Improved space-time forecasting of next day ozone concentrations in the eastern U.S.

Sujit K. Sahu*
School of Mathematics,
University of Southampton,
Southampton, SO17 1BJ, UK.

Stan Yip
Exeter Climate Systems
University of Exeter,
Exeter, EX4 4QJ, UK

David M. Holland
U.S. Environmental Protection Agency
National Exposure Research Laboratory
Research Triangle Park, NC, 27711, USA.

October 14, 2008

Abstract

There is an urgent need to provide accurate air quality information and forecasts to the general public and environmental health decision-makers. This paper develops a hierarchical space-time model for daily 8-hour maximum ozone concentration (O_3) data covering much of the eastern United States. The model combines observed data and forecast output from a computer simulation model known as the Eta Community Multi-scale Air Quality (CMAQ) forecast model in a very flexible, yet computationally fast way, so that the next day forecasts can be computed in real-time operational mode. The model adjusts for spatio-temporal

*Corresponding Author. Email: S.K.Sahu@soton.ac.uk, Telephone number +44-23-8059-5123, Fax Number: +44-23-8059-5147

biases in the Eta CMAQ forecasts and avoids a change of support problem often encountered in data fusion settings where real data have been observed at point level monitoring sites, but the forecasts from the computer model are provided at grid cell levels. The model is validated with a large amount of set-aside data and is shown to provide much improved forecasts of daily O_3 concentrations in the Eastern United States.
Key Words: Bayesian modeling, data fusion, hierarchical model; Markov chain Monte Carlo; spatial interpolation.

1 Introduction

The most direct way to obtain accurate air quality information is from measurements made at surface monitoring stations. However, many areas of the eastern U.S. are not monitored and typically, air monitoring sites are sparsely and irregularly spaced. As the need for spatial prediction has become reality in the regulatory environment, it is now important to combine air monitoring data and numerical model output, in a coherent way for better prediction of air pollution over short time periods. High spatial resolution numerical model output are now available over 12 kilometer grids. Numerical models such as the Eta CMAQ forecast model use emission inventories, meteorological information, and land use to estimate average pollution levels for gridded cells over successive time periods, see e.g. http://www.epa.gov/asmdnerl/CMAQ/. The U.S. Environmental Protection Agency's (USEPA) AIRNow web site (http://airnow.gov) currently displays some air quality information such as the raw CMAQ forecasts and ground level station monitoring data. The Eta CMAQ forecasts often exhibit bias and simple interpolation of monitoring data fail to take into account of spatial and temporal dependencies present in the data. Fusion analyses must be developed to provide improved, in terms of computational efficiency, accuracy, and precision, forecasts of next day 8-hour maximum ozone levels through the AIRNow web site.

This paper builds upon recent advances in Bayesian space-time modeling to provide state-of-the-art forecasts maps of next day daily maximum 8-hour ozone concentration patterns in the Eastern United States. The proposed model uses the Eta CMAQ forecast data in a regression structure avoiding the so called 'change of support problem'. The method

does not require integration of the observed point level monitoring process to a grid level one, see Equation (1) below, unlike other data fusion methods currently available in the literature. Further, this method avoids modeling of the huge amount of Eta CMAQ gridded data in comparison to the number of monitoring sites. Providing a 'data' status to Eta CMAQ output, as often done in data fusion methods, may have the undesired result that the information contained in the Eta CMAQ output overwhelms the information in the monitoring data. Note that about 10,000 CMAQ grid cells cover our study region in the eastern U.S. and modeling temporal data from each of those sites would require processing of an enormous number of data points at each iteration of an iterative model fitting algorithm such as the Gibbs sampler. The associated computational burden will most likely not allow fitting the model in time to display the next day ozone levels on the AIRNow system. Thus, we develop a modeling approach that focuses on treating Eta CMAQ forecast output as a spatially and temporally varying covariate. We fit the model to spatially concurrent Eta CMAQ forecast output and monitoring data, i.e. use CMAQ cells that contain monitoring data, but use the entire Eta CMAQ output to forecast and interpolate next day ozone levels. As a result, the proposed modeling approach does not treat the Eta CMAQ forecasts as data per se, rather those are taken as spatio-temporally varying covariates.

Let $B_i, i = 1, \ldots K$ denote the K CMAQ grid cells. In general, there is no guarantee that the average concentration level in a grid cell B_i at time t denoted by $U(B_i, t)$ is equal to the concentration level which will be observed at any particular site $\mathbf{s}_j$ given by a pair of latitude and longitude values, denoted by $U(\mathbf{s}_j, t)$, within that cell. The so called change of support problem in this context is the problem associated with converting the average concentration of a grid cell B_i to the actual concentration level which will be observed at a particular site $\mathbf{s}_j$ within that cell. Note that

$$U(B_i, t) = \frac{1}{|B_i|} \int_{B_i} U(\mathbf{s}, t) d\mathbf{s} \tag{1}$$

where $|B_i|$ denote the area of the grid cell B_i. In the modeling development of this paper we shall respect this distinction between the average concentration $U(B_i, t)$ and a particular value $U(\mathbf{s}_j, t)$ throughout.

The modeling objective is to predict the actual ozone concentration level denoted by

$U(\mathbf{s},t)$ at a site $\mathbf{s}$ at time t on the basis of the Eta CMAQ forecast, $Q(B(\mathbf{s}),t)$ for the grid cell B containing $\mathbf{s}$. We do not have the actual Eta CMAQ forecasts $Q(\mathbf{s},t)$ at site $\mathbf{s}$ at time t. However, $Q(B(\mathbf{s}),t)$ can be expected to be a good regressor for $U(\mathbf{s},t)$ since $Q(B(\mathbf{s}),t)$ is the forecast model output for the average concentration of the grid cell $B(\mathbf{s})$ containing that particular site $\mathbf{s}$ at that time t. This is also confirmed by Figure 2 which plots both $U(\mathbf{s},t)$ and $Q(B(\mathbf{s}),t)$ at four randomly chosen sites. In this paper we shall model data on the square-root scale so we let $X(B(s),t)$ denote the square-root of the Eta CMAQ forecast value and for convenience we shall abbreviate this notation to be $X(\mathbf{s},t)$, i.e. drop the grid-cell notation B.

By now there is a number of publications discussing space-time modeling of ground level ozone, see, e.g., Guttorp *et al.* (1994), and Carroll *et al.* (1997). Hierarchical Bayesian approaches for spatial prediction of air pollution have been developed, see, e.g. Brown *et al.*, (1994), Huerta *et al.*, (2004), Wikle (2003), Sahu and Mardia (2005), Sahu *et al.* (2006, 2007) and references therein. McMillan *et al.* (2005) propose a regime switching model for ozone forecasting using meteorological variables as covariates and they illustrate using data from April to September in 1999 over a spatial domain covering Lake Michigan.

