through the following steps:

1. We consider a unit square uniformly divided into 900 grid cells.
2. Using the centroids of the grid cells, we generate $\widetilde{R}(A_i)$ $(i = 1, \ldots, I = 900)$ from the CAR model[3] of expression (4.2) where $\tau^2 = 1$.
3. We generate $R(A_i)$ using relation (4.1). We consider different choices for variances of interest $\sigma_r^2(A_i)$:
   - 3.1. $\sigma_r^2(A_i) = 1$, $\forall A_i$;
   - 3.2. $\sigma_r^2(A_i) \sim IG(a, b_*)$, with $b_* = 0.5$;
   - 3.3. $\sigma_r^2(A_i) \sim \text{logCAR}(\tau_*^2)$, with $\tau_*^2 = 0.5$.
4. Then, two different sets of 200 locations are randomly generated within the unit square (hereafter, "Coords1" and "Coords 2").
5. For each location $\mathbf{s}_j$, $(j = 1, \ldots, n = 200)$ belonging to a grid cell A_i, we generate $\widetilde{V}(\mathbf{s}_j)$ using relation (4.4) with $\sigma_{\widetilde{v}}^2 = 1$ and fixed value of decay parameter ϕ. In particular, we set $\phi = 2.8$ or $\phi = 11.25$ corresponding, respectively, to spatial ranges of roughly 80% and 20% of the maximum distance over the region. We also consider the addition of some bias to (4.4) in a portion of the region (top right) when we do the sampling.

[3]The CAR model is not proper, so we add a small constant to make the precision matrix non-singular.

Table 4.1: Sampling/fitting simulation design.

Scenario	Sampling				Fitting[a]
	$\sigma_r^2(A_i)$	Bias to (4.4)	ϕ	Validation data	$\hat{\phi}$
(a)	1, $\forall(A_i)$	NO	2.8	"Coords 1"	2.8
(b)	1, $\forall(A_i)$	YES	2.8	"Coords 1"	2.8
(c)	1, $\forall(A_i)$	NO	11.25	"Coords 1"	11.25
(d)	1, $\forall(A_i)$	YES	11.25	"Coords 1"	11.25
(e)	1, $\forall(A_i)$	NO	2.8	"Coords 1"	11.25
(f)	1, $\forall(A_i)$	NO	2.8	"Coords 2"	2.8
(g)	from IG$(a, 0.5)$	NO	2.8	"Coords 1"	2.8
(h)	from logCAR(0.5)	NO	2.8	"Coords 1"	2.8

[a] In fitting, we consider both priors (4.6) and (4.7).

6. Finally, the validation data $V(\mathbf{s}_j)$ are generated from equation (4.3) where $\sigma_v^2 = 1$.

Given the sampling scheme described in the previous steps, we fit model (4.1) - (4.4) under both prior models (4.6) and (4.7). Moreover, we allow for the case when we fit the model setting the spatial decay parameter ϕ far away from its true value. From all possible sampling/fitting combinations, we consider the scenarios listed in Table 4.1.

We provide, as examples, the sampling design for scenarios (g) and (h) in Figures 4.3 and 4.4. In such figures, the simulated local uncertainties (standard deviations) are shown in the left panel. In the middle panel we have the true errors, i.e. the differences between the simulated $R(A_i)$ and the simulated $\widetilde{R}_{true}(A_i)$. The right panel shows the observed residuals, i.e. $R(A_i) - V(\mathbf{s}_j)$, for each $\mathbf{s}_j \in A_i$ for scenario (g). True errors appear higher at grid cells where the simulated uncertainties are higher, coherently with the idea of uncertainty described in Section 4.3. It is not straightforward to argue similar behavior looking the observed residuals compared to local uncertainties rather it is worth to look at the true errors that we stress can be calculated only a situation like simulation study, where the data generator process is known.

Posterior summaries of model parameters for all scenarios are given in Table

4.2. The table reveals slight differences between the two approaches with respect to parameter recovery across different scenarios, showing that we can however recover the true values of these parameters. Figures 4.5 - 4.12 show the 900 local predicted uncertainties associated with the simulated gridded maps under both the hierarchical approach, (4.6) (left panels) and logCAR model, (4.7) (right panels). In such figures, posterior uncertainty maps resulting from the logCAR prior are smoother relative to those obtained from the hierarchical approach, as we expected. Moreover, the estimated uncertainties obtained using model (4.7) are closer to their true values than the posterior uncertainties obtained using the inverse gamma prior in (4.6).

When we look at the whole set of left and right panels respectively, we note slight differences between the posterior uncertainty maps across the scenarios for both approaches. In particular, the comparison between Figure 4.5 and Figure 4.6 shows what happens when we add some bias to (4.4) in a portion of the region when we do the sampling and we fit a measurement error model which ignores this aspect. The comparison reveals little difference except for somewhat higher estimated uncertainties under scenario (b). We reach the same conclusion by comparing Figure 4.7 and Figure 4.8 corresponding to scenarios (c) and (d), respectively. Comparing the results under scenario (a) and scenario (c), we can learn about the smoothness imparted to σ_r^2's as we change the spatial decay parameter ϕ. From Figure 4.5 and 4.7 it is hard to measure the amount of smoothing since only slight differences are revealed by the figures.

The comparison between Figure 4.9 and 4.5 shows that the posterior variances are not very sensitive to the case when we fit the model setting the spatial decay parameter ϕ far away from its true value, even if we note that we cannot longer recover the true value of parameter σ_v^2 under scenario (e).

Comparing the results under scenario (a) and scenario (f), we can evaluate the effect of the locations of the validation data on the estimated uncertainty. In fact, we assume that the external observed data are independently gathered from the gridded data over the same region. Our stochastic modeling does not attach higher uncertainty to numerical model output corresponding to grid cells that contain sites. Accordingly, Figure 4.5 and Figure 4.10 show that the posterior local uncertainties seem not to be affected by the location of the validation sites.

Finally, Figure 4.11 and 4.12 allow to investigate the performance of our modeling in recovering the true uncertainties also when we do the sampling under further schemes. Recall that the true values of the σ_r^2's under scenario (g) and (h)

are plotted in left panels of Figure 4.3 and Figure 4.4, respectively. The logCAR prior model performs pretty well, leading to smoothed uncertainties quite close to their true values. Instead, the hierarchical approach seems to fail in recovering the true local uncertainties when we do the sampling under scenario (h).

Table 4.3 shows the comparison between the two alternative approaches via criterion (4.9) for all scenarios (here, $c = 1$). In the simulation study, the true average $\widetilde{R}_{true}(A_i)$ is available and it replaces $V(\mathbf{s}_j)$ in the criterion. Again, the table reveals little differences between the two approaches for modeling, a priori, the variances of interest. In general, the logCAR prior model yields to smaller values of criterion (4.9), suggesting slight improved performance upon the hierarchical approach.

4.6 Attaching uncertainty to RUC output

We finally turn to our objective to give a measure of uncertainty associated with numerical model output. Here, we illustrate the data fusion model of Section 4.2 to quantify the uncertainty associated with RUC model output. From Section 3.1, we recall that RUC model produces weather short-term predictions for the conterminous U.S. over grid cells of size 13×13 kilometers. As an example, we consider daily temperature forecasts on August 7th, 2011 obtained as average of 24 hourly temperature forecasts (°F) provided by RUC model from 00:00 to 23:00 on August 7th over the northeastern United States, see Figure 4.2. There are 3,862 RUC grid cells spanning our study region. Moreover, land-based station data over U.S. is provided by the NOAA's National Climatic Data Center (NCDC). Here, we consider 24-hour averages of hourly temperature collected from 163 stations operating in the study region for the same period (Figure 4.2).

We fit model (4.1) - (4.4) under both prior models (4.6) and (4.7). Regarding the spatial decay parameter ϕ, explanatory analysis suggested to fix the parameter at roughly 60% of the maximum distance over the study region. Posterior summaries of the unknown parameters are presented in Table 4.4. The posterior means of true temperature $\widetilde{R}$'s with variances specification (4.6) and (4.7) are shown in Figure 4.13 in left and right panel, respectively. The comparison of the two panels reveals little differences between posterior means of the $\widetilde{R}$'s under the two approaches except for somewhat smoother estimates under prior (4.7). Figure 4.14 shows the estimated uncertainty maps associated with RUC output under prior (4.6) and (4.7) in left and right panel, respectively. The uncertainty map

Table 4.2: True values, posterior mean and 95% credible intervals of model parameters under all scenarios.

