Local Real-time Forecasting of Ozone Exposure using Temperature Data

May 25, 2017

Abstract

Rigorous and prompt assessment of ambient ozone exposure is important for informing the public about ozone levels that may lead to adverse health effects. In this paper, we make use of hierarchical modeling to forecast 8-hour average ozone exposure. Our contribution is to show how incorporating temperature data in addition to observed ozone can significantly improve forecast accuracy, as measured by predictive mean squared error and empirical coverage. Furthermore, our model meets the objective of forecasting in real-time. These advantages are illustrated through modeling data collected at the Village Green monitoring station in Durham, North Carolina.

1 Introduction

Air pollution continues to be a common health concern. There is now growing interest in providing access to real-time monitoring data and using these data to provide air quality forecasts for local communities. Real-time forecasts and personalized smartphone notifications could allow peope to modify their behavior to help reduce pollutant exposures. In the past, continuous, long-term measurement of ambient air pollution has been limited by monitoring restrictions and resource constraints. Recently, there have been monitoring advances in the development of stationary air measurement systems that can provide long-term data collection and be deployed in community environments.

To improve the public understanding of location air pollution, the U.S. Environmental Protection Agency (EPA) developed the Village Green Project in 2015 to demonstrate the capabilities of new real-time monitoring technology for residents and atmospheric scientists to learn more about air quality. The primary monitoring goal is to provide the public and communities with previously unavailable pollution information about local air quality as well as to promote community awareness of air pollution. A pilot station in Durham, North Carolina has demonstrated the ability of the system to monitor several air pollutants in real-time and has the ability to provide access to the data online (https://www.airnow.gov/index.cfm?action =airnow.villagegreen) or by smartphone. This solar and wind-powered station is a park bench structure with instruments that provide minute by minute measurements of ozone, particulate levels, and meteorology. The Village Green project has now expanded to monitoring sites in other communities across the U.S. In particular, other Village Green sites include Washington, D.C., Chicago, and Philadelphia but here, we focus on the Durham site for our model development.

Ground-level ozone, continues to be a health problem in many urban areas of the US. It is well known that temperature influences ozone formation. As a photochemical air pollutant, high temperature wih sunlight in the presence of ozone precursors encourage ozone formation (US EPA, 2013). Both temperature and ground-level ozone are expected in increase in response to expected global climate change (Jacob, and Winner, 2009; US EPA, 2009). This encourages the development of ozone forecasting using temperature information, and consequently motivates the effort and resulting contribution of this paper. We seek to provide forecasts of ozone levels based on temperature observations. Here, we focus on forecasting 8-hour average ozone concentrations defined (at hour t) as the average of the previous four hourly concentrations $(t-4, t-3, t-2, t-1)$, the current hour t, and the next 3 hours $(t+1, t+2, t+3)$. This statistic is chosen in accord with its interest in the environmental and epidemiological communities. The U.S. EPA national ambient air quality standard is based on 4th maxima of 8-hour ozone averages (put web site) and 8-hour averages are used in the computation of EPAs air quality index (AQI) (see http:www.airnow.gov). Using operational weather models, the National Oceanic and Atmospheric Administration (NOAA) Agency displays spatial patterns of average 8-hour ozone across the US (https://www.airnow.gov/index.cfm?action=airnow.noaamaps). Forecasting is accomplished through the development of appropriate hierarchical models, specifically bivariate autoregressive models with conditional and marginal specifications. We compare a variety of different candidate choices by evaluating their predictive performance. Models including temperature as a covariate are shown to outperform autoregressive models with just ozone. All of these models can be implemented in real-time, thus providing almost immediate forecasts of 8-hr average ozone concentrations at a Village Green (VG) site.

We remark that, by now, there is a rich literature on modeling ozone at various temporal scales and also over spatial regions of various sizes. Kalman-filter approaches were considered to improve next day forecasts of ozone concentrations at individual monitoring sites for the summer of 2005 (Kang et al., 2008). An alternative approach for forecasting based on a Bayesian

spatio-temporal model was applied to hourly ozone concentrations (Sahu et al., 2009). However, their model is computationally intensive and not feasible for real-time forecasting. Recent studies explored the idea of fusing observed data with computer model output. An application of the Bayesian melding approach was presented where observed data are combined with numerical model output by introducing a 'true' latent point-level process while monitoring data are linked to the latent process via measurement error model (Fuentes and Raftery, 2005). Univariate and bivariate downscaler models (Berrocal et al., 2010a, b) were developed to relate the monitoring data and numerical model output using a regression model with spatially varying coefficients (Gelfand et al., 2003) modeled using Gaussian processes. A regression model for real-time forecasting was developed using first differences along with spatio-temporally varying coefficients in Paci et al., 2013. In addition, there is previous work explaining ozone levels through temperature (Camalier et al., 2007; Bloomer et al., 2009; Wilson et al., 2014).

However, the specific contribution here is to explore innovative fusions between observed ozone and temperature, which can be used for real-time forecasting. Such forecasting is usually done at hourly scales as described above. However, our models run so quickly that one can envision real-time forecasting at much finer temporal scales, e.g., minute-by-minute updating should that prove to be of interest. Additionally, our temperature driven forecasting can substantially reduce calibration error propagated from computer models such as η-CMAQ (Paci, 2013).