Several papers have appeared in the literature on the topic of data fusion methods for combining ground level observed data and computer model output. Fuentes and Raftery (2005) develop a hierarchical statistical framework to model the "true" pollutant process as jointly Gaussian random fields. They estimate the parameters for the bias of Eta CMAQ output and the parameters of the covariance structure for Eta CMAQ and measurement error processes, then simulate the conditional distribution of the "true" process given both sources of spatial information. Their methodology only applies to spatial processes at a fixed time point, although it can be extended for space-time data. Zimmerman and Holland (2005) consider the problem of optimal spatial prediction of wet deposition data using data from two monitoring networks with network-specific biases and variances. Cowles and Zimmerman (2003) use a Bayesian modeling approach for spatio-temporal data from two monitoring networks that account for possible differences in network measurement error, bias and variances. Jun and Stein (2004) suggest new ways of comparing space-time correlation structure of monitoring observations with Eta CMAQ numerical model output, see

also Yu *et al.* (2007). Eder et al. (2006) discuss statistical metrics to provide an operational evaluation of Eta CMAQ air quality forecasting system.

The remainder of this article is organized as follows. In Section 2 we describe the available data. Modeling developments are presented in Section 3. Prediction details are discussed in Section 4. Section 5 provides the modeling results and analyses. A few summary remarks are provided in Section 6 and an Appendix contains the computational details for Gibbs sampling.

2 Available Data

We use daily O_3 monitoring data for the two week period August 2–14 in the year 2005. The data obtained from http://nsdi.epa.gov/ttn/amtic are recorded in units of parts per billion (ppb) from n =350 monitoring sites spanning the eastern US (see Figure 1). We set aside data from 40 additional randomly chosen sites (numbered 1 to 40 in Figure 1) for model validation purposes. There are about 20% missing values in the data. We model the daily 8-hour maximum data for a running window of seven consecutive days during the two weeks and forecast the nex day's 8-hour maximum in each case. We simply choose to model seven days of data because those data complete a weekly cycle. Inclusion of more distant past data is also possible, but some preliminary analysis (not included here) did not show any significant improvement in interpolation and forecasting.

The output from the Eta CMAQ model are available a day in advance as hourly forecasts on a 12 kilometer grid, see e.g., (http://www.epa.gov/AMD/AQF/index.html). The daily Eta CMAQ value is computed as the maximum of all 8 hourly averages within a day. The 8 hourly averages are centered at the middle of eight hours, for example, the 8-hour average at 4PM is the average value obtained from the 8 hourly measurements observed from 12PM to 7PM.

For prediction purposes, we obtain the Eta CMAQ forecasts for 3000 randomly sampled grid cells out of the available 9119 such grid cells in the eastern US. This is for illustration purposes only and all the available data should be used to produce more accurate forecast maps.

The range of the Eta CMAQ forecasts matches with the range of the ground level observed data. As mentioned above, to compare the Eta CMAQ forecasts with the observed station data we plot data from four randomly chosen stations and Eta CMAQ forecasts from the corresponding grid cells containing the stations, see Figure 2. There is good agreement between the Eta CMAQ forecasts and observed data in some of the sites but there is also large disagreement between them at other sites. This implies that there is bias in the Eta CMAQ forecasts and appropriate modeling is needed to remove these bias structures which may vary in space and time, or just space.

3 Modeling developments

Following Sahu *et al.* (2007), the pollutant process is modeled as a high-resolution space-time process. Let $Z(\mathbf{s}, t)$ denote the square-root of the monitor observation in location $\mathbf{s}$ and at time t, $U(\mathbf{s}, t)$, $t = 1, \ldots, T$. Further, let $O(\mathbf{s}, t)$ denote the true value corresponding to $Z(\mathbf{s}, t)$. We develop models for data from n stations denoted by $\mathbf{s}_1, \ldots, \mathbf{s}_n$, for a running window of $T = 7$ seven days.

The monitoring data are assumed to represent the true ambient levels with random measurement error, but no bias. Expressed as a probability distribution:

$$Z(\mathbf{s}_i, t) = O(\mathbf{s}_i, t) + \epsilon(\mathbf{s}_i, t), \tag{2}$$

for $i = 1, \ldots, n$, $t = 1, \ldots, T$, where $\epsilon(\mathbf{s}_i, t)$ is a white noise process, assumed to follow $N(0, \sigma^2_\epsilon)$ independently. Thus σ^2_ϵ, taken to be homogeneous in space and time, is the so called nugget effect.

Next, we turn to the modeling for $O(\mathbf{s}_i, t)$. Figure 2 shows that there is some auto-correlation between ozone measurements on successive days. That is why our model must include an auto-regressive term. As mentioned above, the Eta CMAQ forecast for the observation grid cell is also relevant for predicting the true ozone so we include a spatially varying regression term with the Eta CMAQ forecasts as predictors. Thus, we assume that,

$$O(\mathbf{s}_i, t) = \xi + \rho\, O(\mathbf{s}_i, t-1) + (\beta_0 + \beta(\mathbf{s}_i))\, x(\mathbf{s}_i, t) + \eta(\mathbf{s}_i, t), \tag{3}$$

for $i = 1, \ldots, n$, $t = 1, \ldots, T$ where ξ is a constant across space and time, $\rho\, O(\mathbf{s}_i, t-1)$ is the auto-regressive term with $0 < \rho < 1$, $(\beta_0 + \beta(\mathbf{s}_i))\, x(\mathbf{s}_i, t)$ is the spatially varying regression term and $\eta(\mathbf{s}_i, t)$ is a spatially correlated, but temporally independent error term. Inclusion of the spatially varying regression term introduces spatial non-stationarity in the model since covariance of $O(\mathbf{s}_i, t)$ and $O(\mathbf{s}_j, t)$ given $O(\mathbf{s}_i, t-1)$ and $O(\mathbf{s}_j, t-1)$ involves both $\mathbf{s}_i$ and $\mathbf{s}_j$, not only their absolute difference in distance. The above model becomes stationary when $\beta(\mathbf{s}) = 0$ for all $\mathbf{s}$. Inclusion of the term $\beta(\mathbf{s}_i)\, x(\mathbf{s}_i, t)$ will lead to a better fitting model than the sub-model corresponding to $\beta(\mathbf{s}) = 0$. However, forecasting and out of sample spatial interpolation may have increased variability due to these additional parameters. In Section 5 we use Bayesian model selection and cross-validation methods to make this decision. The auto-regressive models require an initial condition for $O_1(\mathbf{s}, 0)$, which for convenience we choose to be the grand mean of the data, see Sahu *et al.* (2007) for an alternative method with additional processes and parameters.

We shall use the following vector notations: $\mathbf{Z}_t = (Z(\mathbf{s}_1, t), \ldots, Z(\mathbf{s}_n, t))'$, $\mathbf{O}_t = (O(\mathbf{s}_1, t), \ldots, O(\mathbf{s}_n, t))'$, and $\mathbf{x}_t = (x(\mathbf{s}_1, t), \ldots, x(\mathbf{s}_n, t))'$. Finally, we use X_t to denote a diagonal matrix whose ith diagonal entry is $x(\mathbf{s}_i, t)$. Now we write the above models using vectors and matrices to facilitate computation. The first model equation is obtained from (2):

$$\mathbf{Z}_t = \mathbf{O}_t + \boldsymbol{\epsilon}_t, \tag{4}$$

for $t = 1, \ldots, T$, where $\boldsymbol{\epsilon}_t = (\epsilon(\mathbf{s}_1, t), \cdots, \epsilon(\mathbf{s}_n, t))'$. Let $\mathbf{1}$ be the vector of dimension n with all elements unity and $\boldsymbol{\beta} = (\beta(\mathbf{s}_1), \ldots, \beta(\mathbf{s}_n))'$. From (3) we have:

$$\mathbf{O}_t = \xi \mathbf{1} + \rho\, \mathbf{O}_{t-1} + \beta_0 \mathbf{x}_t + X_t \boldsymbol{\beta} + \boldsymbol{\eta}_t, \tag{5}$$

for $t = 1, \ldots, T$, where $\boldsymbol{\eta}_t = (\eta(\mathbf{s}_1, t), \ldots, \eta(\mathbf{s}_n, t))'$.