Scenario	Parameters	True value	IG(a, b_*)	logCAR(τ_*^2)
	σ_v^2	1	1.099 (0.474, 2.086)	1.157 (0.502, 2.233)
	σ_v^2	1	0.716 (0.471, 0.991)	0.762 (0.513, 1.040)
(a)	τ^2	1	1.314 (0.746, 2.074)	0.959 (0.549, 1.474)
	b_*		1.327 (1.088, 1.564)	
	τ_*^2			0.216 (0.117, 0.376)
	σ_v^2	1	1.231 (0.555, 2.289)	1.297 (0.586, 2.426)
	σ_v^2	1	0.692 (0.448, 0.970)	0.741 (0.493, 1.018)
(b)	τ^2	1	1.292 (0.733, 2.041)	0.893 (0.511, 1.364)
	b_*		1.332 (1.095, 1.570)	
	τ_*^2			0.221 (0.106, 0.391)
	σ_v^2	1	1.075 (0.482, 1.829)	1.134 (0.520, 1.921)
	σ_v^2	1	0.678 (0.317, 1.128)	0.713 (0.335, 1.163)
(c)	τ^2	1	1.385 (0.766, 2.205)	0.964 (0.532, 1.514)
	b_*		1.308 (1.062, 1.551)	
	τ_*^2			0.222 (0.101, 0.416)
	σ_v^2	1	1.233 (0.626, 1.975)	1.301 (0.666, 2.084)
	σ_v^2	1	0.610 (0.279, 1.039)	0.643 (0.292, 1.075)
(d)	τ^2	1	1.366 (0.754, 2.179)	0.941 (0.527, 1.474)
	b_*		1.312 (1.066, 1.556)	
	τ_*^2			0.211 (0.096, 0.402)
	σ_v^2	1	1.112 (0.672, 1.653)	1.184 (0.717, 1.758)
	σ_v^2	1	0.418 (0.204, 0.709)	0.444 (0.214, 0.736)
(e)	τ^2	1	1.327 (0.743, 2.102)	0.926 (0.519, 1.428)
	b_*		1.321 (1.080, 1.561)	
	τ_*^2			0.218 (0.100, 0.439)
	σ_v^2	1	0.860 (0.377, 1.638)	0.848 (0.369, 1.660)
	σ_v^2	1	0.925 (0.670, 1.220)	0.945 (0.694, 1.235)
(f)	τ^2	1	1.159 (0.639, 1.878)	0.895 (0.479, 1.388)
	b_*		1.378 (1.145, 1.609)	
	τ_*^2			0.21 (0.104, 0.365)
	σ_v^2	1	1.177 (0.509, 2.239)	1.213 (0.525, 2.338)
	σ_v^2	1	0.754 (0.505, 1.033)	0.770 (0.518, 1.046)
(g)	τ^2	1	0.989 (0.652, 1.411)	0.865 (0.558, 1.257)
	b_*	0.5	0.547 (0.437, 0.660)	
	τ_*^2			1.361 (0.505, 2.516)
	σ_v^2	1	1.094 (0.476, 2.090)	1.092 (0.470, 2.144)
	σ_v^2	1	0.859 (0.603, 1.151)	0.876 (0.615, 1.162)
(h)	τ^2	1	1.053 (0.734, 1.446)	0.822 (0.587, 1.108)
	b_*		0.244 (0.168, 0.317)	
	τ_*^2	0.5		0.322 (0.136, 0.634)

Table 4.3: Criteria (4.9) for different scenarios under the two alternative prior models.

Scenario	IG(a, b_*)	logCAR(τ_*^2)
(a)	1.226 + 0.162 = 1.389	1.072 + 0.158 = 1.230
(b)	1.231 + 0.163 = 1.394	1.101 + 0.159 = 1.261
(c)	1.210 + 0.164 = 1.374	1.079 + 0.159 = 1.238
(d)	1.214 + 0.164 = 1.378	1.084 + 0.159 = 1.244
(e)	1.222 + 0.161 = 1.383	1.085 + 0.157 = 1.242
(f)	1.266 + 0.151 = 1.417	1.010 + 0.152 = 1.252
(g)	0.531 + 0.116 = 0.647	0.510 + 0.120 = 0.630
(h)	0.231 + 0.010 = 0.331	0.240 + 0.098 = 0.338

resulting from the logCAR prior model on σ_r^2's reveals the spatial variation, as we expected. It also worth noting from Figure 4.13 and 4.14 that the estimated $\widetilde{R}$'s and the attached uncertainties have different spatial patterns. In fact, high values in $\widetilde{R}(A_i)$ does not necessarily imply high uncertainty, rather high uncertainty is linked to large realized residuals. Moreover, we have no reason to believe that the uncertainty should be proportional in some way to true temperature neither that larger variances should be associated with larger responses.

Regarding the comparison between the two alternative approaches we note the following: when we set $c = 1$, criterion (4.9) under inverse gamma prior is 2.41 while under the logCAR prior the criterion results to be 2.43. However, when c increases, i.e increasing weight of the second term in (4.9), we see slight improvement of approach (4.7) upon the approach (4.6), e.g. for $c = 3$ we have 7.23 and 7.14 for prior (4.6) and (4.7), respectively.

4.7 Summary

In this chapter we have developed a hierarchical model to fuse a deterministic output and some validation data in order to quantify the uncertainty associated with the model output. We have allowed for spatially smoothed error variances via a logCAR prior model (4.7) and we have compared this approach against the simpler hierarchical approach (4.6).

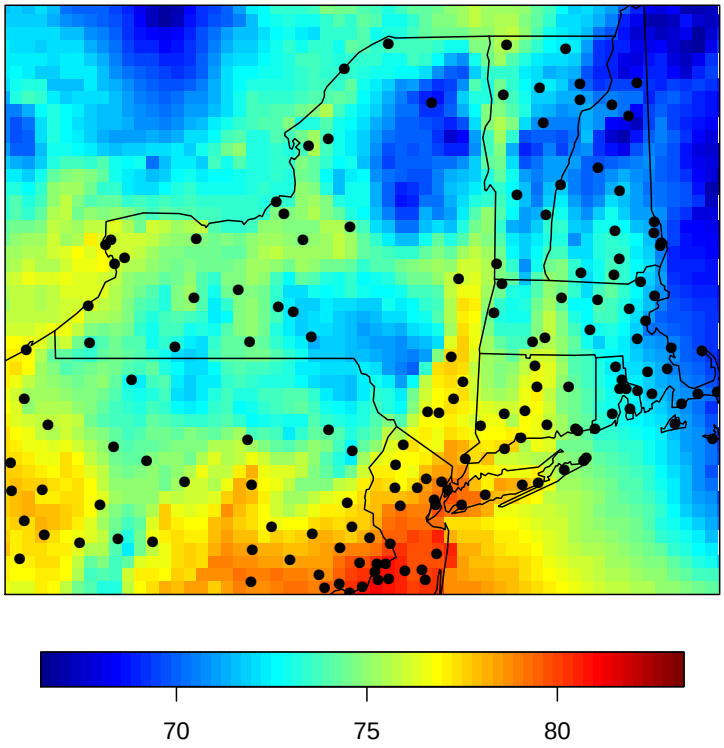

Figure 4.2: Temperature stations (black dots) and daily RUC output on August 7th, 2011 over Northeastern US.

Future work will find us dealing with the calibration issue jointly to the uncertainty assessment of numerical models output. Extension can also concern joint forecast maps, e.g. temperature and precipitation and attaching uncertainty through joint stochastic modeling. Regarding the attached uncertainty to RUC output, we are also interested in seasonal uncertainties, say winter or summer uncertainty maps. Finally, we recall that the RAP weather forecast model took the place of the RUC model on May 1, 2012. Therefore, we are interested to compare uncertainty maps associated with RUC output against those attached to RAP predictions.

Table 4.4: Posterior summary of model parameters.

Parameters	IG(a, b_*)	logCAR(τ_*^2)
σ_v^2	0.931 (0.285, 2.240)	0.902 (0.283, 2.118)
$\sigma_{\tilde{v}}^2$	1.719 (1.235, 2.277)	1.701 (1.231, 2.242)
τ^2	0.791 (0.757, 0.826)	0.728 (0.694, 0.764)
b_*	0.0004 (0.0001, 0.0010)	
τ_*^2		10.868 (7.428, 14.91)

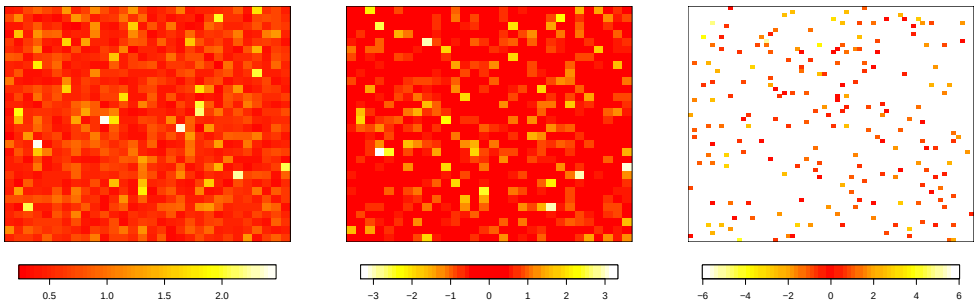

Figure 4.3: Simulated standard deviations (left panel), true errors (middle panel) and observed residuals (right panel) for scenario (g).

