

Exploring Space-Time Structure in Ozone Concentration Using a Dynamic Process Convolution Model

Catherine A. Calder
Christopher Holloman
David Higdon

ABSTRACT Given daily ozone readings from 512 weather stations in the Eastern United States, we are interested in both predicting future ozone concentrations and in gaining insight into the space-time dependence structure of the data. We model ozone concentration as a process that moves across the region over time and exhibits spatial dependence locally in time. Our hope is to better understand the space-time dependence in ozone, so that this information can be used to assess the effectiveness of new monitoring network configurations. Process convolutions not only provide a framework for incorporating time dependence in spatial modeling, but also remain computationally tractable with large datasets. Standard dynamic linear modeling methods can be used to specify the time dependence allowing efficient posterior exploration. We consider a few variations of these space-time process convolution models that incorporate different levels of space-time dependence.

1 Introduction

Ozone concentration levels may depend on many factors including temperature, precipitation, moisture, solar radiation, wind, industrial activity, and auto emissions. Data on some of these variables are often unavailable or difficult to collect over large regions for many points in time. Even if these data can be collected, their effect on ozone levels over space and time can be difficult to specify through a statistical model. We propose a space-time model for ozone concentration that does not incorporate any of these explanatory variables directly, but rather models ozone as depending on an underlying latent process that varies over space and time. This latent process will relate the ozone level in a specific area to the levels in surrounding areas in the past.

Our goal is to be able to identify patterns that are constant over time in the relationship between the latent process and the observed ozone levels. For example, we might find that ozone levels in the Northeast depend heavily on the ozone levels in the Midwest in previous days. As a result, we would propose collecting data on weather variables that are known to move

east over time across the eastern U.S. and constructing a statistical model for ozone that includes them. Alternatively, if we find that ozone levels in the surrounding areas in the past appear to have little effect on the current levels but levels from the same area have a significant influence, we may conclude that variables that are location specific over time are important to consider. Such variables may include industrial activity and auto emissions as opposed to weather variables like wind or temperature that move across the spatial domain. Our model is to be viewed as preliminary search strategy for patterns in the variation of a spatial process over time that can help with model development.

In addition, the ozone level in some geographical regions may typically show dependence on previous ozone levels in nearby areas, while other regions do not show this sort of dependence. Such information could play an important role in the EPA's decision to modify the current ozone monitoring network.

We specify our spatial model using the process convolution method (Higdon, Swall, Kern, 1998). The framework of process convolutions can easily be extended to allow the spatial process to vary over time (Higdon, 2001) in such a way that posterior inference can be performed efficiently by utilizing the theory of dynamic linear models developed by West and Harrison (1997).

2 Process Convolutions

One method commonly used to create a spatial process over a region D is to define a set of points in the region $\mathbf{s} = (s_1, \dots, s_p)'$ and assign a joint normal distribution to the value of the process, ψ , at those points as in equation 2.1. The covariance structure of the process at these points is usually defined such that values of the process at points closer together in space have a higher covariance. For instance, we may choose to define the covariance between process values at locations s_i and $s_j \in D$ as in equation 2.2 using the Euclidean metric denoted by $d(s_i, s_j)$.

$$f(\psi(\mathbf{s} \mid \Sigma)) = (2\pi)^{-p/2} |\Sigma|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} \psi^T \Sigma^{-1} \psi \right\} \quad (2.1)$$

$$\Sigma_{ij} = C(\psi(s_i), \psi(s_j)) = \exp \{ -\theta d(s_i, s_j)^2 \} \quad (2.2)$$

Alternatively, it is possible to create a continuous spatial process over a region by convolving a continuous white noise process with a convolution kernel k (Higdon, 1998). A continuous white noise process x with precision λ_x is one in which

$$Z_A = \int_A x(u) du \sim N(0, \lambda_x^{-1} \text{area}(A)) \quad \forall A \in D \quad (2.3)$$

and

$$\text{cor}(Z_A, Z_B) = \text{area}(A \cap B). \quad (2.4)$$

Thus, we can create a continuous spatial process over the region D defined at any point s by

$$\psi(s) = \int_D k(u - s)x(u)du. \quad (2.5)$$

If k is an isotropic Gaussian kernel, there is a one-to-one mapping between the continuous field defined via a process convolution and a Gaussian process with a covariance function depending only on the distance between two points as in equation 2.2 (Higdon, 2001). In our work, the convolution kernel is the circular normal (see equation 2.6). Because this kernel is symmetric, it creates an isotropic process over the two-dimensional field D . As a result, the kernel only depends on the distance between the evaluation point s and the location of the white noise processes, x .

