Available online at www.sciencedirect.com
SCIENCE DIRECT®
Atmospheric Environment 39 (2005) 1373–1382
ATMOSPHERIC ENVIRONMENT
www.elsevier.com/locate/atmosenv

A hierarchical Bayesian model to estimate and forecast ozone through space and time

Nancy McMillan[a,*], Steven M. Bortnick[a], Mark E. Irwin[b], L. Mark Berliner[c]

[a]*Battelle Memorial Institute, 505 King Avenue, Columbus, OH 43201, USA*
[b]*Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02138, USA*
[c]*The Ohio State University, Department of Statistics, 404 Cockins Hall, 1958 Neil Avenue, Columbus, OH 43210-1247, USA*

Received 8 December 2003; accepted 29 October 2004

Abstract

A Bayesian hierarchical regime switching model describing the spatial–temporal behavior of ozone (O_3) within a domain covering Lake Michigan during spring–summer 1999 is developed. The model incorporates linkages between ozone and meteorology. It is specifically formulated to identify meteorological regimes conducive of high ozone levels and allow ozone behavior during these periods to be different from typical ozone behavior. The model is used to estimate or forecast spatial fields of O_3 conditional on observed (or forecasted) meteorology including temperature, humidity, pressure, and wind speed and direction. The model is successful at forecasting the onset of periods of high ozone levels, but more work is needed to also accurately identify departures from these periods.

Keywords: Statistical model; Space–time models; Air pollution; Ozone; Meteorology

1. Introduction

The United States Environmental Protection Agency (EPA) is interested in developing statistical models describing the spatial–temporal behavior of ambient air pollutants such as ozone (O_3) and particulate matter. Statistical space–time models are useful for illuminating relationships between different air pollutants, quantifying linkages between air pollutants and meteorology, validating air pollution dispersion models, defining the spatial–temporal extent of episodes of dangerous air quality, forecasting urban and area-wide air pollution levels or indices, and evaluating the effectiveness of monitoring networks. This paper offers a hierarchical Bayesian approach to statistical space–time modeling. An application is made to O_3 levels within a domain covering Lake Michigan during spring–summer 1999. The model that is presented here can estimate or forecast spatial fields of O_3 conditional on observed (or forecasted) meteorology including temperature, humidity, pressure, and wind speed and direction.

Section 2 summarizes the data available for analysis. Section 3 summarizes the hierarchical Bayesian statistical model. Section 4 provides graphical depictions of results and model performance.

2. Data

A suitable dataset to conduct the statistical modeling was obtained from the EPA's Office of Air Quality Planning and Standards (OAQPS). These data consist of hourly and daily ozone measurements from EPA's ambient air monitoring program and hourly and daily

*Corresponding author. Fax: +1 614 424 4611.
E-mail address: mcmillann@battelle.org (N. McMillan).

1352-2310/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.atmosenv.2004.10.047

meteorological data from the National Weather Service™ (NWS). EPA's ambient air monitoring program is carried out by State and local agencies and consists of three major categories of monitoring stations including State and Local Air Monitoring Stations (SLAMS), National Air Monitoring Stations (NAMS), and Special Purpose Monitoring Stations (SPMS). The NWS provides weather, hydrologic, and climate forecasts for the United States; NWS data and products form a national information database and infrastructure that can be used by other governmental agencies, the private sector, the public, and the global community.

The domain of the study was restricted to a region covering Lake Michigan for the time period 16 April 1999, through 30 September 1999. This time period roughly corresponded to the ozone season when most of the area's ozone monitors provided data routinely. Fig. 1 provides a geographical summary of the spatial domain, indicated by the rectangular solid outline. Fifty-eight ozone monitoring stations are identified as solid stars and six meteorological stations are identified as solid squares. Ozone stations generally provide hourly data on a daily basis from April through September, whereas meteorological stations generally provide hourly data on a daily basis throughout the entire year. Overlain on this domain in Fig. 1 is a grid, where selected cells have been labeled, as discussed in Section 4.1.

The response variable of interest is the daily 8-h maximum ozone concentration, measured in parts per billion (ppb), and is modeled as a function of past ozone and current and past meteorological conditions. The daily 8-h maximum is derived from hourly ozone data by calculating the maximum of a day's running 8-h averages (i.e., the maximum of the 17 continuous 8-h periods that fall completely within one calendar day). Meteorological variables in the modeling effort include daily maximum temperature (degrees Fahrenheit), 24-h average station pressure (mb), 24-h average specific humidity ($g\,kg^{-1}$), and average surface wind speed ($m\,s^{-1}$) and wind direction (degrees). Average wind speed and wind direction were provided for two different 3-h intervals (i.e., 6–9 a.m. and 1–4 p.m.) each day. The greater of the two wind speeds, and its associated wind direction, were chosen to represent the daily wind speed and direction predictor variables. Note that wind direction was divided into eight 45° classifications (i.e., North, North-East, East, etc.) before being used as a predictor in the statistical model.