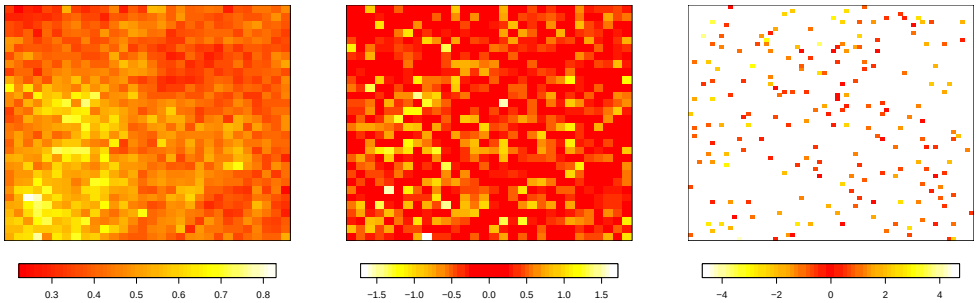

Figure 4.4: Simulated standard deviations (left panel), true errors (middle panel) and observed residuals (right panel) for scenario (h).

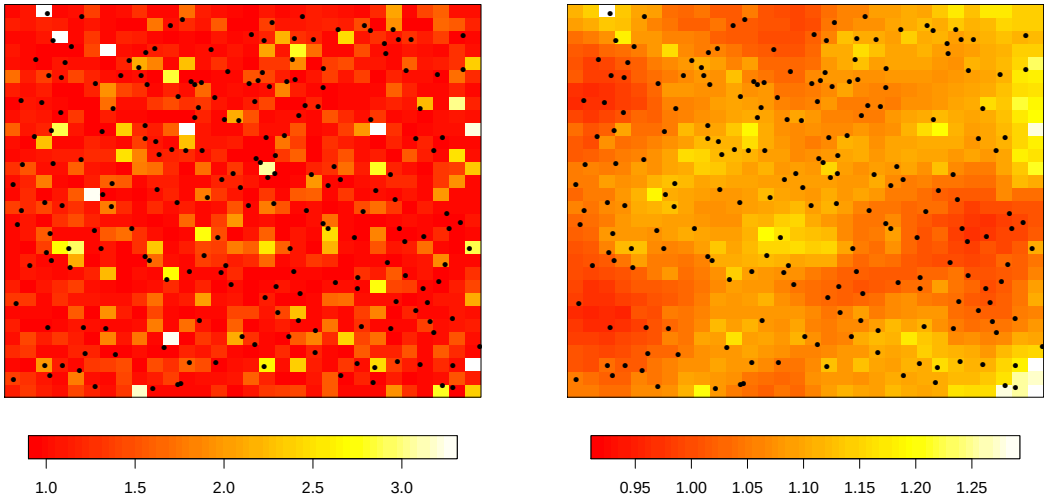

Figure 4.5: Estimated uncertainty under prior (4.6) in the right panel and under prior (4.7) in the left panel for scenario (a). Black dots represent validation sites ("Coords 1").

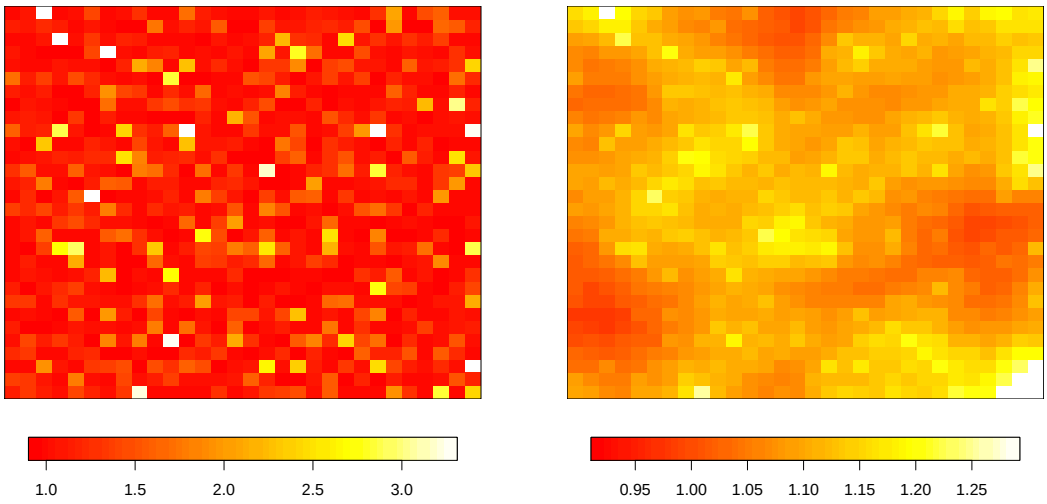

Figure 4.6: Estimated uncertainty under prior (4.6) in the left panel and under prior (4.7) in the right panel for scenario (b).

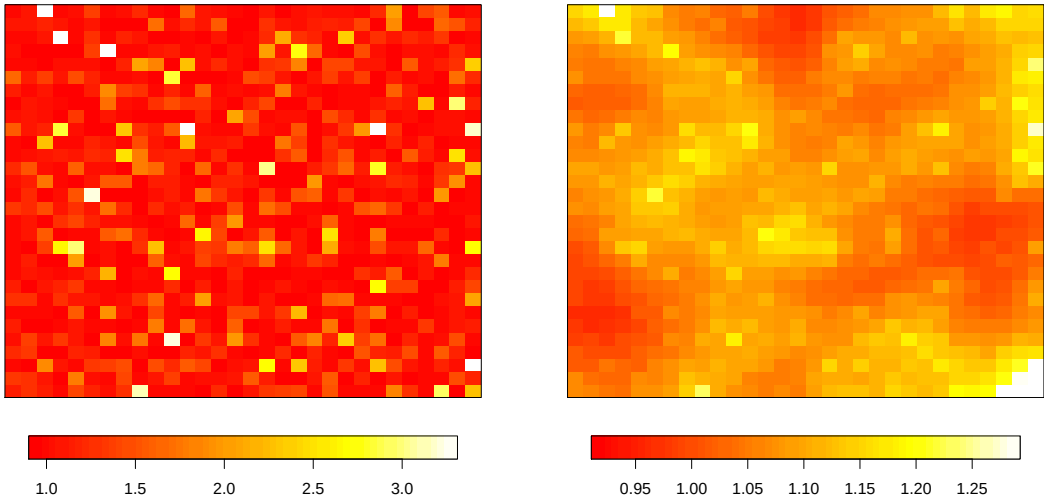

Figure 4.7: Estimated uncertainty under prior (4.6) in the left panel and under prior (4.7) in the right panel for scenario (c).

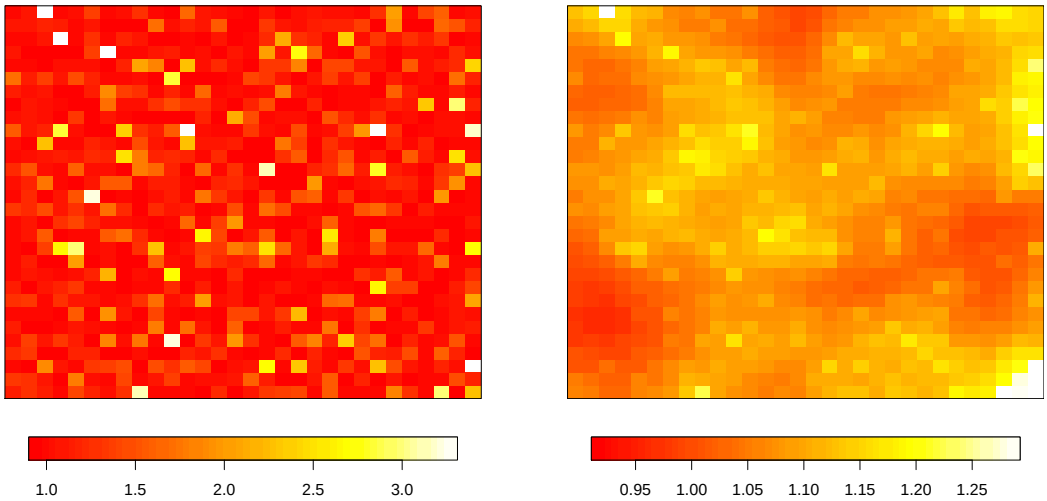

Figure 4.8: Estimated uncertainty under prior (4.6) in the left panel and under prior (4.7) in the right panel for scenario (d).