For the measurement error in (4) we assume that $\boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \sigma^2_\epsilon I_n)$, $t = 1, \ldots, T$, independently, where $\mathbf{0}$ is the vector with all elements zero and I_n is the identity matrix of order n. For the spatially correlated error we assume that $\boldsymbol{\eta}_t \sim N(\mathbf{0}, \Sigma_\eta)$, $t = 1, \ldots, T$ where Σ_η has elements $\sigma_\eta(i, j) = \sigma^2_\eta \rho_\eta(\mathbf{s}_i - \mathbf{s}_j; \phi_\eta)$. We take $\rho_\eta(\mathbf{s}_i - \mathbf{s}_j; \phi_\eta) = \exp(-\phi_\eta d(\mathbf{s}_i, \mathbf{s}_j))$ where $d(\mathbf{s}_i, \mathbf{s}_j)$ is the distance between sites $\mathbf{s}_i$ and $\mathbf{s}_j$, $i, j = 1, \ldots, n$. We acknowledge the simplification associated with choosing the exponential covariance structure, however, other members of the Matèrn family of covariance functions can be chosen.

The spatially varying coefficients $\boldsymbol{\beta} \sim N(\mathbf{0}, \Sigma_\beta)$ where Σ_β has elements $\sigma_\beta(i,j) = \sigma^2_\beta \rho(\mathbf{s}_i - \mathbf{s}_j; \phi_\beta)$. The parameters ϕ_η and ϕ_β are determined using cross-validation as discussed in Section 5. For future use we define S_η and S_β by the relations:

$$\Sigma_\eta = \sigma^2_\eta S_\eta, \quad \Sigma_\beta = \sigma^2_\beta S_\beta.$$

Let $\boldsymbol{\vartheta}_t = \xi \mathbf{1} + \rho\, \mathbf{O}_{t-1} + \beta_0 \mathbf{x}_t + X_t \boldsymbol{\beta}$, for $t = 1, \ldots T$. Further, let $\boldsymbol{\theta}$ denote all the parameters, β_0, $\boldsymbol{\beta}$, ρ, σ^2_ϵ, σ^2_η, σ^2_β and ξ. Let $\mathbf{w}$ denote all the augmented data, $\mathbf{O}_t$ and the missing data, denoted by $z^*(\mathbf{s}_i, t)$, for $i = 1, \ldots, n$, $t = 1, \ldots, T$, and $\mathbf{z}$ denote all the non-missing data $z(\mathbf{s}_i, t)$, for $i = 1, \ldots, n$, $t = 1, \ldots, T$. The log of the posterior distribution, denoted by $\log \pi(\boldsymbol{\theta}, \mathbf{w}|\mathbf{z})$, can be written as

$$\begin{aligned}
&-\tfrac{nT}{2} \log(\sigma^2_\epsilon) - \tfrac{1}{2\sigma^2_\epsilon} \textstyle\sum_{t=1}^T (\mathbf{Z}_t - \mathbf{O}_t)'(\mathbf{Z}_t - \mathbf{O}_t) \\
&-\tfrac{nT}{2} \log(\sigma^2_\eta) - \tfrac{1}{2\sigma^2_\eta} \textstyle\sum_{t=1}^T (\mathbf{O}_t - \boldsymbol{\vartheta}_t)' S_\eta^{-1} (\mathbf{O}_t - \boldsymbol{\vartheta}_t) \\
&-\tfrac{n}{2} \log(\sigma^2_\beta) - \tfrac{1}{2\sigma^2_\beta} \boldsymbol{\beta}' S_\beta^{-1} \boldsymbol{\beta} + \log\left(\pi(\xi, \beta_0, \rho, \sigma^2_\epsilon, \sigma^2_\eta, \sigma^2_\beta)\right)
\end{aligned}$$

where $\pi(\xi, \rho, \beta_0, \sigma^2_\epsilon, \sigma^2_\eta, \sigma^2_\beta)$ denotes the prior distribution. To have a flat prior we assume that ξ and β_0 are independently normally distributed with mean 0 and variance 10^4. The auto-regressive coefficient ρ is also specified as the $N(0, 10^4)$ distribution, but restricted in the interval $I(0 < \rho < 1)$. The inverse of the variance components, $\frac{1}{\sigma^2_\epsilon}$, $\frac{1}{\sigma^2_\eta}$, $\frac{1}{\sigma^2_\beta}$ are assumed to follow $G(a, b)$ independently, where the distribution $G(a, b)$ has mean a/b. In our implementation we take $a = 2$ and $b = 1$ to have a proper prior specification for each of these variance components.

4 Prediction Details

We first develop the methods for spatial interpolation of the ozone levels at a new location $\mathbf{s}'$ and any time t, $t = 1, \ldots, T$. Details for one step-ahead forecasting at time $t = T + 1$ are given at the end of this section. Spatial interpolation at location $\mathbf{s}'$ and time t is based upon the predictive distribution of $Z(\mathbf{s}', t)$ given in the model equations (2) and (3). According to (2), $Z(\mathbf{s}', t)$, has the distribution:

$$Z(\mathbf{s}', t) \sim N\left(O(\mathbf{s}', t),\ \sigma^2_\epsilon\right) \tag{6}$$

and

$$O(\mathbf{s}',t) = \xi + \rho\, O(\mathbf{s}',t-1) + (\beta_0 + \beta(\mathbf{s}'))\, x(\mathbf{s}',t) + \eta(\mathbf{s}',t).$$

It is now clear that $O(\mathbf{s}',t)$ can only be sequentially determined using all the previous $O(\mathbf{s}',t)$ up to time t. Hence, we introduce the notation $\mathbf{O}(\mathbf{s},[t])$ to denote the vector $(O(\mathbf{s},1),\ldots,O(\mathbf{s},t))'$ for $t \geq 1$.

The posterior predictive distribution of $Z(\mathbf{s}',t)$ is obtained by integrating over the unknown quantities in (6) with respect to the joint posterior distribution, i.e.,

$$\begin{aligned}\pi\left(Z(\mathbf{s}',t)|\mathbf{z}\right) &= \int \pi\left(Z(\mathbf{s}',t)|O(\mathbf{s}',[t]),\sigma_\epsilon^2\right)\, \pi\left(O(\mathbf{s}',[t])|\beta(\mathbf{s}'),\boldsymbol{\theta},\mathbf{w}\right) \\ &\quad \pi\left(\beta(\mathbf{s}')|\boldsymbol{\theta}\right)\, dO(\mathbf{s}',[t])\, d\beta(\mathbf{s}')\, d\boldsymbol{\theta}\, d\mathbf{w}. \end{aligned} \tag{7}$$

When using MCMC methods to draw samples from the posterior, the predictive distribution (7) is sampled by composition. Draws from the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{z},\mathbf{w})$, and the conditional distributions $\pi\left(\beta(\mathbf{s}')|\boldsymbol{\theta}\right)$ facilitate evaluation of the above integral, details are provided below.