3. Fundamental modeling strategies and goals

3.1. Hierarchical Bayesian modeling

A fundamental notion in Bayesian statistical analysis is that unknown quantities are viewed as random variables. Hence, a critical challenge is the construction of probability distributions for the unknowns. These

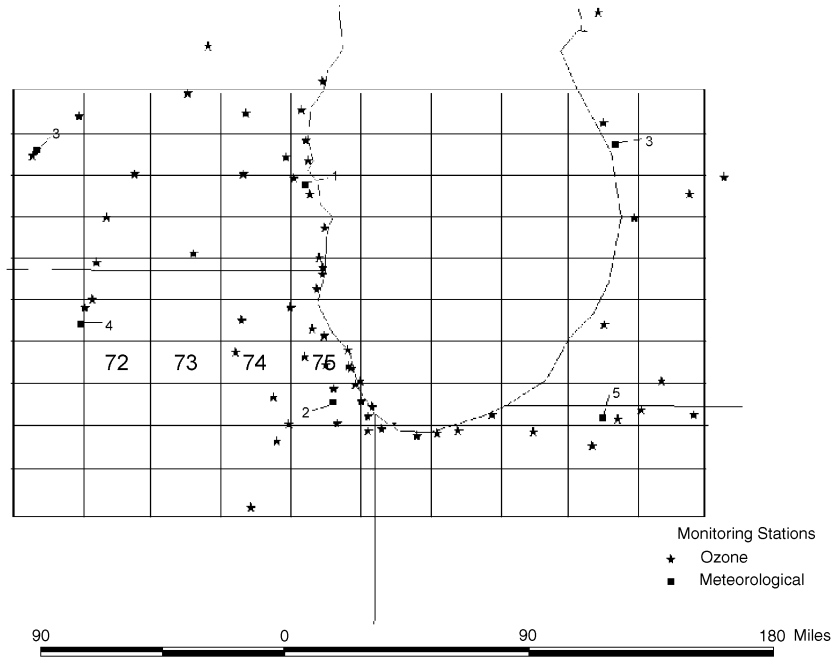

Fig. 1. Lake Michigan region O_3 and meteorological monitoring locations, spring–summer 1999.

distributions are then updated based upon observational data by applying Bayes' Theorem. While the construction of the requisite *prior distributions* is not easy in general, the potential gains in adopting the approach are substantial. First, we have the opportunity to incorporate prior information, reflected in both past data and scientific understanding, in an organized and clearly stated fashion. Second, since primary inputs to and outputs from a Bayesian analysis are probability distributions, the approach focuses on uncertainty description and management. In the context of environmental science, these features are particularly attractive (Berliner, 2000) because of the availability of (i) substantial scientific knowledge regarding the development, response to meteorological behavior, and impacts of environmental hazards; and (ii) observational data regarding both pollutants and meteorology. However, both information sources are subject to uncertainty.

A further limitation in our ability to predict environmental behavior is the complexity of the processes involved. Indeed, the interplay between the origins of uncertainty and the complexity of the environment blur. In the face of complexity, *Bayesian hierarchical modeling* (BHM) is suggested as a strategy for statistical modeling. The idea is easy: one constructs a high-dimensional joint probability distribution for all unknowns as the product of comparatively simple probability models. This strategy has proven particularly useful in space–time problems.

The following general view of a BHM as arising from three basic statistical models is useful (see Berliner, 1996, 2000):

1. *Data model* : distribution of observations, conditional on processes of interest and model parameters,
2. *Process model*: distribution of processes of interest given parameters,
3. *Parameter model*: prior distribution on all model parameters.

By *process*, we mean the actual physical variables of interest; in our setting, the processes include true ozone as represented by average ozone values for grid boxes over the Lake Michigan region. The separation of data and process models allows relatively simple and explicit (i) adjustment for uncertainty in the observations, and (ii) construction of process models including plausible scientific understandings of the interaction between processes, impacts of meteorology, and space–time dynamics.

3.2. Basic modeling strategies—assumptions

We develop models for space–time-gridded values of ozone in the Lake Michigan region during the summer of 1999. First, assume that this spatial domain is gridded, yielding N gridboxes. Time is also discretized and indexes days. Hence, the processes to be modeled can be written as N-vectors: $\mathbf{O}_t$. Elements of these vectors are viewed as spatial averages within gridboxes. Let O denote the space–time-gridded field of ozone over the study period: that is, $O = \{\mathbf{O}_t\colon t \in 0, \ldots, T\}$. Also, let $O^{\tau} = \{\mathbf{O}_t\colon t \in 0, \ldots, \tau\}$, that is, the history of the ozone process through time τ. A similar notation is used to define Z, $\mathbf{Z}_t$, Z^t, M, $\mathbf{M}_t$, and M^t, where Z denotes observational ozone data and M represents meteorological information.

We use the following notation to represent probability distributions: $[x|y]$ represents the distribution of x conditional on y; $[x]$ is the marginal distribution of x. In developing a BHM, we seek parameterized statistical models

$$[Z|O,\ M,\ \boldsymbol{\theta}_z], \tag{1}$$

$$[O|M,\ \boldsymbol{\theta}_o],\ \text{and} \tag{2}$$

$$[\boldsymbol{\theta}_z,\ \boldsymbol{\theta}_o], \tag{3}$$

where $\boldsymbol{\theta}_z$ are parameters of the data model (e.g., measurement error variances, etc.), and $\boldsymbol{\theta}_o$ are parameters in the *process model* for ozone conditional on meteorology.