$$k(\mathbf{u}) = (2\pi\sigma^2)^{-1} \exp\left\{-\frac{1}{2\sigma^2}\mathbf{u}'\mathbf{u}\right\} \quad (2.6)$$

In practice it is very difficult to work with the continuous white noise process. Instead, we use a white noise process defined on a lattice denoted by $\omega = (\omega_1, \dots, \omega_c)'$ that covers the region. In discretizing the process, it is important to consider possible edge effects; evaluations of the field at some point s near the edges of the lattice can be unstable. To avoid such problems, it is useful to define the lattice beyond the edges of the region over which we make predictions as shown in figure 1. If we want to make estimates at the grid points, we could define the lattice as shown. Once the lattice has been defined, the value of the white noise process at location ω_i will be denoted $x(\omega_i)$. Then, equation 2.5 becomes

$$\psi(s | x) = \sum_{i=1}^c k(d(s, \omega_i))x(\omega_i). \quad (2.7)$$

As before, this equation represents a continuous process defined for any $s \in D$. As written, equation 2.7 corresponds to a zero mean Gaussian process. To adjust the mean level of the process, the equation can be written as $\psi(s | x) = \sum_{i=1}^c k(d(\omega_i, s))x(\omega_i) + \mu$.

A one-dimensional illustration of these concepts is shown in figure 2. The vertical lines represent the values of the white noise process $\mathbf{x}(\omega)$ at equally spaced points along the line. The continuous process that results from convolving the white noise values is included as a solid line, and the convolution kernel is included at the top of the figure.

If, for now, we consider locations $s \in D$ at which data have been recorded, it is possible to write equation 2.7 in matrix form (including an extra term to adjust the mean of the process) as

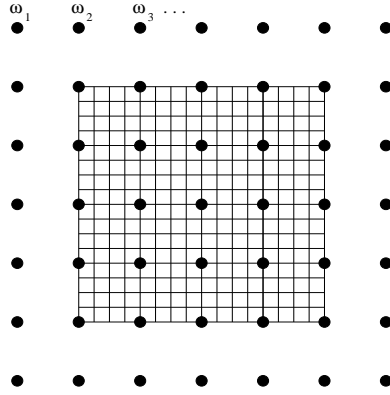


FIGURE 1. White noise process locations and a grid to be estimated.

$$\Psi = \mathbf{K}\mathbf{x} + \mu. \quad (2.8)$$

Here, Ψ denotes the vector $(\psi(s_1), \dots, \psi(s_p))'$, \mathbf{x} denotes the vector $(x(\omega_1), \dots, x(\omega_c))'$, and \mathbf{K} is an $p \times c$ matrix with $K_{ij} = k(d(s_i, \omega_j))$. In our space-time model for ozone, we will find this form particularly useful. Since we only have ozone readings at $p = 512$ locations, we can define those sites as our vector \mathbf{s} . We choose our lattice of white noise values to be a $c = 11 \times 6$ grid over the eastern United States (see figure 5 for the actual locations of the ω 's). One advantage of defining our process this way instead of as a joint normal distribution over all sites in which we are interested is the reduction of dimension of our model (from p to c). Hence, instead of inverting a $p \times p$ matrix, only a much less demanding $c \times c$ inversion is required. This greatly speeds up the MCMC used for posterior exploration. In addition, this process convolution approach readily gives predictive ozone concentrations at any locations beyond just the observed locations. After we have determined values for the white noise process x , we can construct a new matrix \mathbf{K}^* to predict ozone concentrations at any set of points in the region.

This simple spatial model can be extended to model space-time phenomena. We explain how this extension can be applied to this problem using dynamic linear model methods (West and Harrison, 1997) in the next section.