To sample $\beta(\mathbf{s}')$, we have

$$\begin{pmatrix} \beta(\mathbf{s}') \\ \boldsymbol{\beta} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ \mathbf{0} \end{pmatrix}, \sigma_\beta^2 \begin{pmatrix} 1 & S_{\beta,12} \\ S_{\beta,21} & S_\beta \end{pmatrix}\right],$$

where $S_{\beta,12}$ is $1\times n$ with the ith entry given by $\exp(-\phi_\beta d(\mathbf{s}_i,\mathbf{s}'))$ and $S_{\beta,21} = S_{\beta,12}'$. Therefore,

$$\beta(\mathbf{s}')|\boldsymbol{\theta} \sim N\left(S_{\beta,12}S_\gamma^{-1}\boldsymbol{\beta}, \sigma_\beta^2\left(1 - S_{\beta,12}S_\beta^{-1}S_{\beta,21}\right)\right). \tag{8}$$

We draw $O(\mathbf{s}',t)$ from its conditional distribution given $\boldsymbol{\theta},\mathbf{w}$ and $O(\mathbf{s}',[t-1])$. Analogous to (5), we obtain for $t \geq 0$

$$\begin{pmatrix} O(\mathbf{s}',t) \\ \mathbf{O}_t \end{pmatrix} \sim N\left[\begin{pmatrix} \xi + \rho\, O(\mathbf{s}',t-1) + (\beta_0 + \beta(\mathbf{s}'))\, x(\mathbf{s}',t) \\ \xi\mathbf{1} + \rho\, \mathbf{O}_{t-1} + \beta_0\mathbf{x}_t + X_t\boldsymbol{\beta} \end{pmatrix},\ \sigma_\eta^2 \begin{pmatrix} 1 & S_{\eta,12} \\ S_{\eta,21} & S_\eta \end{pmatrix}\right]$$

where $S_{\eta,12}$ is $1\times n$ with the ith entry given by $\exp(-\phi_\eta d(\mathbf{s}_i,\mathbf{s}'))$ and $S_{\eta,21} = S_{\eta,12}'$. Hence,

$$O(\mathbf{s}',t)|\beta(\mathbf{s}'),\mathbf{O}_t,\boldsymbol{\theta},\mathbf{w} \sim N(\chi,\ \Lambda) \tag{9}$$

where $\Lambda = \sigma_\eta^2\left(1 - S_{\eta,12}S_\eta^{-1}S_{\eta,21}\right)$ and

$$\chi = \xi + \rho\, O(\mathbf{s}',t-1) + (\beta_0 + \beta(\mathbf{s}'))\, x(\mathbf{s}',t) + S_{\eta,12}S_\eta^{-1}\left(\mathbf{O}_t - \xi\mathbf{1} - \rho\, \mathbf{O}_{t-1} - \beta_0\mathbf{x}_t - X_t\boldsymbol{\beta}\right).$$

In summary, we implement the following algorithm to predict $Z(\mathbf{s}',t), t = 1,\ldots,T$.

1. Draw a sample $\boldsymbol{\theta}^{(j)}, \mathbf{w}^{(j)}, j \geq 1$ from the posterior distribution.

2. Draw $\beta^{(j)}(\mathbf{s}')$ using (8).

3. Draw $\mathbf{O}^{(j)}(\mathbf{s}', [t])$ sequentially using (9). Note that the initial value value $O^{(j)}(\mathbf{s}', 0)$ is a constant for all $\mathbf{s}'$.

4. Finally draw $Z^{(j)}(\mathbf{s}', t)$ from $N\left(O^{(j)}(\mathbf{s}', t),\ \sigma_\epsilon^{2(j)}\right)$.

The ozone concentration on the original scale is the square of $Z^{(j)}(\mathbf{s}', t)$. If we want the predictions of the smooth ozone concentration process without the nugget term we simply omit the last step in the above algorithm and square the realizations $\mathbf{O}^{(j)}(\mathbf{s}, t)$. We use the median of the MCMC samples and the lengths of the 95% intervals to summarize the predictions. The median as a summary measure preserves the one-to-one relationships between summaries for O and Z, and for O^2 and Z^2.

The one-step ahead Bayesian forecast at a location $\mathbf{s}'$ is given by the posterior predictive distribution of $Z(\mathbf{s}', T+1)$ which is determined by $O(\mathbf{s}', T+1)$. Note that using (9) we already have the conditional distribution of $O(\mathbf{s}', T)$ given $\beta(\mathbf{s}'), \mathbf{O}_t, \boldsymbol{\theta}$, and $\mathbf{w}$. We use model equation (3) to advance this conditional distribution one unit of time in future. The mean of the one step-ahead forecast distribution is given by $\xi + \rho\, O(\mathbf{s}', T) + (\beta_0 + \beta(\mathbf{s}'))\, x(\mathbf{s}', T)$, according to (3), and $O(\mathbf{s}', T+1)$ should be equal to this if we are interested in forecasting the mean. If, however, we want to forecast an observation at location $\mathbf{s}'$ we simulate $O(\mathbf{s}', T+1)$ from the marginal distribution which has the above mean and variance σ_η^2. We work with this marginal distribution rather than the conditional distribution since conditioning with respect to the observed information (i.e. kriging) upto time T at the observation locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$ has already been done in (9), and at the future time $T+1$ there is no new available information to condition on except for the Eta CMAQ output as regressor values. Then we follow the above algorithm and the MCMC output summarization methods to evaluate the forecasts.

5 Analysis

Under weak prior distributions it is not possible to estimate all the parameters in the covariance structure, $\sigma^2_\epsilon, \sigma^2_\eta, \sigma^2_\beta, \phi_\eta$ and ϕ_β consistently, see e.g. Zhang (2004). Moreover, Stein (1999) shows that spatial interpolation is sensitive to the product $\sigma^2\phi$ but not to either one individually. Hence, we use the set-aside validation data from 40 stations to select the three decay parameters ϕ_η and ϕ_β. The variance components are estimated using MCMC conditional on these values. Let $\hat{Z}^2(\mathbf{s}^*_i, t)$ denote the model based validation estimate for $Z^2(\mathbf{s}^*_i, t)$ where $\mathbf{s}^*_i$ denote the ith validation site. Again recall that we model ozone in the square root scale. The validation mean-square error is given by

$$\text{VMSE} = \frac{1}{n_v}\sum_{i=1}^{40}\sum_{t=1}^{T}\left(Z^2(\mathbf{s}^*_i, t) - \hat{Z}^2(\mathbf{s}^*_i, t)\right)^2 I(Z(\mathbf{s}^*_i, t))$$

where $I(Z(\mathbf{s}^*_i, t)) = 1$ if $Z(\mathbf{s}^*_i, t)$ has been observed and 0 otherwise, and n_v is the total number of available observations at the 40 validation sites. We searched for the optimal values in a three dimensional grid formed of the values 0.005, 0.01 and 0.05, and 0.10. The values $\phi_\eta = 0.01$ and $\phi_\beta = 0.05$ provided the smallest estimated VMSE. Although it is possible to further refine the grid in a neighborhood of the best value we do not explore beyond our grid here.