In this article, we treat meteorological variables as fixed and known. In traditional statistical parlance, these variables are viewed as fixed covariates. Our rationale is explained below.

All models considered are forward or coincident in time (e.g., tomorrow's weather is not used to predict today's ozone, though today's and previous days' weather are used). Furthermore, we assume a first-order Markov model in time for ozone fields. That is, given meteorology, the distribution of $\mathbf{O}_t$ conditional on previous ozone depends only on the previous day's values $\mathbf{O}_{t-1}$. Formally, the ozone-given-meteorology model in Eq. (2) is of the form

$$[O|M,\ \boldsymbol{\theta}_o] = [\mathbf{O}_0|\mathbf{M}_0,\ \boldsymbol{\theta}_o]\prod_{t=1}^{T}[\mathbf{O}_t|\mathbf{O}_{t-1},\ \ M^t,\ \boldsymbol{\theta}_o]. \tag{4}$$

Several points merit clarification. First, the term $[\mathbf{O}_0,|\mathbf{M}_0,\ \boldsymbol{\theta}_o]$ is needed to "initialize" the times series model. That is, this prior cannot be identical to the prior for subsequent days, because there is no preceding day's ozone value. The second point involves our use and treatment of meteorology. As indicated by the conditioning on M^t in Eq. (4), our model uses coincident and past meteorology. In predictive contexts, this means we would need to predict $\mathbf{M}_t$ to predict $\mathbf{O}_t$. In our construction, we use the observed values $\mathbf{M}_t$, perhaps corrupted with noise, in this step. Our intent is that in operational uses of the model, one would use the meteorological predictions of meteorologists. The alter-

native would be for us to treat meteorology as random and build a statistical model for it. The approach we currently believe to be more appropriate is to adapt our analyses to incorporate weather predictions from meteorological prediction centers or other forecasters.

Finally, the first-order Markov assumption for ozone can be relaxed, though more general models lead to more computational overhead. Being treated as fixed, there are no specific limitations on the amount of past meteorology upon which we condition. The hope is that higher lags of meteorology serve as a partial surrogate for higher lags of ozone.

3.3. A local ozone model

We next describe a one-day lead forecast model for gridded ozone in a region of Chicago represented in $N = 100$ gridboxes (see Fig. 1). Following the general modeling approach described above, the model is constructed in three primary stages.

Data model: Daily ozone station data from the $n = 58$ stations in the domain are used. Let $\boldsymbol{Z}_t$ denote the n-vector of these data recorded on day t. The statistical model is

$$Z_t = K\mathbf{O}_t + \varepsilon_{z,t}, \tag{5}$$

where $\varepsilon_{z,t} \sim N(\mathbf{0}, \sigma_z^2 I)$; these vectors are assumed to be independent across time. Here, K is an $n \times N$ *mapping matrix* that maps stations to the gridbox that contains them. The model assumes that observations from all stations within the same gridbox have expected value equal to that gridbox ozone level. The errors $\varepsilon_{z,t}$ represent *both* measurement error and subgrid-scale variation of ozone. Such representations are extremely common in spatial statistics; the phenomenon is generally known as the *nugget effect*. For a fixed t, we assume that these errors are independent and have common, unknown variance across stations. The independence assumption should be verified, but it is important to note here that it is plausible. That is, while we expect that realizations of $\boldsymbol{Z}_t$ ought to be replete with spatial dependence, Eq. (5) is the conditional distribution of the observations given $\mathbf{O}_t$. As such, it need not explain that portion of the spatial dependence in the data arising from the spatial dependence present in $\mathbf{O}_t$. See Wikle et al. (1998) and Wikle et al. (2001) for general discussion and related examples.

Process model: Our task is to explicitly develop Eq. (4). This development requires some elaborations. The first of these involves spatial modeling. We expect ozone fields to display spatial–temporal dependence structures. To attack the spatial structure, we use a simple *nearest neighbor* dependence model in which the ozone at a particular site on a particular day depends upon the values at that site on the previous day and its four nearest neighboring sites on the same day. This strategy creates a difficulty; sites along the edges of the domain must be treated carefully. The option selected here is to extend the domain by introducing boundary sites and their associated ozone processes. That is, we consider a boundary process $\mathbf{B}_t$ for each t where $\mathbf{B}_t$ is the vector of 40 ozone values corresponding to the sites needed to complete the spatial model. These values are themselves endowed with a probability model, which includes uncertainty in the value of each boundary site and dependence between neighboring boundary sites; see Wikle et al. (2003) for general discussion. Each term in the product indicated in Eq. (4) is constructed as follows:

$$\mathbf{O}_t = \mu_t\mathbf{1} + H(\boldsymbol{\theta})\mathbf{O}_{t-1} + G(\boldsymbol{\theta})\mathbf{B}_{t-1} + M_t\beta + \varepsilon_{o,t}, \tag{6}$$