Here, we focus primarily on model comparison, exploring a collection of multi-level models incorporating various autoregressive structures as well as periodicity, and heterogeneous variation. Clear benefit emerges by introducing suitable autoregression and periodicity while we find no benefit with heterogeneous variances. Criteria we use to make these assessments include predictive mean square error (PMSE), continuous rank probability scores (CRPS), and predictive interval length. We illustrate the modeling process and model comparison using data collected at the Village Green Durham Station during May and June, 2015. Here, we focus on single site real-time forecasting. However, a spatial version is in development, incorporating the knowledge we have gleaned from the effort here and will be presented in a future manuscript. The spatial version will work with the original time series instead of differenced time series (Paci et al., 2013); this can be shown to reduce forecast uncertainty.

The paper is organized as follows. Section 2 describes data in detail. Section 3 presents the modeling ideas. Section 4 demonstrates modeling results from simulated and real data. Finally, a summary of the results and discussion of future work are presented in Section 5.

2 Data

We use two main sources of data: the first source consists of ground-level ozone concentration in parts per billion units (ppb), and the second source consists of temperature in degrees Celcius. Both sources were collected at the Village Green Durham station at one-minute frequency from May 29^{th} 07:00:00 to June 23^{rd} 07:00:00, 2015, and were aggregated to hourly averages before further modeling. This time period of analysis was chosen so that we have no missingness in data.

Past work has shown success using square root transformation when modeling ozone, and log transformation when modeling temperature. Displays of the raw data on these scales along with the associated autocorrelation and partial autocorrelation functions are shown in Figures 1 and 2. From the running plot of square-root ozone, we can see that the time series is approximately stationary without any apparent trend. Its partial autocorrelation function (PACF) shows

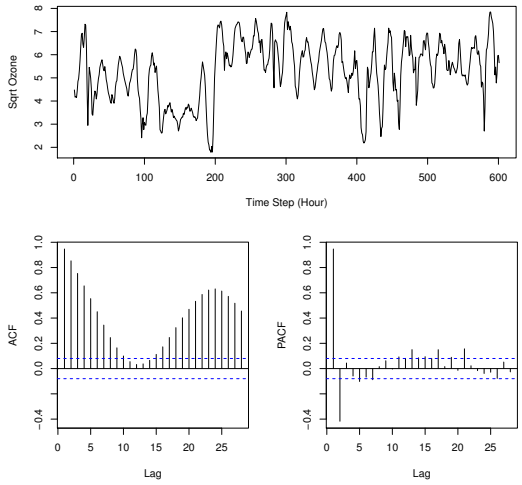


Figure 1: Time series display of square-root ozone with autocorrelation behavior

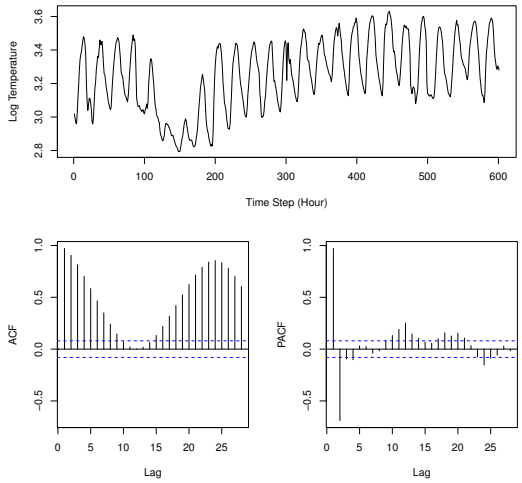
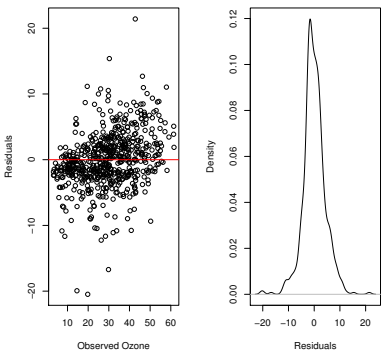


Figure 2: Time series display of log temperature with autocorrelation behavior

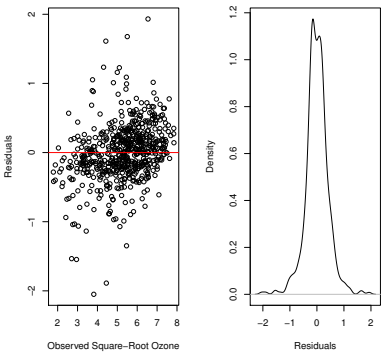
strong correlations in the first two lags and quickly diminishes, providing evidence for including up to two autoregressive terms in specifying models. Its autocorrelation function (ACF) displays an oscillating pattern, with a peak at the 24^{th} lag. This signals the existence of day-long periodicity, providing evidence for including periodic means (Figure 1). Similar behavior can be noticed for log temperature, although its running plot displays a slightly increasing trend, and its first-order autocorrelation is stronger than that of square-root ozone (Figure 2).

We also argue for the viability of such transformations by examining the residual plots of AR(1) regressions of ozone and temperature, before and after their respective transformation. Although the differences among the transformations are small, residuals of transformed ozone and temperature models do seem more normal and homogeneous than their untransformed counterparts (Figure 3). Furthermore, we see a modestly stronger linear relationship between square-root ozone and log temperature than that between ozone and temperature on their original scales (Figure 4).