The spatially varying regression term, $\beta(\mathbf{s}_i)x(\mathbf{s}_i, t)$, although quite attractive theoretically, did not improve the model fitting a great deal. Only a few $\beta(\mathbf{s}_i)$ were significant. We also used the predictive Bayesian model selection criterion of Gelfand and Ghosh (1998) to help make this decision. The criterion was much smaller for the model without the $\beta(\mathbf{s}_i)x(\mathbf{s}_i, t)$ term. Any sort of local lack of model fit for not including $\beta(\mathbf{s}_i)x(\mathbf{s}_i, t)$ gets compensated by the spatio-temporally varying intercept term $\eta(\mathbf{s}_i, t)$. Exclusion of the $\beta(\mathbf{s}_i)x(\mathbf{s}_i, t)$ term, however, *does not* mean that there is no spatio-temporal bias in the Eta CMAQ output – such biases can simply be recovered by the differences between the model based predictions and the Eta CMAQ output. If the intention is to recover the bias using a parametric form then a model omitting the spatially varying intercept and the auto-regressive term $\rho O(\mathbf{s}_i, t-1)$ should be considered.

Figure 3 provides the scatter plot of validation predictions using both the $\beta(\mathbf{s}) = 0$ and the $\beta(\mathbf{s}) \neq 0$ model against the corresponding actual observations. The validation

predictions are for the spatial interpolation of the daily 8-hour maximum ozone values at the 40 validation sites for the seven modeled days August 2–8 and forecast for the next day August 9th. The Eta CMAQ forecasts and the 45^o line are also superimposed in the figure. Clearly, the $\beta(\mathbf{s}) = 0$ model gives better predictions than the other two. In fact, the mean square errors for the Eta CMAQ forecasts is 229.6, the validations for the $\beta(\mathbf{s}) \neq 0$ model and the $\beta(\mathbf{s}) = 0$ model are 229.6, 84.4 and 50.5 respectively. This clearly shows that the $\beta(\mathbf{s}) = 0$ model is much superior than the others. We also have obtained similar plots for the models fitted to data sets obtained from a running window of seven days and forecasting the next day for each of the days during August 10–13. The plots looked similar and the mean square errors were same sort of magnitude apart, see Table 1. Henceforth, we worked with the sub-model corresponding to $\beta(\mathbf{s}) = 0$.

We examine the model performance in more detail in Table 2. In the table, we provide the hit and false alarm rates for the Eta CMAQ and the predictions using our chosen model. Here, hit is defined as the event where both the validation observation and the forecast for it were either both greater or less than 75 ppb, the current O_3 standard. The false alarm, on the other hand, is defined as the event where the actual observation is less than 75 ppb but the forecast is greater than 75 ppb. From the table we see that the model hit rate is more than 90% whereas the Eta CMAQ hit rate is about 80%. The false alarm rate for Eta CMAQ is about 20% compared to it being less than 5% for the proposed model. We repeated the calculations for the threshold values 80 ppb and 70 ppb in place of the above value of 75 ppb. In both the cases the model outperformed the Eta CMAQ forecasts by substantial margins. This is expected since the Eta CMAQ over-estimates the true values and this is observed more for ozone values which are lower than the extremes.

Figures 4 and 5 illustrate the forecast maps for August 9 and 12. As expected, the Eta CMAQ maps show higher daily ozone levels than the observed values. It is clear that the observations are closer to the forecasts than the Eta CMAQ values. The lengths of the 95% forecast intervals for August 9th and 12th are shown in Figure 6. By comparing the model based forecast map (in Figures 4 and 5) and the corresponding length map in Figure 6 we see that on average the higher forecast levels are associated with larger forecast lengths, as is often observed in environmental data.

6 Discussion

We have developed a space-time model in a Bayesian framework that uses both real-time air monitoring data and numerical Eta CMAQ output for forecasting spatial patterns of next day daily 8-hour maximum ozone concentrations across the eastern US. For a two week test period, we have shown model validation results that indicate this model improves upon the forecast results based on sole use of Eta CMAQ forecast output. Moreover, we can attach prediction uncertainties to all of these forecasts. This model appears to have great potential for use in USEPA's AIRNow web site to better inform the U.S. public of next day ozone levels. Accurate air quality information can offer significant health benefits, particularly for people with respiratory diseases, by leading to better environmental decisions. A companion paper, Sahu *et al.* (2008), focuses on predicting eight-hour average ozone levels based on predictions for the previous four hours, current hour, and forecasts for the next three hours.

References

Brown, P. J., Le, N. D., Zidek, J. V. (1994). Multivariate spatial interpolation and exposure to air pollutants. *The Canadian Journal of Statistics*, **22**, 489–510.

Carroll, R. J., Chen, R., George, E. I., Li, T.H., Newton, H.J., Schmiediche, H. and Wang, N. (1997). Ozone exposure and population density in Harris County, Texas. *Journal of the American Statistical Association*, **92**, 392–404.

Cowles, M. K. and Zimmerman, D. L. (2003). A Bayesian space-time analysis of acid deposition data combined from two monitoring networks. *Journal of Geophysical Research-Atomospheres*. **108**, doi: 10.1029/2003JD004001.

Eder, B., Kang, D., Mahur, R., Yu, S., and Schere, K. (2006). An operational evaluation of the Eta-CMAQ air quality forecast model. *Atmospheric Environment*, **40**, 4894–4905.

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.

Guttorp, P., Meiring, W. and Sampson, P. D. (1994). A Space-time Analysis of Ground-level Ozone Data. *Environmetrics*, **5**, 241–254.

Gelfand, A. E. and Ghosh, S. K. (1998). Model Choice: A Minimum Posterior Predictive Loss Approach. *Biometrika*, **85**, 1–11.

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

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

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

Sahu, S. K. and Mardia, K. V. (2005). A Bayesian Kriged-Kalman model for short-term forecasting of air pollution levels. *Journal of the Royal Statistical Society*, Series C, **54**, 223–244.

Sahu, S. K., Gelfand, A. E. and Holland, D. M. (2006). Spatio-temporal modeling of fine particulate matter. *Journal of Agricultural, Biological, and Environmental Statistics,* **11**, 61–86.

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., Yip, S., and Holland, D. M. (2008). A fast Bayesian method for updating and forecasting hourly ozone levels. Technical Report, University of Southampton.

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

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

Yu, S., Mathur, R., Schere, K., Kang, D., Pleim, J., Otte, T. (2007). A detailed evaluation of the Eta-CMAQ forecast model performance for O3, its related precursors, and meteorological parameters during the 2004 ICARTT study. *Journal of Geophysical Research*, **112**, D12S14, doi:10.1029/2006JD007715.