where μ_t is a time-varying mean shift and $\mathbf{1}$ is an N-vector of ones. $H(\boldsymbol{\theta})$ is an $N \times N$ matrix, parameterized to allow for nearest-neighbor spatial dependence among ozone elements. Specifically, for gridbox i, the conditional mean of O_t^i includes $\theta_5\, O_{t-1}^i + \Sigma_{j\in N}\theta_j O_t^j$, where N is the set of the four nearest neighbors of box i (see Royle and Berliner, 1999). $G(\boldsymbol{\theta})$ completes that model for edge sites. M_t is an $N \times d$ matrix whose elements contain relevant meteorological data; namely station observations of (i) daily maximum temperature, (ii) average humidity, (iii) atmospheric pressure, and (iv) wind speeds by direction. β can be thought of as d regression parameters for these covariates. $\varepsilon_{o,t} \sim N(\mathbf{0}, \sigma_o^2 I)$ are vectors assumed to be independent across time. The choice of independent errors is plausible since the mean of the model seeks to explain some spatial dependence through the nearest-neighbor structure as well as spatial dependence inherited from the spatial structure of meteorology.

One of the critical aspects of this model is the role played by the time-varying intercepts μ_t. The idea behind these terms arises from the notion that arrivals, persistence, and departures of synoptic scale weather systems, especially high-pressure systems, significantly impact ozone levels over the entire domain. Fig. 2 illustrates the association between increased ozone levels and high-pressure systems. Specifically, extreme ozone events typically happen only after a high-pressure system has persisted over the area for several days. The pressure signal plotted in Fig. 2 is recursively filtered, areally averaged pressure.

We model the arrival, persistence, and departure of high-pressure systems as a two-regime process: "normal" or typical behavior and "high pressure" system behavior. Conditional on which regime is active, μ_t is modeled as a first-order, autoregressive time series, with regime-dependent mean and autoregression parameters. Regime states and their transitions are then modeled as a first-order Markov chain, whose transition probabilities depend on a recursively filtered, areally averaged

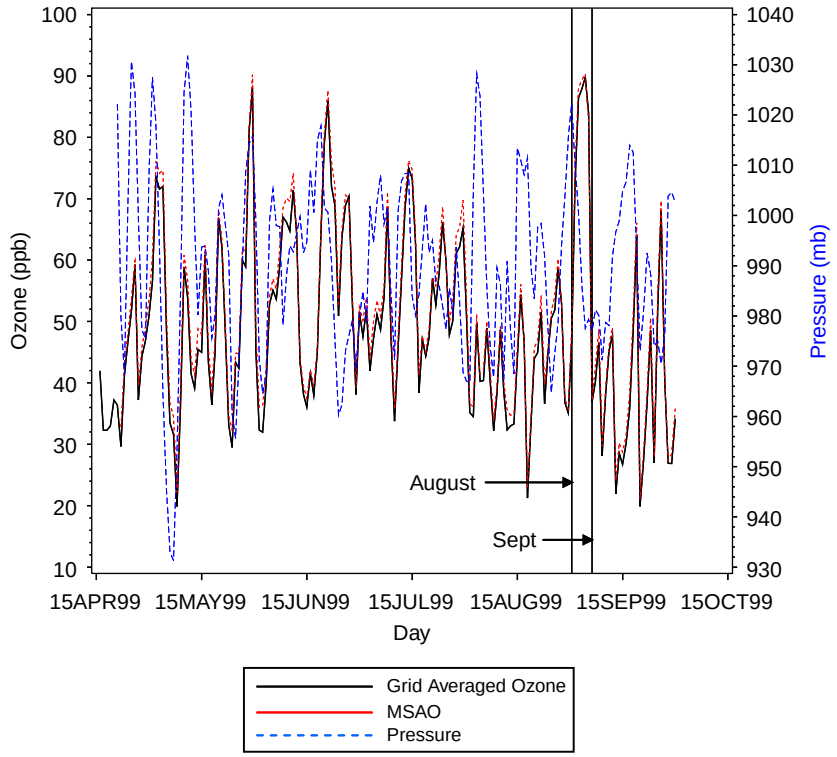

Fig. 2. Monitoring site averaged ozone (MSAO), recursively filtered, areally averaged pressure, and grid averaged ozone.

pressure series. This is an example of powerful methods of Bayesian hierarchical mixture modeling to account for regime switching; see Lu and Berliner (1999) and Berliner et al. (2000) for more discussion and examples.

Parameter prior models: Completion of the Bayesian model requires specification of prior distributions on the model parameters. Model parameters include meteorological regression parameters (β), ozone spatial dependence parameters (θ), autoregressive parameters for the daily mean process, parameters of the Markov process governing regime switching, mean and correlation parameters for the boundary ozone process, and a number of variance parameters capturing measurement error and uncertainty in each layer of the BHM.

In general, our priors tend to be quite vague, except for specific parameters for which vague priors led to poor posteriors. For example, failure to either encourage $\theta_1-\theta_4$ (ozone spatial dependence parameters) to be away from 0 or θ_5 to be near zero led to a poor trade-off between temporal and spatial variability; namely, today's ozone value in a grid cell was explained mostly by yesterday's ozone value in that grid cell, yielding negative and near-zero spatial correlations. Such a model would not allow efficient spatial interpolation of information from data-rich grid boxes to data-poor ones. Also, it was necessary to designate one of the regimes as normal and one as high pressure through specification of the priors on the autoregressive parameters for the daily mean process.