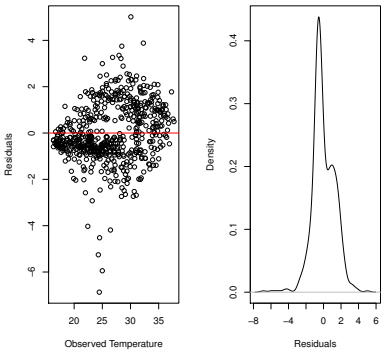
So, in the sequel, we decide to model ozone on the square-root scale, and temperature on the log scale. That is, all Z's in the model specifications that follow are $\sqrt{Z}$, and all X's are



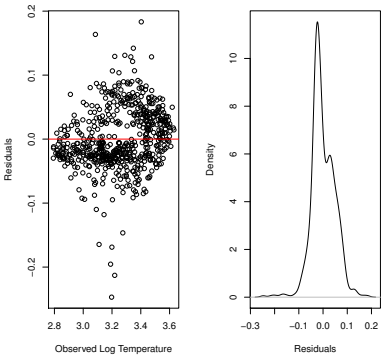
(a)



(b)

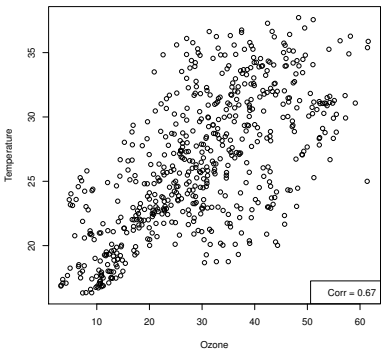


(c)

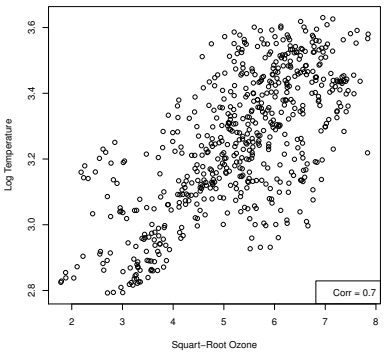


(d)

Figure 3: Scatter plots of residuals by observed values and residual density plots. Plots on the left are from AR(1) models of variables pre-transformation, those on the right are from AR(1) models post-transformation.



(a)



(b)

Figure 4: Correlation plots of ozone and temperature on normal and transformed scales

$\log X$. Nonetheless, all forecasts are transformed back to their original scale for any model evaluation.

3 Modeling Ideas

In this section, we propose four classes of modeling specifications. The classes are *nested*, supplying increasing conditioning. We present their general structures, including the possibility of incorporating a periodic component as well as heterogeneous variances. We also detail the strategy for 8-hour average prediction. All models are hierarchical and we fit them within a Bayesian framework. The 8-hour average ozone prediction is a post model fitting activity using posterior predictive samples. Lastly, we present our model comparison criteria.

The four classes of models are all autoregressive (AR), as is appropriate with the data we employ. The first class of models only uses the ozone data in an AR specification. The second class of models uses an AR structure for temperature with conditionally independent modeling of ozone given temperature. The third class of models specifies AR structure for temperature and AR structure for ozone along with regression on temperature. The fourth class of models specifies AR structure for temperature and AR structure in ozone residuals given temperature. A graphical model of the general structure for each of the classes is given in Figure 5.

The 8-hour average ozone at time t is defined as the average of the previous four hours, the current hour, and the next three hours in the future, that is:

$$Y_t = \frac{1}{8}\sum_{i=-4}^{3} Z_{t+i} \tag{1}$$

Since we are able to observe ozone up to the current hour, our challenge lies in the one-, two- and three-step ahead forecasting of ozone. Throughout this section, we use Y_t to denote the 8-hour average ozone at time t, Z_t to denote the hourly ozone at time t, and X_t to denote the hourly temperature at time t.

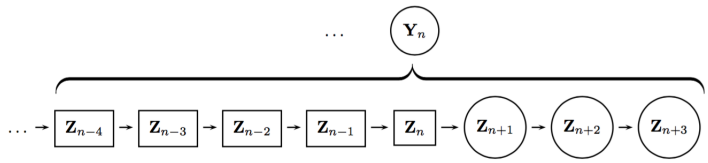
A p^{th} order autoregressive model takes the form:

$$AR(p) : y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + w_t$$

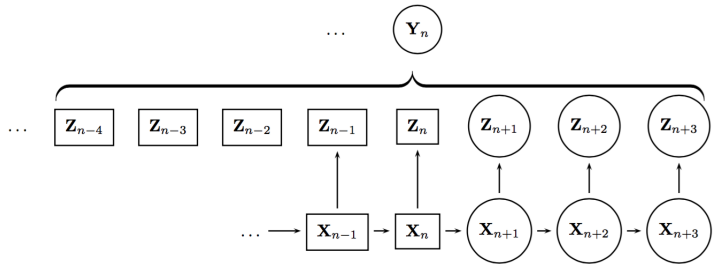
which is often written using the lag operator L, i.e., $Ly_t = y_{t-1}$ and $L^p y_t = y_{t-p}$ as

$$\begin{aligned} y_t &= \delta + \phi_1 L y_t + \phi_2 L^2 y_t + ... + \phi_p L^p y_t + w_t \\ y_t - \phi_1 L y_t - \phi_2 L^2 y_t - ... - \phi_p L^p y_t &= \delta + w_t \\ (1 - \phi_1 L - \phi_2 L^2 - ... - \phi_p L^p) y_t &= \delta + w_t \\ \phi_p(L) y_t &= \delta + w_t \end{aligned}$$

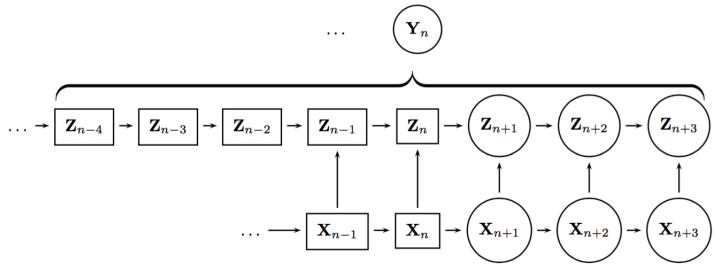
The exploratory section above indicated the possibility of lag 1, lag 2 and possibly lag 24 terms suggesting examination of $AR(p)$ specifications. However, we note here that all of the following models were fitted with these lag terms but, out of sample, only for the class I models and the class II models did $p = 2$ and $q = 2$ outperform $p = 1 and q = 1$, respectively. Therefore, in the interest of space, for all remaining models, we only present results for the first order lag versions.