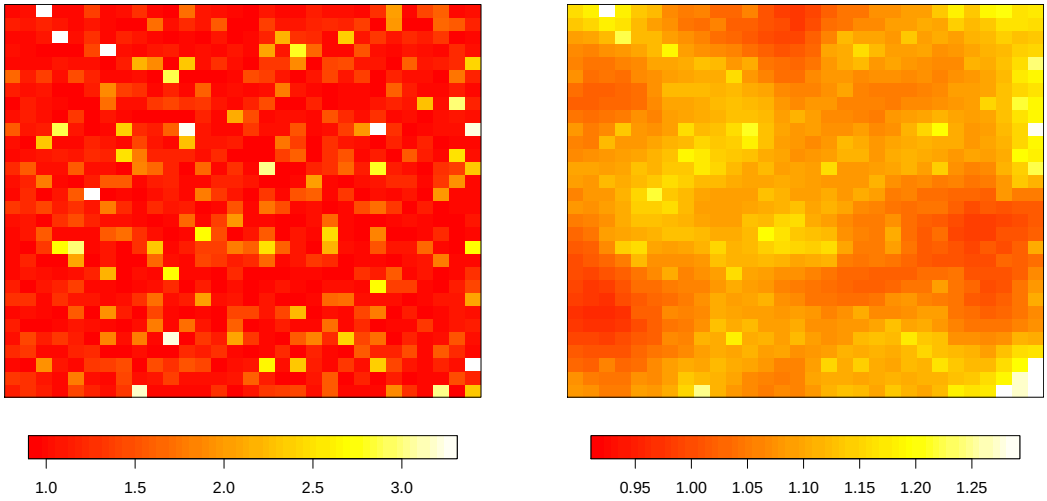

Figure 4.9: Estimated uncertainty under prior (4.6) in the left panel and under prior (4.7) in the right panel for scenario (e).

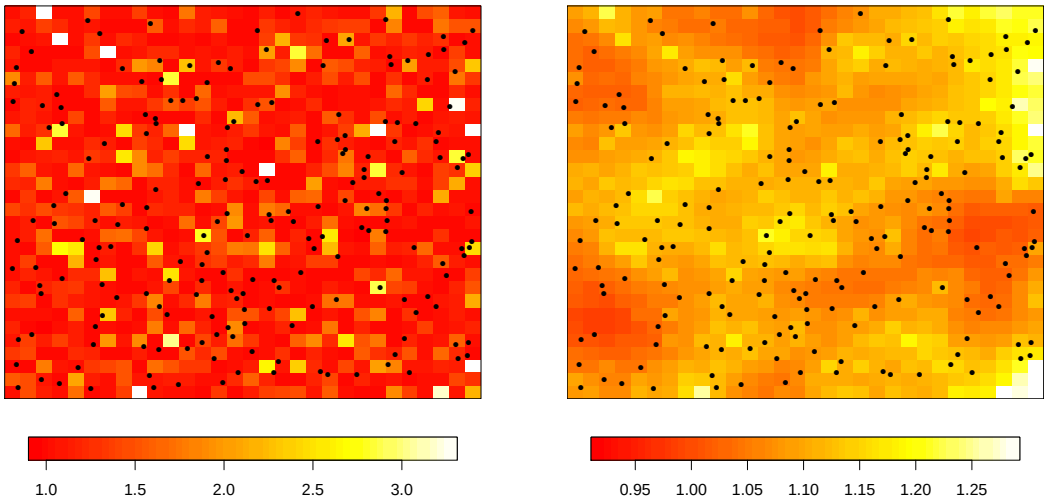

Figure 4.10: Estimated uncertainty under prior (4.6) in the left panel and under prior (4.7) in the right panel for scenario (f). Black dots represent validation sites ("Coords 2").

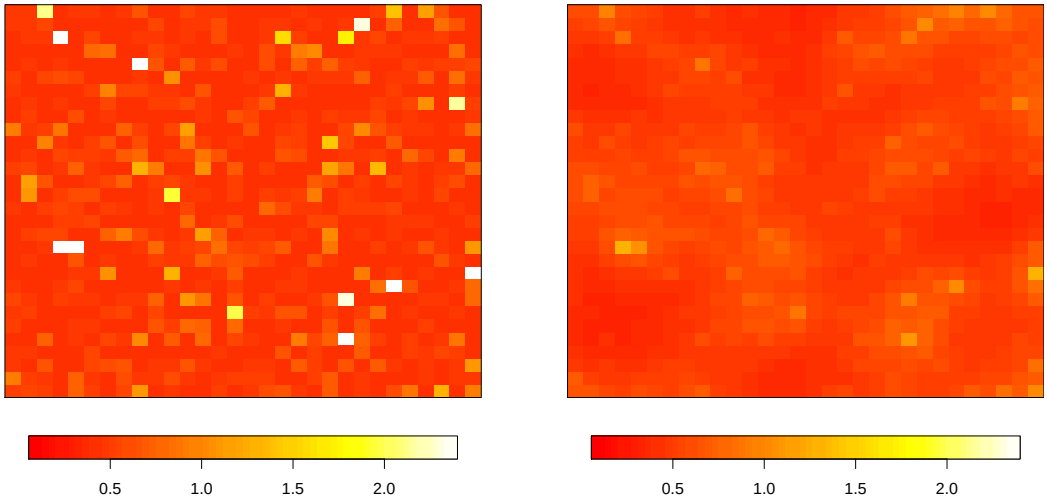

Figure 4.11: Estimated uncertainty under prior (4.6) in the left panel and under prior (4.7) in the right panel for scenario (g).

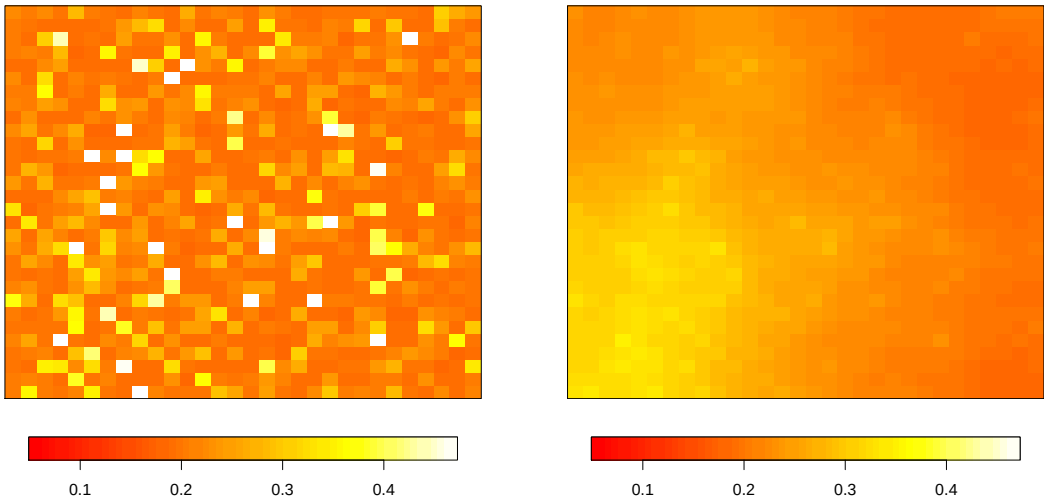

Figure 4.12: Estimated uncertainty under prior (4.6) in the left panel and under prior (4.7) in the right panel for scenario (h).

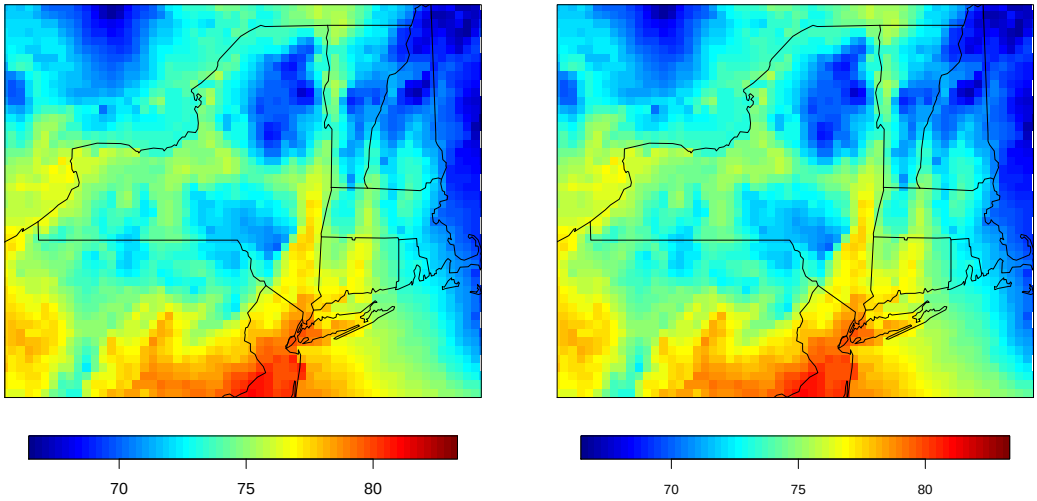

Figure 4.13: Posterior means of $\widetilde{R}(A_i)$ under priors (4.6) and (4.7) in left and right panel, respectively.