3.4. Analysis

The model described here is too complex to allow exact calculation. We relied on numerical simulation approaches, specifically Markov chain Monte Carlo (MCMC). For this application, a Gibbs' Sampler was implemented. The Markov chain was judged to have converged after 200 iterations and was run for an additional 800 iterations. Our results are based on the 800 iterations after convergence; the first 200 iterations were excluded from analysis as a burn-in phase. See Berliner et al. (2000) and Wikle et al. (2001) for discussion and examples.

4. Model estimates and predictions

This section summarizes the performance of the model in both estimating the spatial distribution of ozone and predicting the onset of ozone episodes. Fig. 2 illustrates the close agreement of BHM-estimated ozone to observed ozone by a time series plot of grid-averaged ozone estimates (posterior means) and monitoring site

averaged ozone observations. While these quantities would be expected to track each other closely as two measures of average ozone behavior over the region, they are inherently different measures. Specifically, the ozone monitoring sites are unevenly distributed over the region. They are concentrated at densely populated locations along the western shore of Lake Michigan. There is no ozone monitoring site over Lake Michigan. The grid-averaged ozone uniformly represents the region, including the unpopulated portion, Lake Michigan. Fig. 2 illustrates that grid-averaged ozone is systematically higher than monitoring site averaged ozone at both peaks and valleys of ozone behavior.

One aspect of the model fit that is of particular importance to environmental scientists is the inferred impact of meteorology on ozone levels. Before proceeding, we alert the reader that this is not an easy task. Such interpretations are difficult because of various model complexities. First, the predictive model uses both past ozone levels and meteorology. These variables are in some sense surrogates for each other (i.e., an issue often known as colinearity is active). Second, meteorology, particularly pressure, enters the model in a complicated and nonlinear fashion. Specifically, pressure is used as a conventional regression predictor and also in determining the aforementioned regimes. We believe that the predictive power of the final result justifies enduring such interpretative challenges.

As discussed in Sections 2 and 3, the grid cell ozone process model included the regression of ozone onto pressure, humidity, temperature, lagged temperature and eight wind-speed-by-prevailing-wind-direction variables as measured at the meteorology measurement site closest to each grid cell. Table 1 presents the estimated impact of each of these variables on grid cell ozone levels. We note from Table 1 that the Bayesian posterior 95% credible intervals for the coefficients of pressure and lagged temperature exclude zero, indicating that each of these meteorological variables has a significant impact on ozone. Examination of a scatter plot of temperature and pressure coefficient iterations reveals a much stronger correlation than observed among other sets of meteorological coefficients. Thus, the impact of temperature on ozone may be somewhat masked by pressure. Counterintuitively, the negative coefficient for pressure indicates that higher pressure is associated with lower ozone. The effect, however, is very small from a practical perspective. A 100 mb increase in pressure decreases ozone by less than 1 ppb. Lagged temperature also has a very small impact on ozone from a practical perspective.

Table 1
Estimated effects of meteorology on ozone process

Regressor	Estimate	Standard error
Pressure (mb 10^{-3})	−0.146	0.002
Humidity (g kg^{-1})	**0.226**	**0.119**
Temperature (°F)	**−0.047**	**0.028**
Lagged temperature (°F)	0.002	0.000
N (m s^{-1})	**−0.277**	**0.149**
NE (m s^{-1})	**−0.175**	**0.138**
E (m s^{-1})	**0.061**	**0.135**
SE (m s^{-1})	**−0.103**	**0.127**
S (m s^{-1})	**−0.023**	**0.144**
SW (m s^{-1})	**0.004**	**0.154**
W (m s^{-1})	**0.003**	**0.133**
NW (m s^{-1})	**0.129**	**0.154**

Bold numerals are not statistically significant.

The nature of observational meteorological data is such that the regression coefficient estimates are highly correlated. Posterior correlations between each pair of meteorological regression coefficients were estimated. There were strong correlations between temperature and pressure coefficient estimates and between wind speed/direction coefficients, suggesting that the marginal Bayesian posterior 95% credible intervals for these variables (Table 1) are not sufficient to determine that they have no impact on ozone.

Pressure also enters the ozone model through the areal mean ozone process, μ_t, which is governed by the Markov regime-switching model. Specifically, regime states (normal/high pressure) and their transitions are modeled as a first-order Markov chain, whose transition probabilities depend on a recursively filtered, areally averaged pressure series. The probability of a particular day being a high-pressure ozone day given yesterday's regime is plotted in Fig. 3. Regardless of the previous day's regime, the probability of a particular day being a high-pressure day increases with (recursively filtered, areally averaged) pressure. There are separate mean and auto-regressive parameters of the areal mean ozone process depending on the regime state. The mean and auto-regressive parameters for normal regime days are 51.57 ppb (with a standard error of 4.36 ppb) and 0.66 (with a standard error of 0.03), respectively; for high-pressure regime days, 64.51 ppb (with a standard error of 6.07 ppb) and 0.58 (with a standard error of 0.85), respectively. These parameters dictate higher ozone levels on high-pressure days and lower ozone levels on normal days. Specifically, there is a 13 ppb spread between the intercept of the areal mean ozone process under the high-pressure regime compared to the intercept of the areal mean ozone process under the normal regime. Temporal dependence (as measured by the autoregressive coefficient) under the normal regime is estimated to be very similar to temporal dependence under the high-pressure regime, but because of the small percentage of the days which are classified as high pressure, the variability associated with the auto-regressive coefficient of the high-pressure mean ozone process model is much larger. Investigation of the