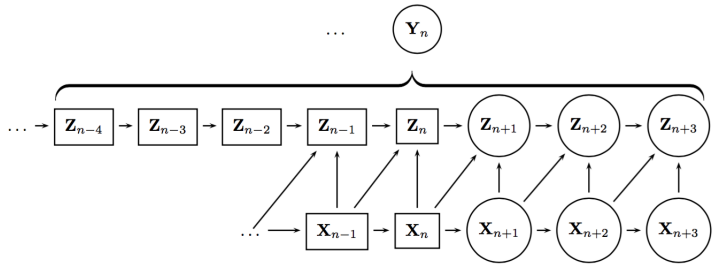
(a) Model Class I



(b) Model Class II



(c) Model Class III



(d) Model Class IV

Figure 5: Graphical illustration of the four model classes, rectangles represent observed variables, circles represent unobserved variables

3.1 Four Model Classes

3.1.1 Model Class I

The first class consists of one-stage autoregressive models. These models use only ozone data, ignoring temperature data.

$$\phi_p(L)Z_t = e_t,\ e_t \sim N(0, \sigma_z^2),\ t = 1, ..., n \tag{2}$$

Hence, the vector of parameters to be estimated is $\boldsymbol{\theta_z} = (\boldsymbol{\phi_z}, \sigma_z^2)$, where $\boldsymbol{\phi_z} = (\phi_{1_z}, \phi_{2_z}, ..., \phi_{p_z})$.

Let $\mathbf{Z_{1:p}}$ denote the vector of observed ozones from time 1 to p, and let $\mathbf{Z^L_{1:p}}$ denote the vector of lagged ozones $(Z_{t-1}, Z_{t-2}, ..., Z_{t-p})$. We model ozone from time $p+1$ conditionally and, therefore, consider $\mathbf{Z_{1:p}}$ as fixed in specifying the model. As noted above, in our real data illustration section (4.2), after considering various choices of $p > 1$, we found no improvement in predictive performance; therefore p is taken as 1 in the following specifications.

By hypothesis, we have that $Z_t|\mathbf{Z^L_{1:p}}, \boldsymbol{\theta_z} \sim N(\sum_{k=1}^p \phi_{k_z} Z_{t-k}, \sigma_z^2)$. Therefore the likelihood function of the model is given by

$$f(\mathbf{Z_{(p+1):n}}|\mathbf{Z_{1:p}}, \boldsymbol{\theta_z}) = \prod_{t=p+1}^{n} \frac{1}{\sqrt{2\pi\sigma_z^2}} exp(-\frac{Z_t - \sum_{k=1}^p \phi_{k_z} Z_{t-k}}{2\sigma_z^2}) \tag{3}$$

Under a Bayesian framework and adding priors for $\boldsymbol{\theta_z}$, the joint posterior distribution of $\boldsymbol{\theta_z}$ is given by

$$\begin{aligned} f(\boldsymbol{\theta_z}|\mathbf{Z_{1:n}}) &\propto f(\mathbf{Z_{(p+1):n}}|\mathbf{Z_{1:p}}, \boldsymbol{\theta_z}) f(\boldsymbol{\theta_z}) \\ &\propto \prod_{t=p+1}^{n} \frac{1}{\sqrt{2\pi\sigma_z^2}} exp(-\frac{Z_t - \sum_{k=1}^p \phi_{k_z} Z_{t-k}}{2\sigma_z^2}) \times \prod_{k=1}^{p} f(\phi_{k_z}) f(\sigma_z^2) \end{aligned} \tag{4}$$

where, conventionally, $f(\phi_{k_z})$ are normal distributions, and $f(\sigma_z^2)$ is an inverse gamma distribution.

When $p = 1$, samples from the one-, two-, three-step ahead posterior predictive distributions are obtained by making draws from each of the following distributions, given a posterior draw of $\boldsymbol{\theta_z}$.

$$\begin{aligned} \hat{Z}_{n+1}|Z_n, \boldsymbol{\theta_z} &\sim N(\phi_{1_z} Z_n, \sigma_z^2) \\ \hat{Z}_{n+2}|\hat{Z}_{n+1}, \boldsymbol{\theta_z} &\sim N(\phi_{1_z} \hat{Z}_{n+1}, \sigma_z^2) \\ \hat{Z}_{n+3}|\hat{Z}_{n+2}, \boldsymbol{\theta_z} &\sim N(\phi_{1_z} \hat{Z}_{n+2}, \sigma_z^2). \end{aligned} \tag{5}$$

3.1.2 Model Class II

The second class consists of two-stage hierarchical models, where the first stage is a linear regression model of square-root ozone on log temperature, and the second stage is a univariate autoregressive model of temperature,

$$\begin{aligned} Z_t &= \alpha_0 + \alpha_1 X_t + e_t,\ e_t \sim N(0, \sigma_z^2) \\ \phi_q(L) X_t &= \eta_t,\ \eta_t \sim N(0, \sigma_x^2) \end{aligned} \tag{6}$$

Hence, the vectors of parameters to be estimated are $\boldsymbol{\theta_z} = (\alpha_0, \alpha_1, \sigma_z^2)$ and $\boldsymbol{\theta_x} = (\boldsymbol{\phi_x}, \sigma_x^2)$, where $\boldsymbol{\phi_x} = (\phi_{1_x}, \phi_{2_x}, ..., \phi_{q_x})$.