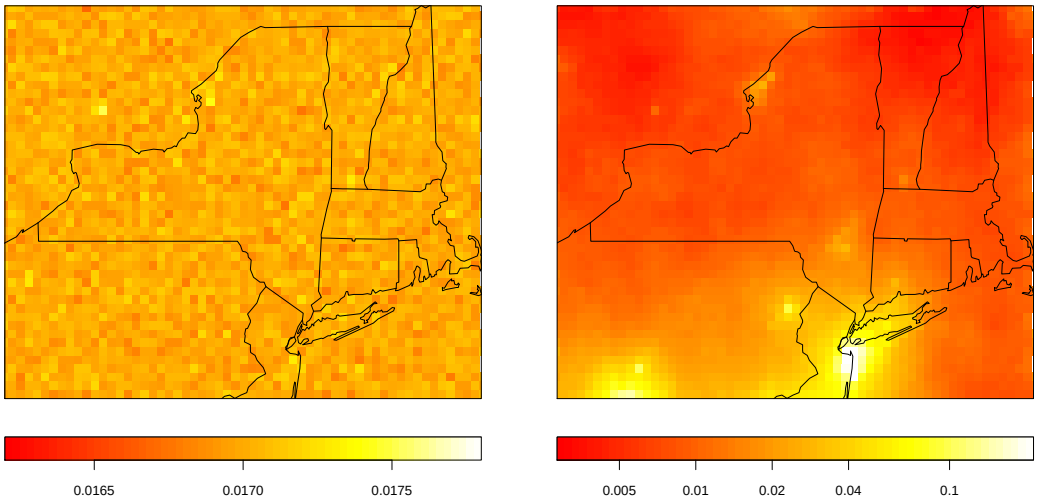

Figure 4.14: Estimated standard deviations under priors (4.6) and (4.7) in left and right panel, respectively.

Chapter 5

Conclusions

In this dissertation, we have proposed and discussed Bayesian modeling to combine monitoring data and numerical model output. Such data fusion has been motivated by different goals, mainly real-time forecasting and uncertainty quantification for numerical model output. We also have addressed the change of support problem encountered in blending observed data with model output via the downscaling approach.

In the first part of this thesis, we have addressed a specific applied challenge, that is real-time forecasting of current 8-hour average ozone levels over the conterminous U.S.. We have combined ozone monitoring data with ozone predictions from the Eta-CMAQ air quality model in order to forecast real-time current 8-hour average ozone level defined as the average of the previous four hours, current hour, and predictions for the next three hours. We have proposed a Bayesian downscaler model based on first differences with a flexible coefficient structure and an efficient computational strategy such that it is feasible for real-time implementation. We have used historical data from a large subregion of U.S. to show that our approach enables significant improvement in the accuracy of ozone forecasting relative to the predictions provided by the current forecasting system.

Furthermore, we have considered the introduction of temperature data into our downscaler for real-time forecasting. In particular, ozone monitoring data has been fused with real-time temperature data arising as output from a weather forecast model. Again, we have exploited first differences to expedite computation. Model validation for the eastern United States showed consequential improvement of our fully inferential approach compared with the existing implementations and the previous downscaler. One added advantage of our method is its easy and cheap

implementation allowed by free access to the RUC (RAP) output at the NOAA's web site. Hence, our strategy is currently being implemented by EPA to provide the public and experts with real-time current 8-hour average ozone predictions.

In the last part of this work, we have developed a data fusion model to quantify the uncertainty associated with numerical model output, regardless of how it was created. In fact, as many authors have noted, to be fully helpful an inference or prediction must also have an uncertainty assessment attached to it. However, computer models do not usually provide any uncertainty measure associated with their predictions, since they are deterministic models. We have proposed a Bayesian hierarchical model to provide spatially smoothed uncertainty associated with numerical model output, showing that we can learn about such uncertainty through suitable stochastic data fusion modeling using some external validation data. Our model has been successfully applied to attach uncertainty to temperature output over the northeastern United States.

Model developments presented in this thesis enable accurate forecasting along with appropriate uncertainty quantification and so they might be helpful to advance the knowledge about several environmental phenomena. For instance, accurate and instantaneous forecasting of ozone exposure can better inform the public and environmental decision-makers about air pollution levels that may lead to harmful health effects. Similarly, a suitable uncertainty assessment may provide useful information to guide environmental agencies in thinning and improving computer models.

Finally, it is worth to give a general overview of this dissertation. In fact, we can envision a link between the first part of this work concerning forecasting and last part focusing on uncertainty assessment for deterministic maps. We recall that the downscaler for real-time forecasting proposed in Chapters 2 and 3 takes directly the model output as covariate. However, we have shown that the model output is affected by uncertainty and it might be relevant to take into account such uncertainty within the downscaler. In other words, we can be interested to quantify the uncertainty associated with the model output and then look at the effect of such uncertainty on the ozone forecasting.

Appendix A

Full conditional distributions

Here, we derive the full conditional distributions under model (2.9) - (2.10). The full conditional distributions for the inverse of the variance parameters τ^2, σ^2 and ξ^2 are:

$$\frac{1}{\tau^2}\,|rest \sim Ga\Bigg(a+\frac{(T-3)n}{2},\\ b+\frac{1}{2}\sum_{t=1}^{T-3}\sum_{i=1}^{n}\Big(\Delta_t^Z(\mathbf{s}_i)-\beta_0(\mathbf{s}_i)\beta_{0,t}-\beta_1\Delta_t^x(\mathbf{s}_i)-\beta_1(\mathbf{s}_i)\Delta_t^x(\mathbf{s}_i)\Big)^2\Bigg)$$

$$\frac{1}{\sigma^2}\,|rest \sim Ga\Bigg(a+\frac{n}{2}, b+\frac{1}{2}\mathbf{B}_0^{\prime(s)}H^{-1}(\phi)\mathbf{B}_0^{(s)}\Bigg)$$

$$\frac{1}{\xi^2}\,|rest \sim Ga\Bigg(a+\frac{n}{2}, b+\frac{1}{2}\mathbf{B}_1^{\prime(s)}H^{-1}(\phi)\mathbf{B}_1^{(s)}\Bigg)$$

The full conditional distribution for the global slope parameter β_1 is: $\beta_1\,|rest \sim N(vg, v)$ where

$$v^{-1}=\frac{1}{\tau^2}\sum_{t=1}^{T-3}\sum_{i=1}^{n}(\Delta_t^x(\mathbf{s}_i))^2+\frac{1}{g^2}$$

$$g=\frac{1}{\tau^2}\sum_{t=1}^{T-3}\sum_{i=1}^{n}\Delta_t^x(\mathbf{s}_i)\Big(\Delta_t^Z(\mathbf{s}_i)-\beta_0(\mathbf{s}_i)\beta_{0,t}-\beta_1(\mathbf{s}_i)\Delta_t^x(\mathbf{s}_i)\Big)$$

Let $\mathbf{\Delta}_t^Z=\big(\Delta_t^Z(\mathbf{s}_1),\ldots,\Delta_t^Z(\mathbf{s}_n)\big)'$ and $\mathbf{\Delta}^Z(\mathbf{s}_i)=\big(\Delta_1^Z(\mathbf{s}_i),\ldots,\Delta_{T-3}^Z(\mathbf{s}_i)\big)'$ the vectors that collect spatial series and temporal series, respectively. The full conditional distribution for the intercept spatial effect is a normal distribution $\mathbf{B}_0^{(s)}\,|rest \sim$

$N(Vd, V)$ where

$$V^{-1} = \frac{1}{\tau^2}\sum_{t=1}^{T-3}\big(\beta_{0,t}\big)^2 + \frac{1}{\sigma^2}H^{-1}(\phi)$$
$$d = \frac{1}{\tau^2}\sum_{t=1}^{T-3}\beta_{0,t}\big(\boldsymbol{\Delta}_t^Z - \beta_1\boldsymbol{\Delta}_t^x - D_t^x\mathbf{B}_1^{(s)}\big)$$

and the matrix D_t^x is diagonal with $\big(D_t^x\big)_{ii'} = \Delta_t(\mathbf{s}_i)$. For the slope spatial effect we have[1] $\mathbf{B}_1^{(s)}\,|rest \sim N(Vd, V)$ where

$$V^{-1} = \frac{1}{\tau^2}\sum_{t=1}^{T-3}\big(D_t^x\big)^2 + \frac{1}{\xi^2}H^{-1}(\phi)$$
$$d = \frac{1}{\tau^2}\sum_{t=1}^{T-3}D_t^x\big(\boldsymbol{\Delta}_t^Z - \beta_{0,t}\mathbf{B}_0^{(s)} - \beta_1\boldsymbol{\Delta}_t^x\big)$$

Finally, the full conditional distribution for the temporal effect is a normal distribution $\mathbf{B}_0^{(t)}\,|rest \sim N(Vd, V)$ where

$$V^{-1} = \frac{1}{\tau^2}\sum_{i=1}^{n}\big(\beta_0(\mathbf{s}_i)\big)^2 + K^{-1}(\varphi)$$
$$d = \frac{1}{\tau^2}\sum_{i=1}^{n}\beta_0(\mathbf{s}_i)\big(\boldsymbol{\Delta}(\mathbf{s}_i)^Z - \beta_1\boldsymbol{\Delta}(\mathbf{s}_i)^x - \beta_1(\mathbf{s}_i)\boldsymbol{\Delta}(\mathbf{s}_i)^x\big)$$

[1] We re-use the same symbols for notation simplicity.