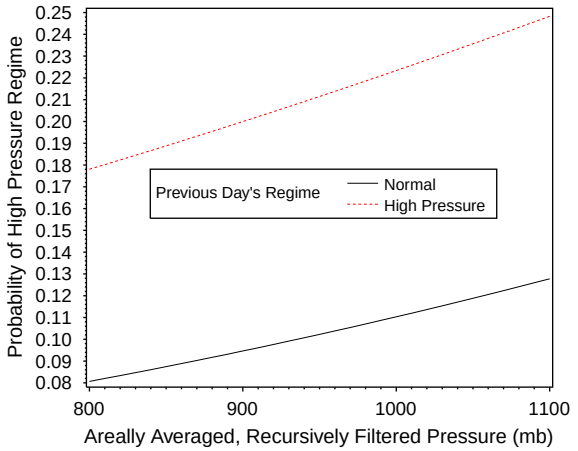

Fig. 3. Effect of areally averaged, recursively filtered pressure on probability of a particular day being a high-pressure regime day given previous day's regime.

posterior expectation of the mean process given the day's classification resulted in the determination that, for days with significant probability of being a high-pressure day, the mean process on high-pressure days was indeed higher than on normal days.

4.1. Spatial interpolation

Given the large number of days over which the model was applied, the discussion here is limited to a sequence of 7 days from 31 August to 6 September during the most significant area-wide ozone episode observed in the region for the 1999 season (See Fig. 2). Results in Fig. 4 correspond to fitting the model to the entire dataset and demonstrate the model's ability to spatially interpolate the response ozone. The grid cells of the spatial ozone model are shaded to indicate the model-estimated (posterior mean) ozone level. The observed ozone at each of the 58 ozone monitoring sites is depicted by the shading of the circle located at the monitoring site. Fig. 4 illustrates the model's ability to switch regimes and capture the large shifts in ozone behavior that occurred between 31 August and 1 September (days 132 and 133), and then again between 5 and 6 September (days 137 and 138).

One appealing feature of the proposed model for ozone is its ability to provide uncertainty estimates associated with both model estimates and predictions. The uncertainty associated with each of the grid cells of Fig. 4 varies based on the quantity and variability of the data within and around that cell. To illustrate this feature of the model, consider Fig. 5, in which model estimates of ozone are provided for four grid cells over the 7-day ozone episode from 31 August to 6 September. As was shown in Fig. 1, the four grid cells considered are 72, 73, 74, and 75 where grid cells are numbered from the upper left corner; for example, grid cell 72 is in the seventh row from the top and the second column from the left. The grid cells considered illustrate a wide range of data availability conditions. Grid cell 72 and its four nearest neighbors have no ozone monitoring sites within them. Grid cell 73 contains no monitoring sites, but its neighbor to the right has two. Grid cell 73 contains two monitoring sites. Grid cell 74 contains five monitoring sites. Fig. 5 illustrates clearly that there is much less uncertainty associated with ozone grid cell estimates when there are monitoring sites within the grid cell. The uncertainty decreases even further when there are multiple monitoring sites within the grid cell.

4.2. Temporal prediction

The excellent fit of the model to the observed monitoring data during the 31 August ozone episode, exhibited in Fig. 4, is not sufficient to demonstrate the model's ability to predict ozone episodes because the fit used the ozone and meteorological data on and after 31 August. This is not appropriate in a prediction scenario. Thus, additional MCMC runs were completed fitting the model up to a certain day, "forecasting" meteorology for the next day, then in turn "forecasting" and spatially interpolating the response ozone for that next day. Thus, fair predictive results were obtained.

Fig. 6 examines the predictive performance of the model. Specifically, the grid average of one-day ahead predicted ozone and observed monitoring site averaged ozone are graphed over the period of the 31 August to 5 September ozone episode. In order to better understand model behavior, the areal mean ozone process, μ_t, and the probability of the high regime are also plotted over this time period. This figure indicates that the model slightly under-predicts ozone at the onset of a switch in regime for one day, but that the model's regime switching component accurately captures the onset. The model appears to predict ozone levels during episodes quite well, once the episodes have commenced. The model's regime switching component captures the departure of the high-pressure system (on 3 September), which precedes the return to lower ozone levels by a few days (6 September).