3 A Space-Time Model for Ozone Concentration

3.1 The Data

The ozone data set we analyzed using a model based on equation 3.12 consists of maximum eight hour average ozone concentrations taken over

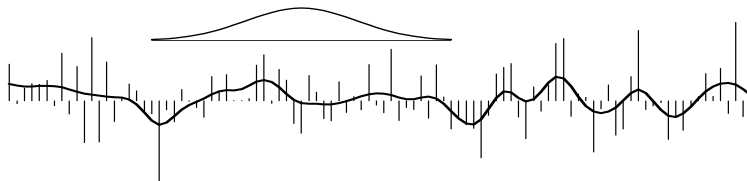


FIGURE 2. Smoothing kernel, white noise process, and resulting continuous process.

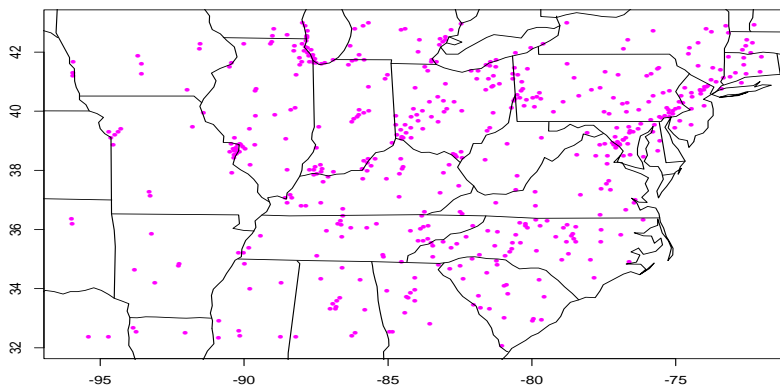


FIGURE 3. Stations where ozone measurements were collected.

30 consecutive days in 1999. The specific measurements are calculated by taking the maximum eight hour average ozone concentration for a particular day at each location. We have these measurements at 512 locations scattered across the eastern United States. Figure 3 shows the locations of the measurement sites.

3.2 A Simple Space-Time Model Using Process Convolutions

A spatial model constructed using process convolutions can easily be extended to a space-time model by allowing the white noise process, $x(\omega)$, to evolve over time. This can be done by specifying that the $x(\omega_i)$'s follow independent Gaussian random walks over time, i.e.

$$x(\omega_i, t) = x(\omega_i, t - 1) + \nu_{i,t}, \quad (3.9)$$

where $\nu_{i,t} \stackrel{iid}{\sim} N(0, 1/\lambda_\nu)$. This extension results in the following space-time model for the field Ψ at all locations in space $s \in D$ and for all $t \in T$:

$$\psi(s, t|x) = \sum_{i=1}^c k(\omega_i - s)x(\omega_i, t) + \mu + \epsilon_{s,t} \quad (3.10)$$

$$x(\omega_i, t) = x(\omega_i, t-1) + \nu_{i,t}, \quad (3.11)$$

where $\epsilon_{s,t} \stackrel{iid}{\sim} N(0, 1/\lambda_\epsilon)$ and $\nu_{i,t} \stackrel{iid}{\sim} N(0, 1/\lambda_\nu)$. This field can be rewritten as

$$\Psi(t) = \mathbf{K}\mathbf{x}(t) + \mu + \epsilon(t) \quad (3.12)$$

$$\mathbf{x}(t) = \mathbf{x}(t-1) + \nu(t) \quad (3.13)$$

where $\mathbf{x}(t)$ and \mathbf{K} are defined in equation 2.8 and $\epsilon(t)$ and $\nu(t)$ are vectors of the error terms.

Figure 4 depicts the way this model relates the data in space and in time:

- The temporal dependence is captured within the latent process $x(\omega_i, t)$.
- The spatial dependence is obtained by smoothing the latent process $x(\omega_i, t)$ in space using the convolution kernel, k .

The linear structure and normal error distributions of equation 3.12 allow the full conditional distributions of the $\mathbf{x}(t)$'s to be computed in closed form given the other values of the model parameters $\lambda_\epsilon, \lambda_\nu$, and μ . This can be done using the Forward Filtering Backward Sampling Algorithm (FFBS), defined in Theorem 16.1 in West and Harrison (1997). Also, given the values of the latent $\mathbf{x}(t)$'s, the full conditional distributions of λ_ϵ , λ_ν , and μ can be found in closed form. As a result of this model structure, full posterior inference can be performed using a simple Gibbs sampler.