Let $\mathbf{X^L_{1:q}}$ denote the vector of lagged temperatures $(X_{t-1}, X_{t-2}, ..., X_{t-q})$. Similarly, we model temperature from time $q+1$, although in the real data illustration $q = 1$ is the adopted specification. By hypothesis, we have $Z_t|X_t, \boldsymbol{\theta_z} \sim N(\alpha_0+\alpha_1 X_t, \sigma_z^2)$ and $X_t|\mathbf{X^L_{1:q}}, \boldsymbol{\theta_x} \sim N(\sum_{k=1}^{q} \phi_{k_x} X_{t-k}, \sigma_x^2)$. Therefore the likelihood function of the model is given by

$$\begin{aligned} f(\mathbf{Z_{1:n}}, \mathbf{X_{(q+1):n}}|\mathbf{X_{1:q}}, \boldsymbol{\theta_z}, \boldsymbol{\theta_x}) &= f(\mathbf{Z_{1:n}}|\mathbf{X_{1:n}}, \boldsymbol{\theta_z}) f(\mathbf{X_{(q+1):n}}|\mathbf{X_{1:q}}, \boldsymbol{\theta_x}) \\ &= \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi\sigma_z^2}} exp(-\frac{Z_t - (\alpha_0 + \alpha_1 X_t)}{2\sigma_z^2}) \\ &\times \prod_{t=q+1}^{n} \frac{1}{\sqrt{2\pi\sigma_x^2}} exp(-\frac{X_t - \sum_{k=1}^{q} \phi_{k_x} X_{t-k}}{2\sigma_x^2}) \end{aligned} \tag{7}$$

Again, under a Bayesian framework and adding priors for $\boldsymbol{\theta_z}$ and $\boldsymbol{\theta_x}$, their joint posterior distribution is given by

$$\begin{aligned} f(\boldsymbol{\theta_z}, \boldsymbol{\theta_x}|\mathbf{Z_{1:n}}, \mathbf{X_{1:n}}) &\propto f(\mathbf{Z_{1:n}}, \mathbf{X_{(q+1):n}}|\mathbf{X_{1:q}}, \boldsymbol{\theta_z}, \boldsymbol{\theta_x}) f(\boldsymbol{\theta_z}) f(\boldsymbol{\theta_x}) \\ &\propto \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi\sigma_z^2}} exp(-\frac{Z_t - (\alpha_0 + \alpha_1 X_t)}{2\sigma_z^2}) \\ &\times \prod_{t=q+1}^{n} \frac{1}{\sqrt{2\pi\sigma_x^2}} exp(-\frac{X_t - \sum_{k=1}^{q} \phi_{k_x} X_{t-k}}{2\sigma_x^2}) \\ &\times f(\alpha_0) f(\alpha_1) f(\sigma_z^2) \times \prod_{k=1}^{q} f(\phi_{k_x}) f(\sigma_x^2) \end{aligned} \tag{8}$$

where $f(\alpha_0), f(\alpha_1)$, and $f(\phi_{k_x})$ are normal distributions, while $f(\sigma_z^2)$ and $f(\sigma_x^2)$ are inverse gamma distributions.

When $q = 1$, sampling the one-, two-, three-step ahead posterior predictive distributions for temperature is handled with posterior draws of $\boldsymbol{\theta_z}$, $\boldsymbol{\theta_x}$, and the following distributions:

$$\begin{aligned} \hat{X}_{n+1}|X_n, \boldsymbol{\theta_x} &\sim N(\phi_{1_x} X_n, \sigma_x^2) \\ \hat{X}_{n+2}|\hat{X}_{n+1}, \boldsymbol{\theta_x} &\sim N(\phi_{1_x} \hat{X}_{n+1}, \sigma_x^2) \\ \hat{X}_{n+3}|\hat{X}_{n+2}, \boldsymbol{\theta_x} &\sim N(\phi_{1_x} \hat{X}_{n+2}, \sigma_x^2) \end{aligned} \tag{9}$$

and

$$\hat{Z}_{n+h}|\hat{X}_{n+h}, \boldsymbol{\theta_z} \sim N(\alpha_0 + \alpha_1 \hat{X}_{n+h}, \sigma_z^2),\ h = 1, 2, 3. \tag{10}$$

3.1.3 Model Class III

The third class consists of two-stage bivariate autoregressive models, where the first stage regresses square-root ozone on its lagged values and log temperature, and the second stage is an autoregressive model of temperature,

$$\begin{aligned} \phi_p(L) Z_t &= \alpha_0 + \alpha_1 X_t + e_t,\ e_t \sim N(0, \sigma_z^2) \\ \phi_q(L) X_t &= \eta_t,\ \eta_t \sim N(0, \sigma_x^2) \end{aligned} \tag{11}$$

Hence, the vectors of parameters to be estimated are $\boldsymbol{\theta_z} = (\boldsymbol{\phi_z}, \alpha_0, \alpha_1, \sigma_z^2)$ where $\boldsymbol{\phi_z} = (\phi_{1_z}, \phi_{2_z}, ..., \phi_{p_z})$, and $\boldsymbol{\theta_x} = (\boldsymbol{\phi_x}, \sigma_x^2)$ where $\boldsymbol{\phi_x} = (\phi_{1_x}, \phi_{2_x}, ..., \phi_{q_x})$.