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

Zimmerman, D. L. and Holland, D. M. (2005). Complementary co-kriging: spatial prediction using data combined from several environmental monitoring networks. *Environmetrics*, **16**, 219–234.

Appendix: Distributions for Gibbs sampling

Conditional Distributions for: σ_ϵ^2, σ_η^2, $\mathbf{O}_t$, ρ, β_0 and $\boldsymbol{\beta}$

Any missing value, $Z(\mathbf{s},t)$ is to be sampled from $N(O(\mathbf{s},t),\sigma_\epsilon^2)$, $t=1,\ldots,T$. Straightforward calculation yields the following complete conditional distributions:

$$\begin{aligned}
&\tfrac{1}{\sigma_\epsilon^2} \sim G\left(\tfrac{nT}{2}+a, b+\tfrac{1}{2}\textstyle\sum_{t=1}^{T}(\mathbf{Z}_t-\mathbf{O}_t)'(\mathbf{Z}_t-\mathbf{O}_t)\right),\\
&\tfrac{1}{\sigma_\eta^2} \sim G\left(\tfrac{nT}{2}+a, b+\tfrac{1}{2}\textstyle\sum_{t=1}^{T}(\mathbf{O}_t-\boldsymbol{\vartheta}_t)'S_\eta^{-1}(\mathbf{O}_t-\boldsymbol{\vartheta}_t)\right),\\
&\tfrac{1}{\sigma_\beta^2} \sim G\left(\tfrac{n}{2}+a, b+\tfrac{1}{2}\boldsymbol{\beta}'S_\beta^{-1}\boldsymbol{\beta}\right).
\end{aligned}$$

Let $Q_\eta=\Sigma_\eta^{-1}$. The full conditional distribution of $\mathbf{O}_t$ is $N(\Lambda_t\boldsymbol{\chi}_t,\,\Lambda_t)$ where

Case 1: For $1\le t<T-1$:

$$\Lambda_t^{-1}=\frac{I_n}{\sigma_\epsilon^2}+(1+\rho^2)Q_\eta,$$

$$\boldsymbol{\chi}_t=\frac{\mathbf{Z}_t}{\sigma_\epsilon^2}+Q_\eta\left\{\xi\mathbf{1}+\rho\mathbf{O}_{t-1}+\beta_0\mathbf{x}_t+X_t\boldsymbol{\beta}+\rho\left(\mathbf{O}_{t+1}-\xi\mathbf{1}-\beta_0\mathbf{x}_{t+1}-X_{t+1}\boldsymbol{\beta}\right)\right\}.$$

Case 2: For $t=T$

$$\Lambda_t^{-1}=\frac{I_n}{\sigma_\epsilon^2}+Q_\eta,$$

$$\boldsymbol{\chi}_t=\frac{\mathbf{Z}_t}{\sigma_\epsilon^2}+Q_\eta\left\{\xi\mathbf{1}+\rho\mathbf{O}_{t-1}+\beta_0\mathbf{x}_t+X_t\boldsymbol{\beta}\right\}.$$

The full conditional distribution of ρ is $N(\Lambda\chi,\ \Lambda)$ where

$$\Lambda^{-1} = \sum_{t=1}^{T} \mathbf{O}'_{t-1} Q_\eta \mathbf{O}_{t-1} + 10^{-4},\ \chi = \sum_{t=1}^{T} \mathbf{O}'_{t-1} Q_\eta (\mathbf{O}_t - \xi\mathbf{1} - \beta_0 \mathbf{x}_t - X_t\boldsymbol{\beta}),$$

restricted in the interval (0, 1).

The full conditional distribution of β_0 is $N(\Lambda\chi,\ \Lambda)$ where

$$\Lambda^{-1} = \sum_{t=1}^{T} \mathbf{x}'_t Q_\eta \mathbf{x}_t + 10^{-4},\ \chi = \sum_{t=1}^{T} \mathbf{x}'_t Q_\eta (\mathbf{O}_t - \xi\mathbf{1} - \rho\mathbf{O}_{t-1} - X_t\boldsymbol{\beta}),$$

The full conditional distribution of $\boldsymbol{\beta}$ is $N(\Lambda\boldsymbol{\xi},\ \Lambda)$ where

$$\Lambda^{-1} = \sum_{t=1}^{T} X'_t Q_\eta X_t + \Sigma_\beta^{-1},\ \text{and}$$

$$\boldsymbol{\xi} = \sum_{t=1}^{T} X'_t Q_\eta (\mathbf{O}_t - \xi\mathbf{1} - \rho\mathbf{O}_{t-1} - \beta_0 \mathbf{x}_t).$$

The full conditional distribution of ξ is $N(\Lambda\chi,\ \Lambda)$ where

$$\Lambda^{-1} = \frac{1}{\sigma_\xi^2} + T\mathbf{1}'Q_\eta\mathbf{1},\quad \chi = \mathbf{1}'Q_\eta \sum_{t=1}^{T} \mathbf{a}_t$$

where $\mathbf{a}_t = \mathbf{O}_t - \rho\mathbf{O}_{t-1} - \beta_0\mathbf{x}_t - X_t\boldsymbol{\beta}$.

Disclaimer:

The U.S. Environmental Protection Agency's Office of research and Development partially collaborated in the research described here. Although it has been reviewed by EPA and approved for publication, it does not necessarily reflect the Agency's policies or views.

List of Tables

List of Figures

Table 1: Mean Square Errors

Validation Days	Eta CMAQ	$\beta(\mathbf{s}) \neq 0$	$\beta(\mathbf{s}) = 0$
Aug 2–9	229.6	84.4	50.5
Aug 3–10	246.4	58	50
Aug 4–11	260.5	77.8	64.5
Aug 5–12	253.4	99.1	62.1
Aug 6–13	240.6	72.5	45.4

Table 2: Hit and False Alarm percentages for O_3 exceeding 75 ppb.

Validation Days	Eta CMAQ		Model: $\beta(\mathbf{s}) = 0$	
	Hit	False Alarm	Hit	False Alarm
Aug 2–9	81.70	17.07	90.86	4.88
Aug 3–10	79.58	19.37	92.67	3.66
Aug 4–11	78.97	20.00	93.85	2.56
Aug 5–12	80.71	18.78	93.40	1.52
Aug 6–13	79.90	19.60	92.97	2.51

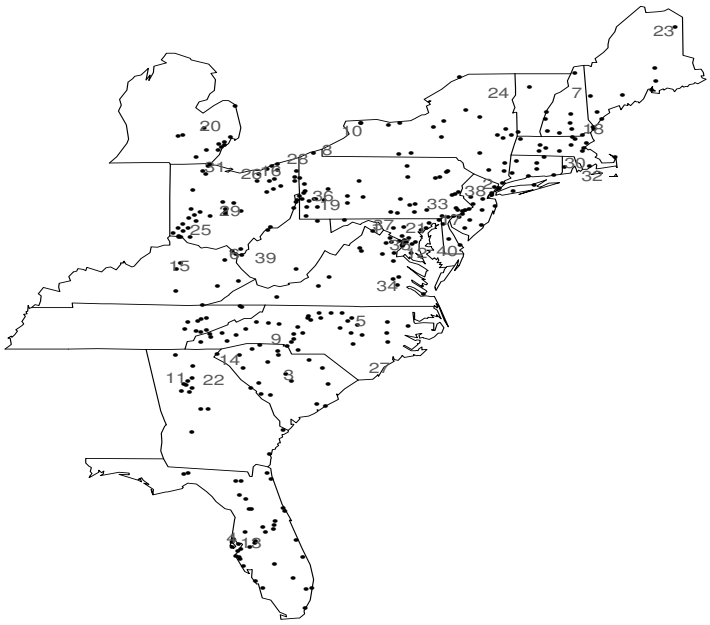

Figure 1: The 350 data and 40 validation sites (1 to 40).