3.3 The Model for Ozone Concentration

Our model for ozone concentration is an extension of the simple model proposed above. We define \mathbf{Y}_t to be the vector of the *log* of the ozone readings at time t . For $t = 1, 2, \dots, T$,

$$\mathbf{Y}_t = \mathbf{K}\mathbf{x}_t + \mu + \epsilon_t \quad \epsilon_t \sim N(0, (1/\lambda_\epsilon)\mathbf{I}) \quad (3.14)$$

$$\mathbf{x}_t = \mathbf{G}(\beta)\mathbf{x}_{t-1} + \nu_t \quad \nu_t \sim N(0, (1/\lambda_\nu)\mathbf{I}). \quad (3.15)$$

Prior distributions for other parameters of the model are taken to be independent and conjugate for simplicity; the priors for μ and \mathbf{x}_0 are normal, and the priors for the precision parameters λ_ϵ and λ_ν are diffuse gamma distributions.

The function $G(\beta)$ is constructed so that

$$\begin{aligned} x_{(i,j),t} = & \beta_{(i,j)}^0 x_{(i,j),t-1} + \beta_{(i,j)}^N x_{(i,j+1),t-1} + \beta_{(i,j)}^E x_{(i+1,j),t-1} \\ & + \beta_{(i,j)}^S x_{(i-1,j),t-1} + \beta_{(i,j)}^W x_{(i-1,j),t-1} + \nu_{(i,j),t} \end{aligned} \quad (3.16)$$

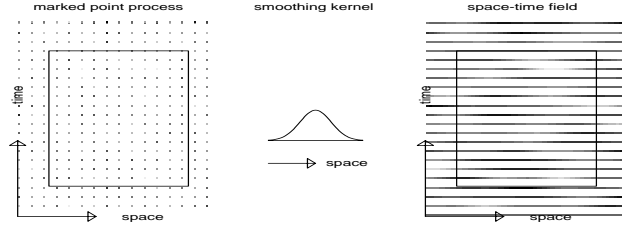


FIGURE 4. Smoothing a latent process over space as the process evolves over time.

- $\beta_{(i,j)}^0$ represents the amount of influence that the ozone level at location (i, j) at time $t - 1$ has on the level at (i, j) at time t .
- $\beta_{(i,j)}^k$ represents the amount of influence that the ozone level from the k of (i, j) , for $k \in ((N)orth, (E)ast, (S)outh, (W)est)$, at time $t - 1$ has on the level at time t .

The β^k for $k \in (0, N, E, S, W)$ are constant over time but not across the spatial field. Consequently, they represent the amount of influence ozone levels from neighboring locations have on the current level on average over the 30 days we are considering. We should also note that the locations on the edges of our spatial field do not have a complete set of β 's; we only fit β^k 's that correspond to locations within our spatial field.

The priors on the β^k 's are Gaussian and independent for all combinations of values of $k \in (0, N, E, S, W)$ but are dependent for a specific value of k . The covariance matrices of these prior distributions incorporates into the model dependence between each of the $\beta_{(i,j)}^k$'s, for a fixed k , and its neighbors. Specifically, the priors take the form of equations 2.1 and 2.2 with a fixed $\theta = .05$. This choice of prior was somewhat arbitrary. Larger values of θ tended to remove all prior influence making directions of influence over large regions hard to discern. Smaller values were avoided to prevent the prior from overwhelming the data.

The posterior distributions of the β 's can be found within the framework of the Gibbs sampler discussed above. Given the values of all of the other model parameters, including the \mathbf{x} 's, the full conditional distributions of the β 's will be multivariate normal and so can be sampled directly.

Ozone readings at a couple of the stations are missing from the data set at a few time points. This is not a problem in the Bayesian setting; we treat the missing values as parameters in the model and find their posterior distributions within our Gibbs sampler.