Appendix B

Pseudo algorithm

Suppose that today is August 2nd, 2012 and the current hour, T, is 07:00 (UTC). We would forecast the current 8-hour average ozone level $Z_T(\mathbf{s})$ at 7AM on August 2nd, corresponding to the average of hourly ozone data from 3AM to 10AM. Suppose we have already collect historical data up to 6AM on August 2nd. Then, at the current hour T, we would need to implement the following steps:

1. Download the RAP output from the ftp:
   ftp://ftpprd.ncep.noaa.gov/../../pub/data/nccf/com/rap/prod/.

   Usually, the RAP file are specified as

   rap.yyyymmdd/rap.thhz.awp130bgrfff.grib2

   where yyyymmdd is the date, hh is the model cycle runtime, 130 is the spatial resolution (13-km) and ff is the forecast hour.
   For instance, rap.20120801/rap.t00z.awp130bgrf01.grib2 is the 1-hour ahead RAP output from model which is run at 00:00 on August 1st.

   In principle, at 7AM, we would need the following files:

   - rap.20120802/rap.t07z.awp130bgrf00.grib2 (corresponding to 7AM)
   - rap.20120802/rap.t07z.awp130bgrf01.grib2 (corresponding to 8AM)
   - rap.20120802/rap.t07z.awp130bgrf02.grib2 (corresponding to 9AM)
   - rap.20120802/rap.t07z.awp130bgrf03.grib2 (corresponding to 10AM)

   However, the RAP files could be delayed of some hours. In this case, we can download the last available hour plus the k-step ahead forecasts corresponding to the data we need. In general for hh$= T - k$, we download the file

corresponding to ff$= 0 : 3 + k$.
In this example, we need the RAP output (hourly) from 12AM on August 1st to 7AM on August 2nd, plus 1-hour, 2-hour and 3-hour ahead forecasts corresponding to 8AM, 9AM and 10AM on August 2nd.

2. Extract the variable "temperature 2-m above ground" and the coordinates from the files downloaded. Let $R_t(B)$ denotes the hourly temperature RAP output over the grid cell B.

3. Given the 8-hour average ozone data, remove each site with at least one missing values in the 24-hour window.

4. Associate to each site $\mathbf{s}$ the grid cell B that contains $\mathbf{s}$ ($\mathbf{s} \in B$).

5. Compute the ozone monitoring data differences in (2.5) and the RAP data differences as in (3.1) and consider model (3.2).

6. Set aside data from m randomly chosen sites (about 10% or less) to estimate the decay parameters ϕ and φ.

7. Given a set of values for ϕ and φ, fit model (3.2) via a Gibbs sampling for each combination of the two decay parameters (parallel computation).

8. Predict $\Delta_t^Z(\mathbf{s}'_j)$ for each $t = 1, \ldots, 24$ and $j = 1, \ldots, m$ (at the hold-out sites).

9. Compute the VMSE in (2.17) and choose the combination of ϕ and φ which leads to the smallest VMSE.

10. Predict $\Delta_{T-2}^Z(\mathbf{s}_i)$, $\Delta_{T-1}^Z(\mathbf{s}_i)$ and $\Delta_T^Z(\mathbf{s}_i)$ for each $i = 1, \ldots, n$ (at the monitoring sites) and sum each predictive posterior draw to $Z_T(\mathbf{s}_i)$.

11. Obtain the $Z_T(\mathbf{s})$ at the Eta-CMAQ centroids via an ordinary kriging using a fast available package.

12. Get the predicted surface of the current 8-hour average ozone level as average of the posterior predictive distribution of the kriged $Z_T(\mathbf{s})$.

Appendix C

MCMC details

We partition $\widetilde{\mathbf{R}} = \big(\widetilde{\mathbf{R}}^{(1)}, \widetilde{\mathbf{R}}^{(2)}\big)$, where $\widetilde{\mathbf{R}}^{(1)}$ corresponds to the numerical model output for the n grid cells where monitoring stations are, while $\widetilde{\mathbf{R}}^{(2)}$ is the vector containing the numerical model output at $(I-n)$ grid cells where no observations are made.

The full conditional distributions for the inverse of the variance parameters σ_v^2, $\sigma_{\tilde{v}}^2$, τ^2 and τ_*^2 are:

$$\begin{aligned}
\frac{1}{\sigma_v^2}\,|rest &\sim Ga\bigg(a_\sigma + \frac{n}{2},\ b_\sigma + \frac{1}{2}\Big(\mathbf{V}-\widetilde{\mathbf{V}}\Big)' H^{-1}(\phi)\Big(\mathbf{V}-\widetilde{\mathbf{V}}\Big)\bigg)\\
\frac{1}{\sigma_{\widetilde{v}}^2}\,|rest &\sim Ga\bigg(a_\sigma + \frac{n}{2},\ b_\sigma + \frac{1}{2}\Big(\widetilde{\mathbf{V}}-\widetilde{\mathbf{R}}^{(1)}\Big)'\Big(\widetilde{\mathbf{V}}-\widetilde{\mathbf{R}}^{(1)}\Big)\bigg)\\
\frac{1}{\tau^2}\,|rest &\sim Ga\bigg(a_\tau + \frac{I}{2},\ b_\tau + \frac{1}{2}\widetilde{\mathbf{R}}'\big(D_w - W\big)\widetilde{\mathbf{R}}\bigg)\\
\frac{1}{\tau_*^2}\,|rest &\sim Ga\bigg(a_\tau + \frac{I}{2},\ b_\tau + \frac{1}{2}\big(\log(\boldsymbol{\sigma}_r^2)\big)'\big(D_w - W\big)\big(\log(\boldsymbol{\sigma}_r^2)\big)\bigg)
\end{aligned}$$

where $\boldsymbol{\sigma}_r^2 = \big(\sigma_r^2(A_1), \ldots, \sigma_r^2(A_I)\big)'$, $D_w = diag(w_i)$ and W is the proximity matrix[1].

The posterior conditional distribution for $\widetilde{\mathbf{V}}$ is a multivariate normal distribution $N(D_{\tilde{v}} d_{\tilde{v}}, D_{\tilde{v}})$, where

$$\begin{aligned}
D_{\tilde{v}}^{-1} &= \frac{1}{\sigma_v^2}H^{-1}(\phi) + \frac{1}{\sigma_{\tilde{v}}^2}I_n\\
d_{\tilde{v}} &= \frac{1}{\sigma_v^2}H^{-1}(\phi)\mathbf{V} + \frac{1}{\sigma_{\tilde{v}}^2}\widetilde{\mathbf{R}}^{(1)}
\end{aligned}$$

We sample the elements of $\widetilde{\mathbf{R}}$ using univariate sampling scheme as following. If $\widetilde{R}(A_i) \in \widetilde{\mathbf{R}}^{(1)}$, the full conditional distribution for $\widetilde{R}(A_i)$ is normal distribution

[1] Adjacent cells according to the rook's neighborhood structure.

$N(D_{r1}d_{r1}, D_{r1})$, where

$$\begin{aligned} D_{r1}^{-1} =& \frac{1}{\sigma_r^2(A_i)} + \frac{1}{\sigma_{\tilde{v}}^2} + \frac{w_i}{\tau^2} \\ d_{r1} =& \frac{R(A_i)}{\sigma_r^2(A_i)} + \frac{\widetilde{V}(A_i)}{\sigma_{\tilde{v}}^2} + \frac{1}{\tau^2}\sum_{i'\sim i}\widetilde{R}(A_{i'}) \end{aligned}$$

If $\widetilde{R}(A_i) \in \widetilde{\mathbf{R}}^{(2)}$, the full conditional distribution for $\tilde{R}(A_i)$ is normal distribution $N(D_{r2}d_{r2}, D_{r2})$, where

$$\begin{aligned} D_{r2}^{-1} =& \frac{1}{\sigma_r^2(A_i)} + \frac{w_i}{\tau^2} \\ d_{r2} =& \frac{R(A_i)}{\sigma_r^2(A_i)} + \frac{1}{\tau^2}\sum_{i'\sim i}\widetilde{R}(A_{i'}) \end{aligned}$$

Given the common inverse gamma prior $IG(a, b_*)$, the full conditional distribution for each $\sigma_r^2(A_i)$ is:

$$\frac{1}{\sigma_r^2(A_i)}\,|rest \sim Ga\left(a + \frac{1}{2}, b_* + \frac{1}{2}\Big(R(A_i) - \tilde{R}(A_i)\Big)^2\right) \tag{C.1}$$

The Gamma distribution falls in the conjugate class of prior densities for b_*, then it is straightforward to show that its full conditional distribution is

$$b_*\,|rest \sim Ga\left(c + aI, d + \frac{1}{\sum_{i=1}^{I}\sigma_r^2(A_i)}\right)$$

Finally, using the logCAR prior model for $\sigma_r^2(A_i)$, the full conditionals cannot be obtained in closed form; so we use a random walk Metropolis proposal steps for individual grid cell to generate samples from its posterior.

Bibliography

Arbia, G. (1989). Statistical effect of spatial data transformations: a proposed general framework. In Goodchild, M. and Gopal, S., editors, *The Accuracy of Spatial Data Bases*, pages 249–259. Taylor & Francis, London.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC, Boca Raton, FL.