Let r be the larger of p and q. Since we have $Z_t|\mathbf{Z^L_{1:p}}, X_t, \boldsymbol{\theta_z} \sim N(\alpha_0+\alpha_1 X_t+\sum_{k=1}^{p}\phi_{k_z} Z_{t-k}, \sigma_z^2)$ and $X_t|\mathbf{X^L_{1:q}}, \boldsymbol{\theta_x} \sim N(\sum_{k=1}^{q}\phi_{k_x} X_{t-k}, \sigma_x^2)$, the likelihood function of the model is given by,

$$
\begin{aligned}
&f(\mathbf{Z_{(r+1):n}}, \mathbf{X_{(r+1):n}}|\mathbf{Z_{1:r}}, \mathbf{X_{1:r}}, \boldsymbol{\theta_z}, \boldsymbol{\theta_z}) \\
&= f(\mathbf{Z_{(r+1):n}}|\mathbf{X_{(r+1):n}}, \mathbf{Z_{1:r}}, \boldsymbol{\theta_z}) f(\mathbf{X_{(r+1):n}}|\mathbf{X_{1:r}}, \boldsymbol{\theta_x}) \\
&= \prod_{t=r+1}^{n} \frac{1}{\sqrt{2\pi\sigma_z^2}} exp(-\frac{Z_t - (\alpha_0 + \alpha_1 X_t + \sum_{k=1}^{p}\phi_{k_z} Z_{t-k})}{2\sigma_z^2}) \\
&\times \prod_{t=r+1}^{n} \frac{1}{\sqrt{2\pi\sigma_x^2}} exp(-\frac{X_t - \sum_{k=1}^{q}\phi_{k_x} X_{t-k}}{2\sigma_x^2})
\end{aligned}
$$

Under a Bayesian framework and adding priors for $\boldsymbol{\theta_z}$ and $\boldsymbol{\theta_x}$, their joint posterior distribution is given by

$$
\begin{aligned}
f(\boldsymbol{\theta_z}, \boldsymbol{\theta_x}|\mathbf{Z_{1:n}}, \mathbf{X_{1:n}}) &\propto f(\mathbf{Z_{(r+1):n}}, \mathbf{X_{(r+1):n}}|\mathbf{Z_{1:r}}, \mathbf{X_{1:r}}, \boldsymbol{\theta_z}, \boldsymbol{\theta_x}) f(\boldsymbol{\theta_z}) f(\boldsymbol{\theta_x}) \\
&\propto \prod_{t=r+1}^{n} \frac{1}{\sqrt{2\pi\sigma_z^2}} exp(-\frac{Z_t - (\alpha_0 + \alpha_1 X_t + \sum_{k=1}^{p}\phi_{k_z} Z_{t-k})}{2\sigma_z^2}) \\
&\times \prod_{t=r+1}^{n} \frac{1}{\sqrt{2\pi\sigma_x^2}} exp(-\frac{X_t - \sum_{k=1}^{q}\phi_{k_x} X_{t-k}}{2\sigma_x^2}) \\
&\times \prod_{k=1}^{p} f(\phi_{k_z}) f(\alpha_0) f(\alpha_1) f(\sigma_z^2) \times \prod_{k=1}^{q} f(\phi_{k_x}) f(\sigma_x^2)
\end{aligned} \tag{12}
$$

where $f(\alpha_0), f(\alpha_1)$, $f(\phi_{k_z})$ and $f(\phi_{k_x})$ are normal distributions, while $f(\sigma_z^2)$ and $f(\sigma_x^2)$ are inverse gamma distributions.

For forecasting, when $p = q = 1$, we need to draw samples from the one-, two-, and three-step ahead posterior predictive distributions. With posterior draws of $\boldsymbol{\theta_x}$ the distributions for temperature are the same as in (9). Given draws of $\boldsymbol{\theta_z}$, those for ozone are given by

$$
\begin{aligned}
\hat{Z}_{n+1}|Z_n, \hat{X}_{n+1}, \boldsymbol{\theta_z} &\sim N(\alpha_0 + \alpha_1 \hat{X}_{n+1} + \phi_{1_z} Z_n, \sigma_z^2) \\
\hat{Z}_{n+2}|\hat{Z}_{n+1}, \hat{X}_{n+2}, \boldsymbol{\theta_z} &\sim N(\alpha_0 + \alpha_1 \hat{X}_{n+2} + \phi_{1_z} \hat{Z}_{n+1}, \sigma_z^2) \\
\hat{Z}_{n+3}|\hat{Z}_{n+2}, \hat{X}_{n+3}, \boldsymbol{\theta_z} &\sim N(\alpha_0 + \alpha_1 \hat{X}_{n+3} + \phi_{1_z} \hat{Z}_{n+2}, \sigma_z^2).
\end{aligned} \tag{13}
$$

3.1.4 Model Class IV

The fourth class consists of two-stage bivariate autoregressive models, where the first stage models ozone as conditionally autoregressive given temperature, and the second stage models temperature as marginally autoregressive.

$$
\begin{aligned}
\phi_p(L)(Z_t - (\alpha_0 + \alpha_1 X_t)) = e_t,\ e_t \sim N(0, \sigma_z^2) \\
\phi_q(L) X_t = \eta_t,\ \eta_t \sim N(0, \sigma_x^2)
\end{aligned} \tag{14}
$$

Hence, the vectors of parameters to be estimated are $\boldsymbol{\theta_z} = (\boldsymbol{\phi_z}, \alpha_0, \alpha_1, \sigma_z^2)$ where $\boldsymbol{\phi_z} = (\phi_{1_z}, \phi_{2_z}, ..., \phi_{p_z})$, and $\boldsymbol{\theta_x} = (\boldsymbol{\phi_x}, \sigma_x^2)$ where $\boldsymbol{\phi_x} = (\phi_{1_x}, \phi_{2_x}, ..., \phi_{q_x})$.