3.4 Results

After running an MCMC for 3000 iterations, we calculated posterior means for the 296 β 's estimated. The means for the directional β 's (β^N , β^E , β^S and β^W) are summarized in figure 5. The influence from the same location at previous time steps (β^0) is not included. At each location ω_i in the lattice, a number of arrows are plotted corresponding to the number of directional β 's available for that point in the lattice. The ω_i 's in the center have information from all directions, but the ω_i 's along the edges of the lattice have only 2 or 3 neighbors. The arrows are plotted so that their direction and color indicate the sign of their posterior means; the influence on the latent process at $\omega_{(i,j)}$ from the process at $\omega_{(i-1,j)}$ (its neighbor to the west) would be represented by a black arrow to the east if the value is positive or a grey arrow back toward the west if the value is negative. The sizes of the arrows are scaled such that one degree of latitude or longitude corresponds to .2 on the β scale. Thus, an arrow of length 2° in figure 5 corresponds to $\beta = .4$.

Examining figure 5, some patterns are immediately obvious. In the Northeast, the latent process shows movement to the north and east. Further to the west in the Midwest, the trend is more to the east and south. Through the plains and the South, influence from surrounding areas does not appear to be as strong.

In this analysis, we only considered the influence on the latent process at some location ω_i by the latent process at lattice locations directly surrounding ω_i . It is possible to construct $G(\beta)$ to allow for inference from locations further away. We also considered models using second and third order neighborhoods, but the results did not yield additional information.

Figure 6 shows the interpolated ozone concentration for eight consecutive days. We constructed a fine grid over the region and calculated a new transformation matrix K^* to link the white noise processes and the fine grid points as in equation 2.8. Using the posterior means of the white noise processes, $\hat{x}(\omega_i)$, the predicted values on the fine grid are $K^*\hat{x}(\omega_i)$.

4 Conclusion

Our exploratory analysis has given some insight into the space-time dependence in ozone concentrations in different parts over the eastern United States. For instance, in the Northeast, there appears to be some north-eastern movement of the latent process over time. This observation might indicate that weather variables like wind or temperature would be helpful in predicting true ozone concentrations in those areas. In the South, the influence from other regions at previous time steps is not as apparent. In such situations, we might look at more location specific influences like auto emissions or industrial activity to predict ozone levels.