Benjamin, S. G., Dèvènyi, D., Weygandt, S. S., Brundage, K. J., Brown, J. M., Grell, G. A., Kim, D., Schwartz, B. E., Smirnova, T. G., Smith, T. L., and Manikin, G. S. (2004). An hourly assimilation-forecast cycle: the RUC. *Monthly Weather Review*, 132:495–518.

Berrocal, V. J., Gelfand, A. E., and Holland, D. M. (2010a). A bivariate spatio-temporal downscaler under space and time misalignment. *Annals of Applied Statistics*, 4:1942–1975.

Berrocal, V. J., Gelfand, A. E., and Holland, D. M. (2010b). A spatio-temporal downscaler for output from numerical models. *Journal of Agricultural, Biological and Environmental Statistics*, 14:176–197.

Berrocal, V. J., Gelfand, A. E., and Holland, D. M. (2012). Space-time data fusion under error in computer model output: an application to modeling air quality. *Biometrics*, 68:837–848.

Berrocal, V. J., Raftery, A. E., and Gneiting, T. (2007). Combining spatial statistical and ensemble information in probabilistic weather forecasts. *Monthly Weather Review*, 135:1386–1402.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, 36(2):192–236.

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.

Bloomer, B. J., Stehr, J. W., Piety, C. A., Salawitch, R. J., and Dickerson, R. R. (2009). Observed relationships of ozone air pollution with temperature and emissions. *Geophysical research letters*, 36, DOI: 10.1029/2009GL037308.

Bloomfield, P., Royle, J. A., Steinberg, L. J., and Yang, Q. (1996). Accounting for meteorological effects in measuring urban ozone levels and trends. *Atmospheric Environment*, 30(17):3067 – 3077.

Bruno, F., Cocchi, D., Greco, F., and Scardovi, E. (2013a). Spatial reconstruction of rainfall fields from rain gauge and radar data. *Stochastic Environmental Research and Risk Assessment*, in press.

Bruno, F., Cocchi, D., and Paci, L. (2013b). A practical approach for assessing the effect of grouping in hierarchical spatio-temporal models. *Advances in Statistical Analysis*, 97:93–108.

Bruno, F., Guttorp, P., Sampson, P. D., and Cocchi, D. (2009). A simple non-separable and non-stationary spatiotemporal model for ozone. *Environmental and Ecological Statistics*, 16:515–529.

Bruno, F. and Paci, L. (2013). Hierarchical spatio-temporal models for short-term predictions of air pollution data. In Brentari, E. and Carpita, M., editors, *Advances in Latent Variables*. Vita e Pensiero, Milan, Italy.

Camalier, L., Cox, W., and Dolwick, P. (2007). The effects of meteorology on ozone in urban areas and their use in assessing ozone trends. *Atmospheric Environment*, 41(33):7127 – 7137.

Cameletti, M., Ghigo, S., and Ignaccolo, R. (2011). A spatio-temporal model for air quality mapping using uncertain covariates. In Cafarelli, B., editor, *Spatial Data Methods for Environmental and Ecological Processes - 2nd Edition. Proceedings*, Foggia. CDP Service Edizioni.

Carroll, S. S., Day, G. N., Cressie, N. A. C., and Carroll, T. R. (1995). Spatial modeling of snow water equivalent using airborne and ground-based snow data. *Environmetrics*, 6:127–139.

Chaloner, K. (1994). Residual analysis and outliers in Bayesian hierarchical models. In Smith, A. F. M. and Freeman, P. R., editors, *Aspects of Uncertainty*, pages 149–157. John Wiley & Sons, Chichester, UK.

Chaloner, K. and Brant, R. (1988). A Bayesian approach to outlier detection and residual analysis. *Biometrika*, 75:651–659.

Chiles, J. and Delfiner, P. (1999). *Geostatistics: Modeling Spatial Uncertainty*. JohnWiley and Sons, New York.

Cocchi, D. and Trivisano, C. (2013). Ozone. In El-Shaarawi, A. H. and Piegorsch, W., editors, *Encyclopedia of Environmetrics*. John Wiley & Sons Ltd, Chichester, UK. DOI: 10.1002/9780470057339.vao022.pub2.

Cowles, M. K. and Zimmerman, D. L. (2003). A Bayesian space-time analysis of acid deposition data combined for two monitoring networks. *Journal of Geophysical Research*, 108:90–106.

Cowles, M. K., Zimmerman, D. L., Christ, A., and McGinnis, D. L. (2002). Combining snow water equivalent data from multiple sources to estimate spatio-temporal trends and compare measurement system. *Journal of Agricultural, Biological and Environmental Statistics*, 7:536–557.

Cox, W. M. and Chu, S. H. (1993). Meteorologically adjusted trends in urban areas: a probabilistic approach. *Atmospheric Environment*, 27(4):425–434.

Cox, W. M. and Chu, S. H. (1996). Assessment of interannual ozone variation in urban areas from a climatological perspective. *Atmospheric Environment*, 30(14):2615 – 2625.

Craig, P. S., Goldstein, M., Seheult, A. H., and Smith, J. A. (1998). Constructing partial prior specifications for models of complex physical systems. *The Statistician*, 47:37–53.

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

Cumming, J. A. and Goldstein, M. (2010). Bayes linear uncertainty analysis for oil reservoirs based on multiscale computer experiments. In Gelfand, A. E., Fuentes, M., Guttorp, P., and Diggle, P. J., editors, *Handbook of Spatial Statistics*, pages 241–270. CRC Press, Boca Raton, FL.

Di Narzo, A. F. and Cocchi, D. (2010). A Bayesian hierarchical approach to ensemble weather forecasting. *Journal of the Royal Statistical Society, Series C*, 59:405–422.

Dominici, F., Daniels, M., Zeger, S. L., and Samet, J. M. (2002). Air pollution and mortality: Estimating regional and national dose-response relationships. *Journal of the American Statistical Association*, 97:100–111.

Dou, Y., Le, N., and Zidek, J. (2010). Modeling hourly ozone concentration fields. *Annals of Applied Statistics*, 4:1183–1213.

Draper, D., Pereira, A., Prado, P., Saltelli, A., Cheal, R., Eguilior, S., Mendes, B., and Tarantola, S. (1999). Scenario and parametric uncertainty in GESAMAC: A methodological study in nuclear waste disposal risk assessment. *Computer Physics Communications*, 117:142 – 155.

Fuentes, M. and Foley, M. (2008). A statistical framework to combine multivariate spatial data and physical models for hurricane surface wind prediction. *Journal of Agricultural, Biological and Environmental Statistics*, 13:37–59.

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

Fuentes, M., Reich, B., and Lee, G. (2008). Spatial-temporal mesoscale modeling of rainfall intensity using gage and radar data. *Annals of Applied Statistics*, 2:1148–1169.

Fuentes, M., Song, H.-R., Ghosh, S. K., Holland, D. M., and Davis, J. M. (2006). Spatial association between speciated fine particles and mortality. *Biometrics*, 62:855–863.

Fuller, W. A. (1987). *Measurement Error Models*. Wiley, New York.

Gelfand, A. E. (2010). Misaligned spatial data: The change of support problem. In Gelfand, A. E., Fuentes, M., Guttorp, P., and Diggle, P. J., editors, *Handbook of Spatial Statistics*, pages 517–539. CRC Press, Boca Raton, FL.

Gelfand, A. E. and Ghosh, S. K. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika*, 85:1–11.

Gelfand, A. E., Kim, H.-J., Sirmans, C. F., and Banerjee, S. (2003). Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, 98:387–396.

Gelfand, A. E. and Sahu, S. K. (2010). Combining monitoring data and computer model output in assessing environmental exposure. In O'Hagan, A. and West, M., editors, *Handbook of Applied Bayesian Analysis*, pages 482–510. Oxford University Press.

Gelfand, A. E., Schmidt, A. M., Banerjee, S., and Sirmans, C. (2004). Nonstationary multivariate process modeling through spatially varying coregionalization. *TEST*, 13:263–312.

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

Ghosh, S., Gelfand, A. E., and Mølhave, T. (2012). Attaching uncertainty to deterministic spatial interpolations. *Statistical Methodology*, 9:251–264.

Givens, G. H., Raftery, A. E., and Zeh, J. E. (1993). Benefits of a Bayesian approach for synthesizing multiple sources of evidence and uncertainty linked by a deterministic model. *Report of the International Whaling Commission*, 43:495–500.

Gneiting, T., Raftery, A. E., Westveld, A. H., and Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation. *Monthly Weather Review*, 133:1098–1118.

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

Gotway, C. A. and Young, L. J. (2007). A geostatistical approach to linking geographically aggregated data from different sources. *Journal of Computational and Graphical Statistics*, 16:115–135.

Greco, F. P., Lawson, A. B., Cocchi, D., and Temples, T. (2005). Some interpolation estimators in environmental risk assessment for spatially misaligned health data. *Environmental and Ecological Statistics*, 12:379–395.

Guillas, S., Bao, J., Choi, Y., and Wang, Y. (2008). Statistical correction and downscaling of chemical transport model ozone forecasts over atlanta. *Atmospheric Environment*, 12:1338–1348.