Again, r is the larger of p and q. Since we have $Z_t|\mathbf{Z^L_{1:p}}, X_t, \mathbf{X^L_{1:p}}, \boldsymbol{\theta_z} \sim N(\alpha_0 + \alpha_1 X_t + \sum_{k=1}^{p}\phi_{k_z}(Z_{t-k} - (\alpha_0 + \alpha_1 X_{t-k})), \sigma_z^2)$ and $X_t|\mathbf{X^L_{1:q}}, \boldsymbol{\theta_x} \sim N(\sum_{k=1}^{q}\phi_{k_x} X_{t-k}, \sigma_x^2)$, the likelihood

function of the model is given by,

$$
\begin{aligned}
&f(\mathbf{Z_{(r+1):n}}, \mathbf{X_{(r+1):n}} | \mathbf{Z_{1:r}}, \mathbf{X_{1:r}}, \boldsymbol{\theta_z}, \boldsymbol{\theta_z}) \\
&= f(\mathbf{Z_{(r+1):n}} | \mathbf{X_{1:n}}, \mathbf{Z_{1:r}}, \boldsymbol{\theta_z}) f(\mathbf{X_{(r+1):n}} | \mathbf{X_{1:r}}, \boldsymbol{\theta_x}) \\
&= \prod_{t=r+1}^{n} \frac{1}{\sqrt{2\pi\sigma_z^2}} exp(-\frac{Z_t - (\alpha_0 + \alpha_1 X_t + \sum_{k=1}^{p} \phi_{k_z}(Z_{t-k} - (\alpha_0 + \alpha_1 X_{t-k})))}{2\sigma_z^2}) \\
&\times \prod_{t=r+1}^{n} \frac{1}{\sqrt{2\pi\sigma_x^2}} exp(-\frac{X_t - \sum_{k=1}^{q} \phi_{k_x} X_{t-k}}{2\sigma_x^2})
\end{aligned}
$$

Once again under a Bayesian framework and adding priors for $\boldsymbol{\theta_z}$ and $\boldsymbol{\theta_x}$, their joint posterior distribution is given by

$$
\begin{aligned}
f(\boldsymbol{\theta_z}, \boldsymbol{\theta_x} | \mathbf{Z_{1:n}}, \mathbf{X_{1:n}}) &\propto f(\mathbf{Z_{(r+1):n}}, \mathbf{X_{(r+1):n}} | \mathbf{Z_{1:r}}, \mathbf{X_{1:r}}, \boldsymbol{\theta_z}, \boldsymbol{\theta_x}) f(\boldsymbol{\theta_z}) f(\boldsymbol{\theta_x}) \\
&\propto \prod_{t=r+1}^{n} \frac{1}{\sqrt{2\pi\sigma_z^2}} exp(-\frac{Z_t - (\alpha_0 + \alpha_1 X_t + \sum_{k=1}^{p} \phi_{k_z}(Z_{t-k} - (\alpha_0 + \alpha_1 X_{t-k})))}{2\sigma_z^2}) \\
&\times \prod_{t=r+1}^{n} \frac{1}{\sqrt{2\pi\sigma_x^2}} exp(-\frac{X_t - \sum_{k=1}^{q} \phi_{k_x} X_{t-k}}{2\sigma_x^2}) \\
&\times \prod_{k=1}^{p} f(\phi_{k_z}) f(\alpha_0) f(\alpha_1) f(\sigma_z^2) \times \prod_{k=1}^{q} f(\phi_{k_x}) f(\sigma_x^2)
\end{aligned} \tag{15}
$$

where $f(\alpha_0), f(\alpha_1)$, $f(\phi_{k_z})$ and $f(\phi_{k_x})$ are normal distributions, while $f(\sigma_z^2)$ and $f(\sigma_x^2)$ are inverse gamma distributions.

For forecasting, when $p = q = 1$, we again need to draw samples from the one-, two-, and three-step ahead posterior predictive distributions. With posterior draws of $\boldsymbol{\theta_x}$ the distributions for temperature are the same as in (9). Given draws of $\boldsymbol{\theta_z}$, those for ozone are given by

$$
\begin{aligned}
\hat{Z}_{n+1} | Z_n, \hat{X}_{n+1}, X_n, \boldsymbol{\theta_z} &\sim N(\alpha_0 + \alpha_1 \hat{X}_{n+1} + \phi_{1_z}(Z_n - (\alpha_0 + \alpha_1 X_n)), \sigma_z^2) \\
\hat{Z}_{n+2} | \hat{Z}_{n+1}, \hat{X}_{n+2}, \hat{X}_{n+1}, \boldsymbol{\theta_z} &\sim N(\alpha_0 + \alpha_1 \hat{X}_{n+2} + \phi_{1_z}(\hat{Z}_{n+1} - (\alpha_0 + \alpha_1 \hat{X}_{n+1})), \sigma_z^2) \\
\hat{Z}_{n+3} | \hat{Z}_{n+2}, \hat{X}_{n+3}, \hat{X}_{n+2}, \boldsymbol{\theta_z} &\sim N(\alpha_0 + \alpha_1 \hat{X}_{n+3} + \phi_{1_z}(\hat{Z}_{n+2} - (\alpha_0 + \alpha_1 \hat{X}_{n+2})), \sigma_z^2).
\end{aligned} \tag{16}
$$