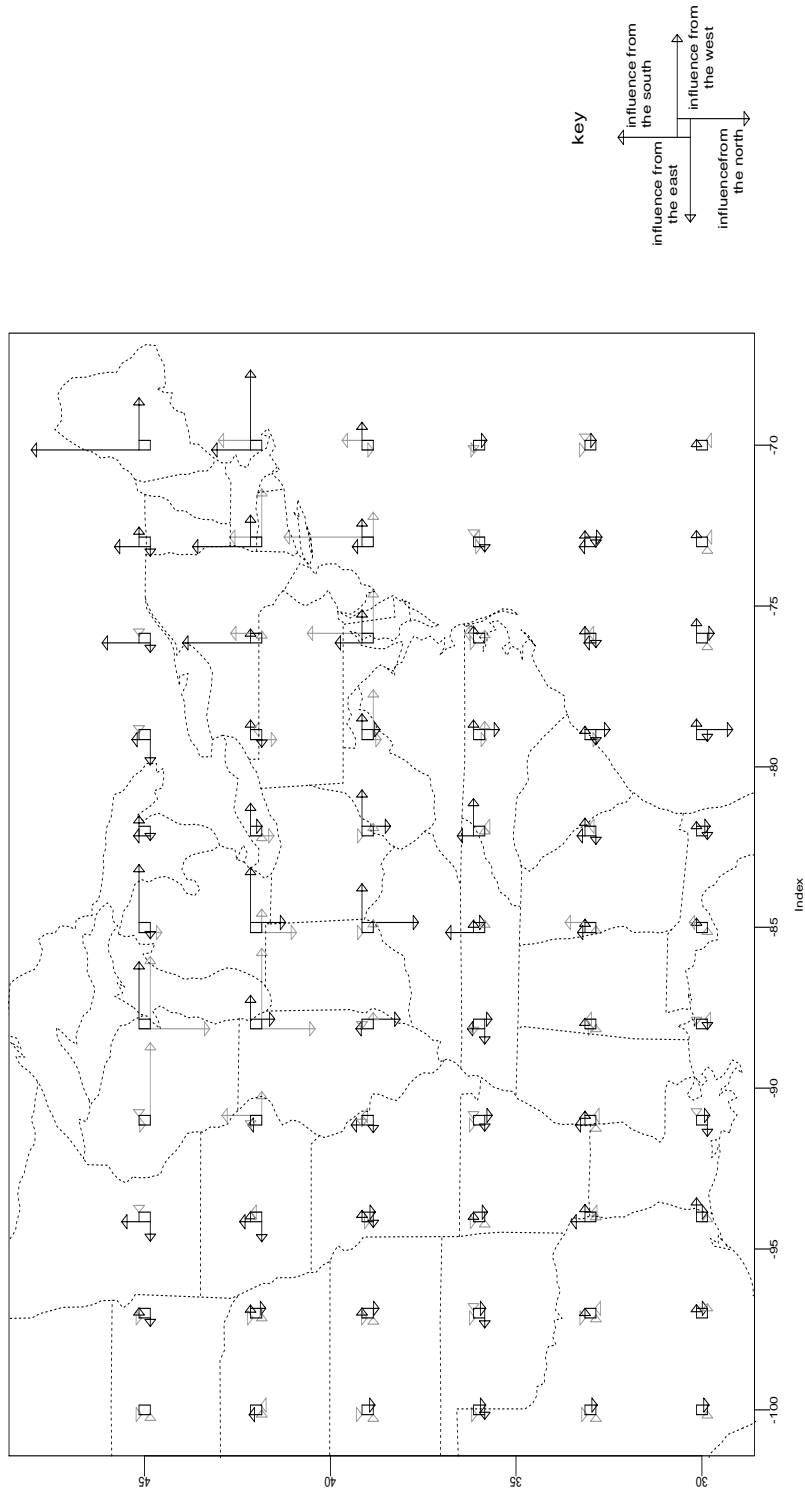


FIGURE 5. Posterior means of the β 's in arrow form. A black arrow represents a positive posterior mean of a β , and a grey arrow represents a negative posterior mean of a β . These arrows indicate strength and direction of influence of neighboring ozone concentration levels from the previous day.