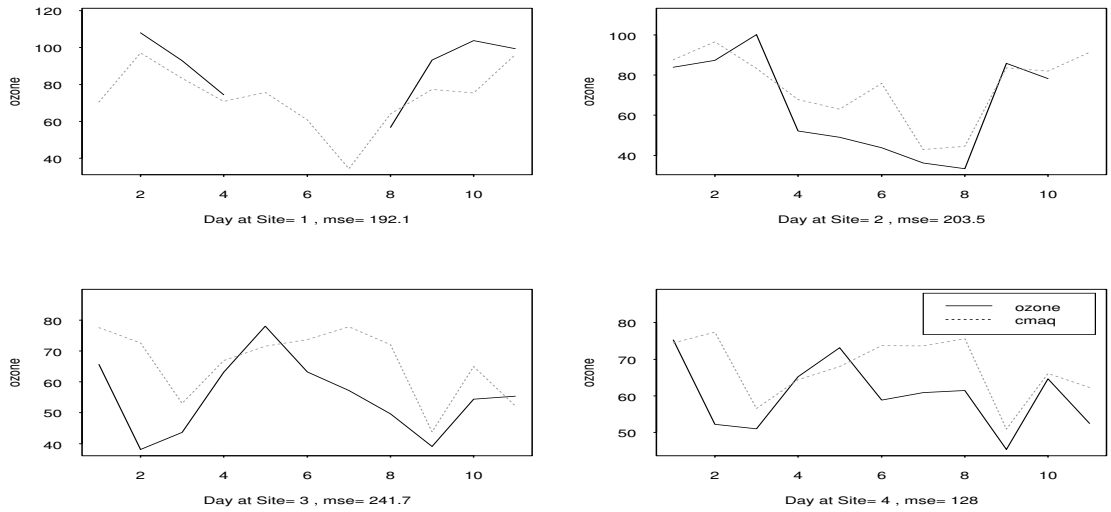

Figure 2: Observed data are dotted lines and Eta CMAQ forecasts are dashed lines at four sites. The mse for each plot is the mean square error between the data and the Eta CMAQ forecasts.

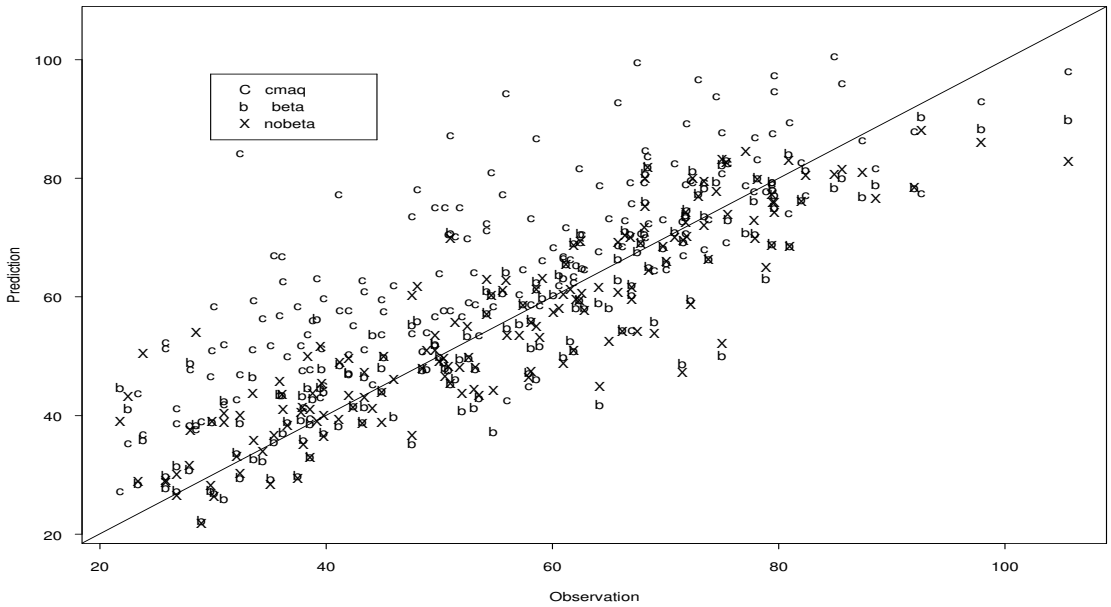

Figure 3: Scatter plot of validation predictions against observations The symbols 'C', 'b' and 'X' denote Eta CMAQ forecasts, the validation predictions from the model with non-zero $\beta(\mathbf{s})$ and the validations predictions from the model with $\beta(\mathbf{s}) = 0$, respectively.