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

Heaton, M., Holland, D. M., and Leininger, T. (2012). User's manual for downscaler fusion software. *U.S. Environmental Protection Agency, Washington, DC, EPA/600/C-12/002*.

Huerta, G., Sanso, B., and Stroud, J. (2004). A spatiotemporal model for Mexico City ozone levels. *Journal of the Royal Statistical Society, Series C*, 53:231–248.

Ignaccolo, R., Ghigo, S., and Giovenali, E. (2008). Analysis of air quality monitoring networks by functional clustering. *Environmetrics*, 19:672–686.

Jacob, D. J. and Winner, D. A. (2009). Effect of climate change on air quality. *Atmospheric Environment*, 43:51–63.

Jun, M. and Stein, M. L. (2004). Statistical comparison of observed and cmaq modeled daily sulfate levels. *Atmospheric Environment*, 38:4427–4436.

Kang, D., Mathur, R., Rao, S. T., and Yu, S. (2008). Bias adjustment techniques for improving ozone air quality forecasts. *Journal of Geophysical Research*, 113(D23):DOI:10.1029/2008JD010151.

Katul, G. G., Ruggeri, F., and Vidakovic, B. (2006). Denoising ozone concentration measurements with BAMS filtering. *Journal of Statistical Planning and Inference*, 136:2395–2405.

Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models (with discussion). *Journal of the Royal Statistical Society, Series B*, 63:425–464.

Kleiber, W., Raftery, A. E., and Gneiting, T. (2011). Locally calibrated probabilistic quantitative precipitation forecasting. *Journal of the American Statistical Association*, 106(496):1291–1303.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach (with discussion). *Journal of the Royal Statistical Society, Series B*, 73(4):423–498.

Liu, Z., Le, N. D., and Zidek, J. V. (2011). An empirical assessment of Bayesian melding for mapping ozone pollution. *Environmetrics*, 22:340–353.

Massart, B. G. and Kvalheim, O. M. (1998). Ozone forecasting from meteorological variables: Part i. predictive models by moving window and partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 42:179 – 190.

McMillan, N., Bortnick, S., Irwin, M., and Berliner, M. (2005). A hierarchical Bayesian model to estimate and forecast ozone through space and time. *Atmospheric Environment*, 39:1373–1382.

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

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

Mugglin, A. S., Carlin, B. P., and Gelfand, A. E. (2000). Fully model based approaches for spatially misaligned data. *Journal of the American Statistical Association*, 95:877–887.

Nychka, D. W. and Anderson, J. L. (2010). Data assimilation. In Gelfand, A. E., Fuentes, M., Guttorp, P., and Diggle, P. J., editors, *Handbook of Spatial Statistics*, pages 241–270. CRC Press, Boca Raton, FL.

Nychka, D. W., Wikle, C., and Royle, J. A. (2002). Multiresolution models for nonstationary spatial covariance functions. *Statistical Modelling*, 2:315–332.

Oakley, J. E. and O'Hagan, A. (2004). Probabilistic sensitivity analysis of complex models: a Bayesian approach. *Journal of the Royal Statistical Society, Series B*, 66(3):751–769.

O'Hagan, A. (2006). Bayesian analysis of computer code outputs: A tutorial. *Reliability Engineering and System Safety*, 91:1290–1300.

Orasi, A., Jona Lasinio, G., and Ferrari, C. (2009). Comparison of calibration methods for the reconstruction of space-time rainfall fields during a rain enhancement experiment in southern italy. *Environmetrics*, 20:812–834.

Paci, L., Gelfand, A. E., and Holland, D. M. (2013). Spatio-temporal modeling for real-time ozone forecasting. *Spatial Statistics*, 4:79–93.

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., Givens, G. H., and Zeh, J. E. (1995). Inference from a deterministic population dynamics model for bowhead whales. *Journal of the American Statistical Association*, 90:402–416.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133:1155–1174.

Reich, B. J. and Hodges, J. S. (2008). Modeling longitudinal spatial periodontal data: a spatially adaptive model with tools for specifying priors and checking fit. *Biometrics*, 64:790–799.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15:351–357.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent gaussian model by using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B*, 71:319–392.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989). Design and analysis of computer experiments. *Statistical Sciences*, 4(4):409–435.

Sahu, S. K. and Bakar, K. (2012). A comparison of Bayesian models for daily ozone concentration levels. *Statistical Methodology*, 9:144 – 157.

Sahu, S. K., Gelfand, A. E., and Holland, D. M. (2007). High-resolution space-time ozone modeling for assessing trends. *Journal of the American Statistical Association*, 102:1221–1234.

Sahu, S. K., Gelfand, A. E., and Holland, D. M. (2010). Fusing point and areal level space-time data with application to wet deposition. *Journal of the Royal Statistical Society, Series C*, 59:77–103.

Sahu, S. K., Yip, S., and Holland, D. M. (2009a). A fast Bayesian method for updating and forecasting hourly ozone levels. *Environmental and Ecological Statistics*, 18:185–207.

Sahu, S. K., Yip, S., and Holland, D. M. (2009b). Improved space-time forecasting of next day ozone concentrations in the eastern us. *Atmospheric Environment*, 43:494–501.

Sloughter, J. M., Gneiting, T., and Raftery, A. E. (2010). Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *Journal of the American Statistical Association*, 105:25–35.

Sloughter, J. M., Gneiting, T., and Raftery, A. E. (2013). Probabilistic wind vector forecasting using ensembles and bayesian model averaging. *Monthly Weather Review*, 141:2017–2119.

Sloughter, J. M., Raftery, A. E., Gneiting, T., Albright, M., and Baars, J. (2007). Probabilistic quantitative precipitation forecasting using bayesian model averaging. *Monthly Weather Review*, 135:3209–3220.

Smith, B. J. and Cowles, M. K. (2007). Correlating point-referenced radon and areal uranium data arising from a common spatial process. *Applied Statistics*, 56:313–326.

Smith, R. L., Tebaldi, C., Nychka, D. W., and Mearns, L. O. (2009). Bayesian modeling of uncertainty in ensembles of climate models. *Journal of the American Statistical Association*, 104:97–116.

Sørbye, S. H. and Rue, H. (2013). Scaling intrinsic Gaussin Markov random field priors in spatial modelling. *Spatial Statistics*, http://dx.doi.org/10.1016/j.spasta.2013.06.004.

Stein, M. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Verlag, New York.

Stroud, J., Müller, P., and Sansó, B. (2001). Dynamic models for spatio-temporal data. *Journal of the Royal Statistical Society, Series B*, 63:673–689.

Swall, J. L. and Davis, J. M. (2006). A Bayesian statistical approach for the evaluation of CMAQ. *Atmospheric Environment*, 40(26):4883 – 4893.

Tebaldi, C. and Smith, R. L. (2010). Characterizing the uncertainty of climate change projections using hierarchical models. In Gelfand, A. E., Fuentes, M., Guttorp, P., and Diggle, P. J., editors, *Handbook of Spatial Statistics*, pages 545–594. CRC Press, Boca Raton, FL.

Thompson, M. L., Reynolds, J., Cox, L. H., Guttorp, P., and Sampson, P. D. (2001). A review of statistical methods for the meteorological adjustment of tropospheric ozone. *Atmospheric Environment*, 35:617–623.

Van de Kassteele, J. and Stein, A. (2006). A model for external drift kriging with uncertain covariates applied to air quality measurements and dispersion model output. *Environmetrics*, 17(4):309–322.

Wackernagel, H. (2003). *Multivariate Geostatistics: an introduction with applications (3rd edition)*. Springer, Berlin.

Wakefield, J. (2003). Sensitivity analyses for ecological regression. *Biometrics*, 59:9–17.

Wakefield, J. (2008). Ecological studies revisited. *Annual Review of Public Health*, 29:75–90.

Wikle, C. K. (2003). Hierarchical models in environmental science. *International Statistical Review*, 71:181–199.

Wikle, C. K. and Berliner, L. M. (2005). Combining information across spatial scales. *Technometrics*, 47:80–91.

Yan, J. (2007). Spatial stochastic volatility for lattice data. *Journal of Agricultural, Biological, and Environmental Statistics*, 12(1):25–40.

Yu, S., Mathur, R., Sarwar, G., Kang, D., Tong, D., Pouliot, G., and Pleim, J. (2010). Eta-cmaq air quality forecasts for o3 and related species using three different photochemical mechanisms (CB4, CB05, SAPRC-99): comparisons with measurements during the 2004 ICARTT study. *Atmospheric Chemistry and Physics*, 10:3001–3025.

Zellner, A. (1975). Bayesian analysis of regression error terms. *Journal of the American Statistical Association*, 70(349):138–144.

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

Zhu, L., Carlin, B. P., and Gelfand, A. E. (2003). Hierarchical regression with misaligned spatial data: relating ambient ozone and pediatric asthma ER visits in atlanta. *Environmetrics*, 14:537–557.

Zidek, J. V., Le, N. D., and Liu, Z. (2012). Combining data and simulated data for space-time fields: application to ozone. *Environmental and Ecological Statistics*, 19:37–56.