Fig. 6 provides interesting insight into the interpretation of the regime model, which is helpful in understanding why the intercepts of the normal and high-pressure regime mean ozone processes differ by only 13 ppb. The regime model picks up the transition of ozone behavior from normal to high-pressure regime, but the high-pressure regime ends before the ozone drops, i.e., high-pressure events precede high ozone events. Thus, the transition to high ozone levels is classified as high-pressure regime and the anticipated

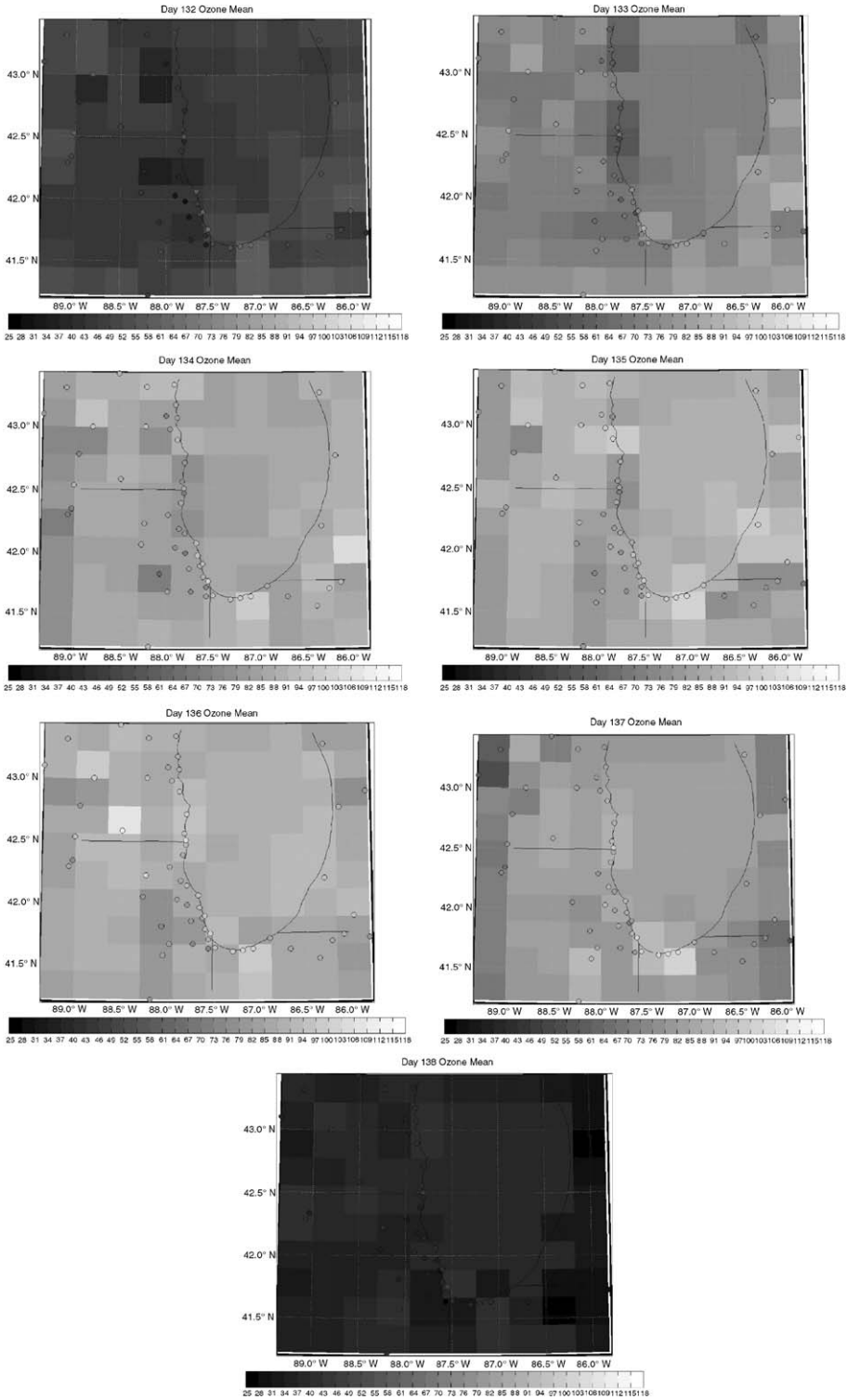

Fig. 4. Model estimates (spatial interpolation using full data set) for 31 August 1999 through 6 September 1999.

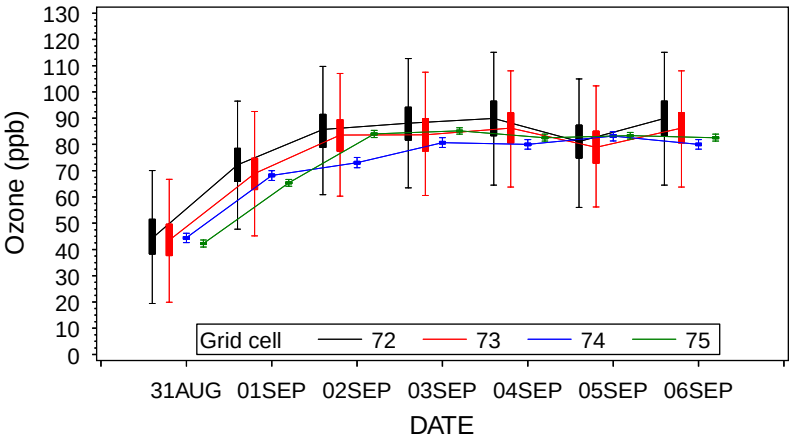

Fig. 5. Trend and variability associated with four grid cell ozone estimates during one high ozone episode.