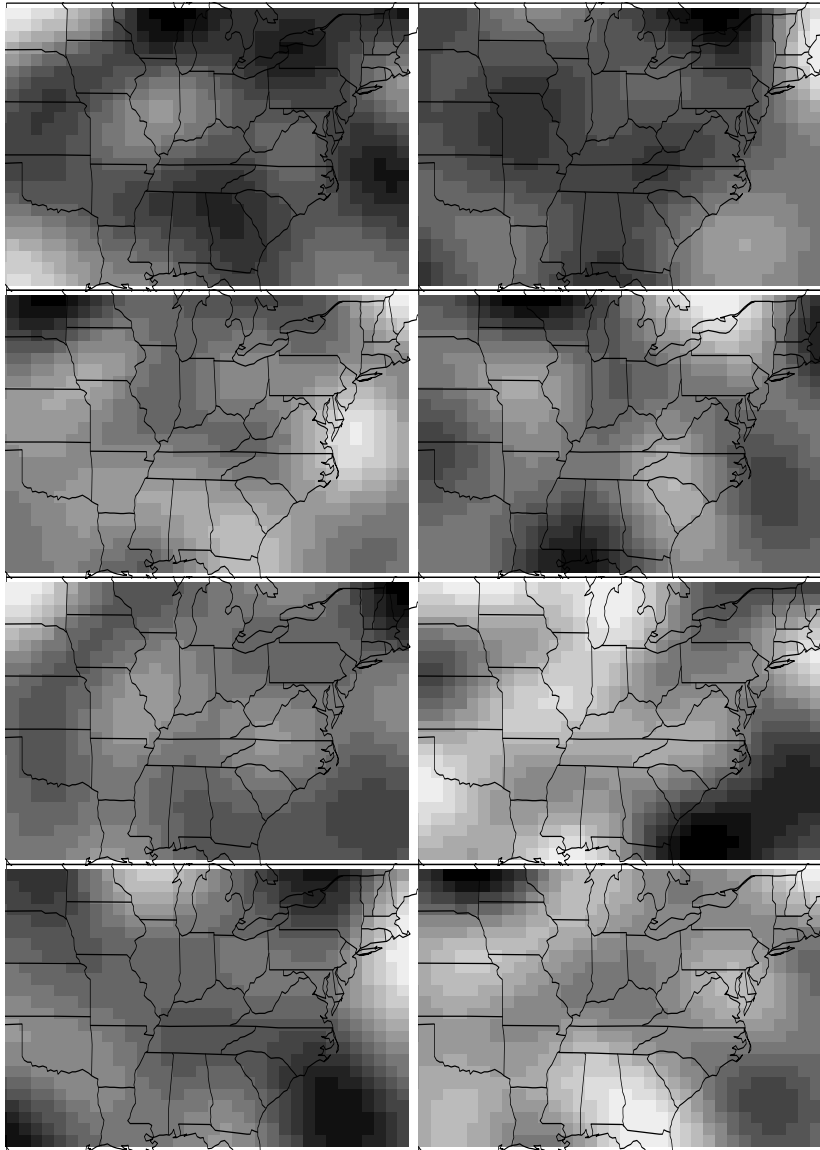


FIGURE 6. Interpolated ozone concentration levels for eight consecutive days based on the posterior means of the white noise processes, $x(\omega_i)$'s. Darker shades represent lower ozone concentrations.

These results also may give insight into a more appropriate allocation of ozone monitoring sites in the eastern United States. For instance, fewer stations may be adequate in the Northeast. Ozone levels in areas in this region can be fairly accurately predicted based on the concentration levels

to the southwest in the previous few days. Alternatively, in the South where there appears to be little influence from the surrounding areas, it may be necessary to add more stations to monitor ozone levels.

Two possible extensions of this model involve incorporating nonstationarity and anisotropy. With regard to stationarity, we have taken the mean level of ozone to be constant across the 30 day period. For such a short time span, such an assumption may not impair the ability of our model to pick out patterns of ozone movement. However, over longer time periods, we would want to model the changes in mean ozone level over time. In the same way, our assumption of isotropy could be changed by allowing asymmetric kernels. Swall (1999) addresses the topic of process convolutions with anisotropic and spatially evolving kernels. If the form of the kernels is taken to be known, the computational cost is the same as in our model. However, fitting parameters that specify the density anisotropic kernels can be computationally expensive, especially over such a large area. Our model trades some accuracy for computational feasibility in this regard.

In the model we have considered here, the spacing of the ω 's in the lattice and the neighborhood structure are somewhat confounded. For instance, we could choose a lattice that is twice as dense (a 21×11 lattice) and define $G(\beta)$ to consider only influence from the latent process two grid steps away instead of one. The modeling technique introduced here is not intended to determine the optimal model for the situation but to be an exploratory technique that might give insight into the underlying processes driving actual ozone concentrations. Our experimentation with other grid sizes has produced results similar to those shown, so combining process convolutions and linear modeling appears to be an effective technique for exploring the underlying processes driving ozone concentration levels.

Acknowledgments

Thanks to Dave Holland, an EPA statistician who provided valuable insight as well as the data for this application. This research was funded in part by a cooperative agreement with the EPA.

References

- Cressie, N.A.C. (1991). *Statistics for Spatial Data*. New York: Wiley-Interscience.
- Higdon, D.M. (1998). A process-convolution approach to modeling temperatures in the north Atlantic Ocean. *Journal of Environmental and Ecological Statistics*, **5**, 173-190.
- Higdon, D.M., Swall, J. and Kern, J. (1999). Non-stationary spatial modeling. In *Bayesian Statistics 6. Proceedings of the Sixth Valencia In-*

ternational Meeting, 761-768. Oxford University Press.

- Higdon, D.M. (2001). Space and space-time modeling using process convolutions. Discussion Paper 01-03, Institute of Statistics and Decision Sciences, Duke University.
- Stroud, J., Müller, P. and Stanso, B. (1999). Dynamic Models for Spatio-Temporal Data. Technical Report 99-20, Institute of Statistics and Decision Sciences, Duke University.
- Swall, J. (1999) A process convolution approach to modeling non-stationary spatial dependence. Ph.D. Thesis. Duke University, Durham, NC 27708.
- Wackernagel, H. (1995). *Multivariate Geostatistics. An Introduction with Applications*. New York: Springer-Verlag.
- West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models (Second Edition)*. New York: Springer-Verlag.