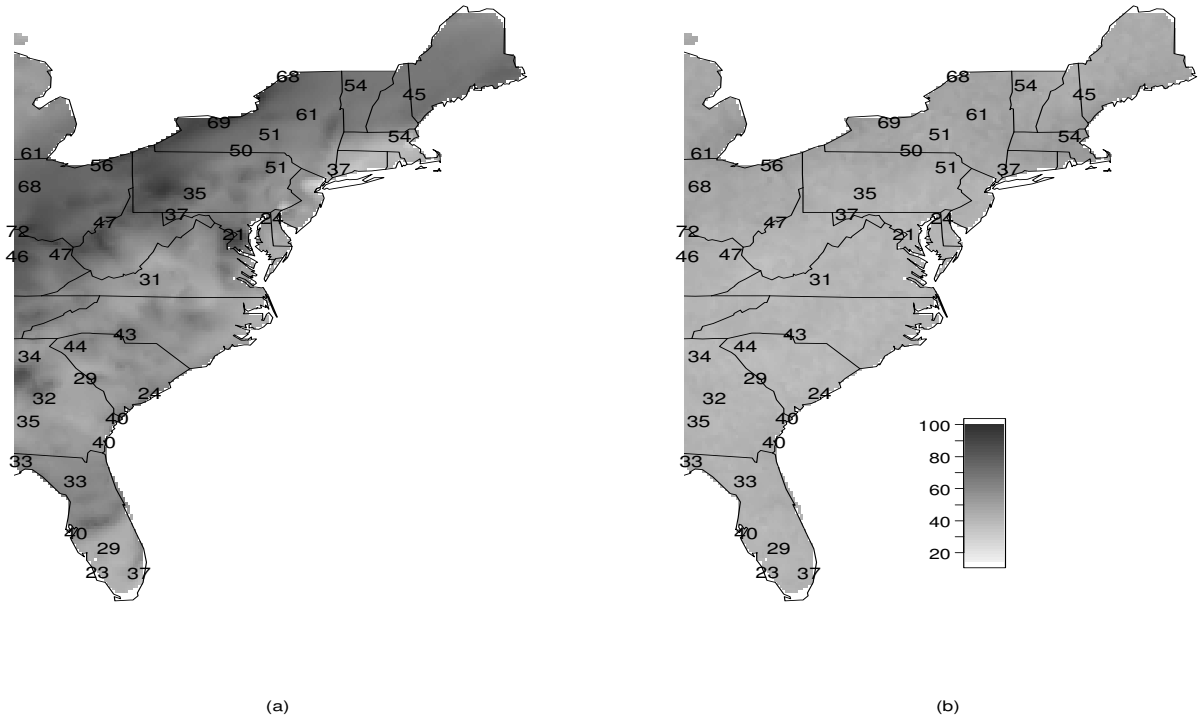

Figure 4: Forecast maps for August 9: (a) is for the Eta CMAQ and (b) is for the model with $\beta(\mathbf{s}) = 0$. Observed ozone values from some selected sites are superimposed. (For visual clarity we present only a subset of the monitoring data.)

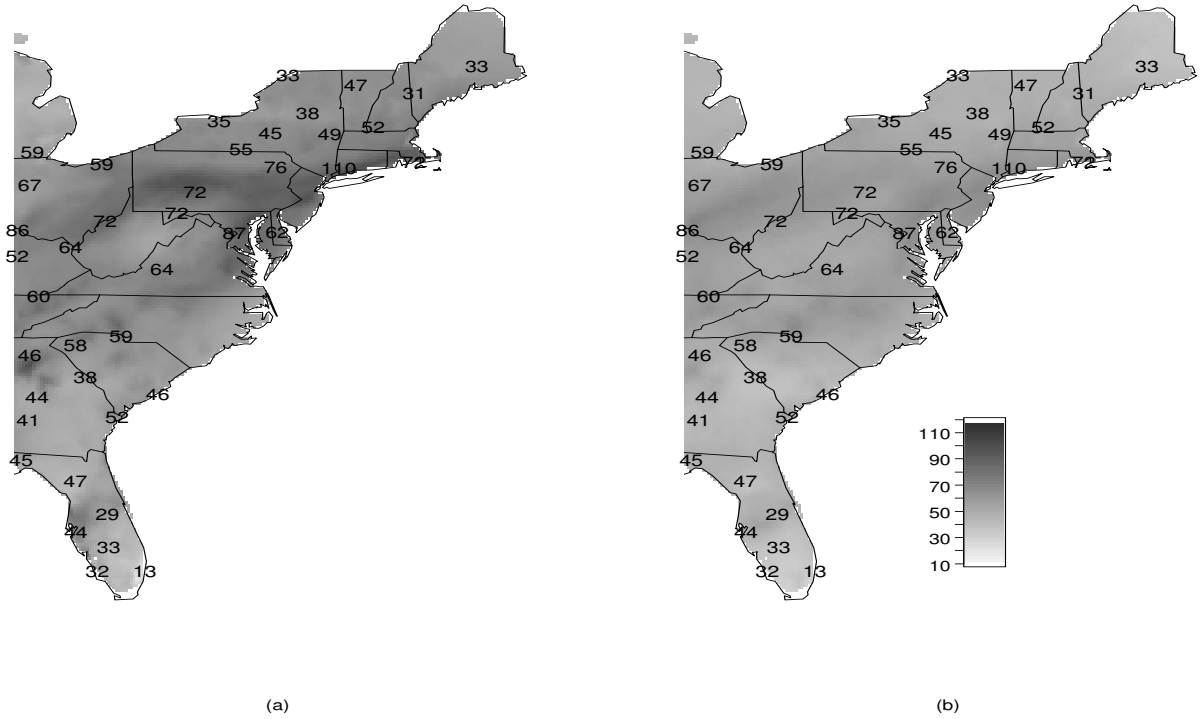

Figure 5: Forecast maps for August 12: (a) is for the Eta CMAQ and (b) is for the model with $\beta(\mathbf{s}) = 0$. Observed ozone values from some selected sites are superimposed. (For visual clarity we present only a subset of the monitoring data.)

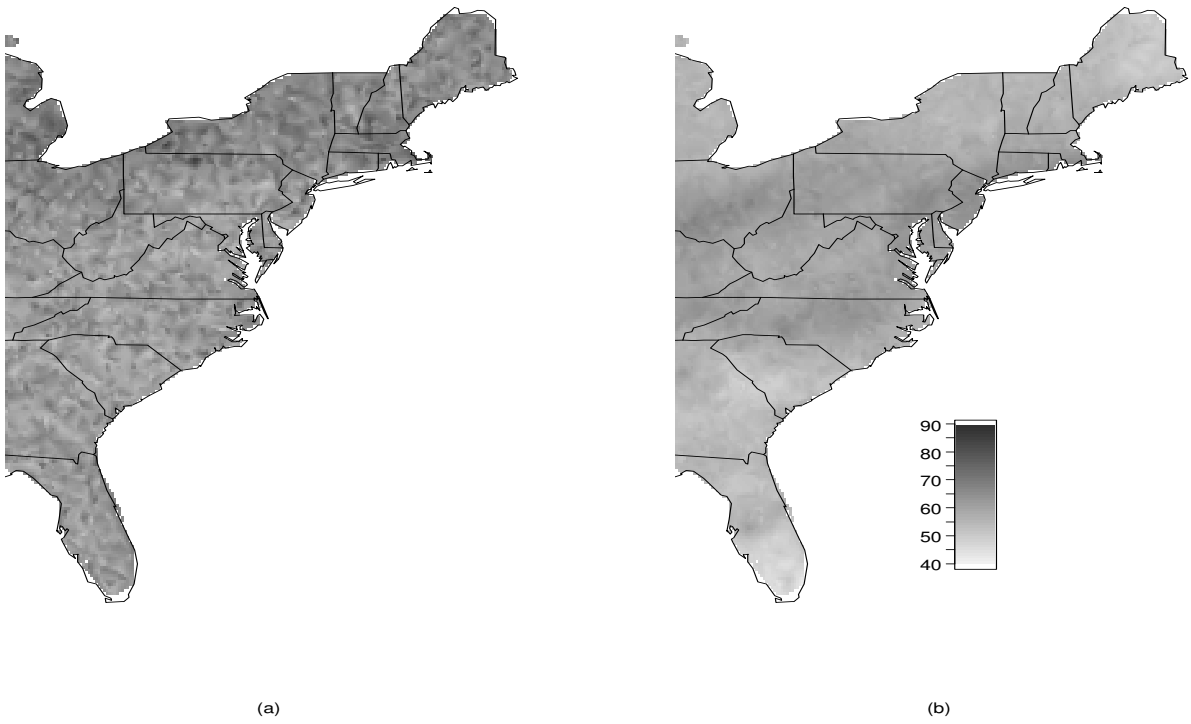

Figure 6: Lengths of the 95% intervals for the forecasts: (a) is for the forecast map on August 9th and (b) is for August 12th.