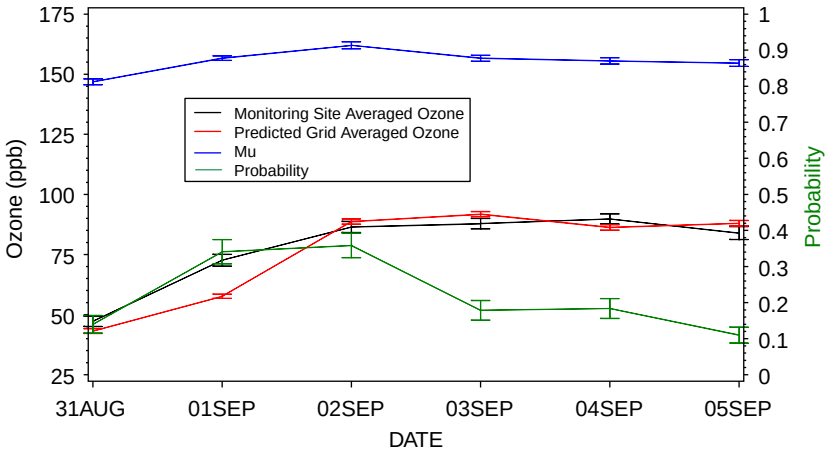

Fig. 6. Model predictive performance.

behavior is increasing ozone levels. The duration of high ozone levels is classified as normal regime and the anticipated behavior is steady, high ozone. There is no specific piece of the model that picks up on falling ozone levels after the high-pressure system has passed.

To improve the ability of the model to predict a regime switch ahead of time, several pressure signals were considered in the model. Specifically, the effect of the filter window location and the advantages of multiple pressure variables were investigated. The model proved to be quite insensitive to the exact pressure summary selected.

5. Conclusions

The Bayesian hierarchical model proposed simultaneously captures a number of key ozone behaviors. Spatial dependence is modeled in a manner that provides estimated ozone levels over the entire modeling domain based upon unevenly distributed monitoring data, while quantitatively accounting for uncertainty in those estimates. Uncertainty in ozone is observed to be greater in grid cells removed from monitoring sites. The meteorological dependence of ozone on temperature, humidity, pressure, and wind speed/direction is modeled. The inclusion of meteorology in the model is used to improve predictions of tomorrow's ozone through the use of forecast meteorology.

The pivotal ozone behavior captured in the proposed model is regime switching. The model postulates that there are two ozone behavior patterns. One behavior pattern is associated with a high-pressure system settling over the area. The other is active on "normal" pressure days. The posterior distribution describes the probability each individual day is a high-pressure day, as well as describing the behavior of the day's ozone. The onset of

ozone events of sufficient magnitude to raise health concerns can be forecast more accurately because of the regime switching model component.

Acknowledgements

The authors were supported in part by the US EPA's Office of Air Quality Planning and Standards (OAQPS) through Contract No. 68-D-98-030. Professor Berliner was supported in part by the US EPA's Science to Achieve Results (STAR) program, Assistance Agreement R827257-01-1. The authors wish to acknowledge helpful comments and suggestions from Bill Cox of OAQPS's Emissions, Monitoring, and Analysis Division and Shelley Eberly of EPA's Office of Research and Development (ORD) National Exposure Research Laboratory (NERL).

References

Berliner, L.M., 1996. Hierarchical Bayesian time series models. In: Hanson, K.M., Silver, R.N. (Eds.), Maximum Entropy and Bayesian Methods. Kluwer Academic Publishers (Springer), New York, pp. 15–22.

Berliner, L.M., 2000. Hierarchical Bayesian modeling in the environmental sciences. Allgemeines Statistisches Archiv, Journal of the German Statistical Society 84, 141–153.

Berliner, L.M., Wikle, C.K., Cressie, N., 2000. Long-lead prediction of Pacific SSTs via Bayesian dynamic modeling. Journal of Climate 13, 3953–3968.

Lu, Z.-Q., Berliner, L.M., 1999. Markov switching time series models with applications to a daily runoff series. Water Resources Research 35, 523–534.

Royle, J.A., Berliner, L.M., 1999. A hierarchical approach to multivariate spatial modeling and prediction. Journal of Agricultural, Biological, and Environmental Statistics 4, 1–28.

Wikle, C.K., Berliner, L.M., Cressie, N.A.C., 1998. Hierarchical Bayesian space–time analysis. Journal of Environmental and Ecological Statistics 5, 117–154.

Wikle, C.K., Milliff, R.F., Nychka, D., Berliner, L.M., 2001. Spatiotemporal hierarchical Bayesian blending of tropical ocean surface wind data. Journal of the American Statistical Association 96, 382–397.

Wikle, C.K., Berliner, L.M., Milliff, R.F., 2003. Hierarchical Bayesian approach to boundary value problems with stochastic boundary conditions. Monthly Weather Review 131, 1051–1062.