3.2 Periodicity

Meteorological data tend to exhibit periodic behavior. In particular, we suspect a diurnal pattern in temperature, i.e., the plot of temperature over time will follow a periodic curve where it rises in the morning, peaks in the early afternoon, and drops at night. Moreover, since ground-level ozone concentration is positively correlated with temperature, we expect to see similar patterns there as well. In order to model the diurnal pattern, we introduce a daily periodic mean along with the AR structure. A daily periodic mean at the t^{th} hour takes the form,

$$
\phi \cos(\frac{2\pi(t \bmod 24)}{24} - \theta), \ t = 1, 2, ..., n \tag{17}
$$

where ϕ is the amplitude, $\frac{1}{24}$ is the frequency (so that the period is 24 hours, or a day), and θ is the phase shift.

We can write (14) as a linear model specification using $\cos(u - v) = \cos(u)\cos(v) - \sin(u)\sin(v)$ so that

$$
\phi \cos(\frac{\pi(t \bmod 24)}{12} - \theta) = a \cos(\frac{\pi(t \bmod 24)}{12}) - b \sin(\frac{\pi(t \bmod 24)}{12})
$$

where $a = \phi\cos(\theta), b = \phi\sin(\theta)$. Evidently, $\phi = \sqrt{a^2+b^2}, \theta = \tan^{-1}(\frac{b}{a})$. However, retrieving ϕ and θ is not necessary for forecasting. More complex periodic functions can be examined, e.g., using Fourier series approximations (Wei, 1994) with $\cos(\frac{k\pi(t \bmod 24)}{12})$ and $\sin(\frac{k\pi(t \bmod 24)}{12}$, for $k = 1, 2, ...$ However, here a simple single daily period is sufficient.

An example of an ozone model with periodic means is,

$$\phi_{p_z}(L)(Z_t - (a\cos(\frac{\pi(t \bmod 24)}{12}) + b\sin(\frac{\pi(t \bmod 24)}{12})) = e_t$$

We model periodicity for ozone and temperature respectively. In addition, we model periodicity in both stages, in case ozone is not conditionally independent of periodicity given temperature.

3.3 Heterogeneous Variance

An assumption in all of the above modeling is constant variance over time. We seek to learn whether a heterogeneous variance specification is more appropriate. Although elaborate stochastic volatility models have been developed in past work (Achar, et al), for considerations of model complexity and runtime, we only consider simple implementations of heterogeneous variance in this paper.

In our ozone models, instead of specifying $e_t \sim N(0, \sigma_z^2)$, we experimented with the following parametrizations that connect model uncertainty to the magnitude of temperature,

$$e_t \sim N(0, \sigma_z^2 X_t)$$
$$e_t \sim N(0, \sigma_z^2 X_t^2)$$

Temperatures are introduced on the log scale. During our window in the ozone season, log temperatures are always above 0. These models are compared to models with homogeneous variance.

3.4 Comparison Criteria

The comparison criteria used in this paper are predictive mean squared error (PMSE), continuous ranked probability score (CRPS), and average length of 90% predictive intervals. Furthermore, in the case of temperature models, the overall PMSE is broken down into one-, two-, and three-step ahead PMSEs to allow for more detailed examination. All criteria are obtained from the samples of the posterior predictive distributions associated with a given $Y_{t,obs}$.

We randomly chose 5 starting hours, and ran 48 hours' worth of forecasts from each of the 5 choices. Criteria arise by averaging over a total of 240 forecasts,

$$\text{PMSE} = \frac{\sum_{t=1}^{240}(\hat{Y}_t - Y_{t,obs})^2}{240}$$
$$\text{CRPS} = \frac{\sum_{t=1}^{240}\int(F(Y_t|\text{Data}) - \mathbf{1}(Y_t \geq Y_{t,obs}))^2 dY_t}{240}$$
$$\text{Average Interval Length} = \frac{\sum_{t=1}^{240}\text{Length of 90\% Predictive Interval at Time } t}{240}$$

Here, $\hat{Y}_t$ is the posterior predictive mean. For CRPS, we use the equivalent expression, $E_F|Y_t - Y_{t,obs}| - \frac{1}{2}E_F|Y_t - Y_t^{'}|$ which enables simple Monte Carlo integration using posterior predictive samples. For all three criteria, small values are preferred.

3.5 Model Fitting

All models are fit using JAGS in R with 40,000 iterations, with a burn-in of 30,000 and a thinning rate of $\frac{1}{10}$ for the last 10,000. For every forecast of 8-hour average ozone, we use a 5-day, 120-hour fitting window. To connect with real time forecasting, the fitting window is shifted one hour forward for the next forecast, and new priors are centered around posterior means learned from the last window.

In the case of one-stage models, samples of one-, two-, and three-step ahead ozone forecasts are drawn directly from their posterior predictive distributions. From the samples, we can obtain 8-hour average ozone forecasts using (5), (10) or (13). In the case of two-stage models, a sample of one-step ahead temperature forecasts is drawn and fed into the first stage ozone model to generate a sample of one-step ahead ozone forecasts. The process works similarly for the needed two- and three-step ahead ozone forecasts.

4 Illustrations

4.1 Simulated Data

In this section, our intent is to demonstrate that, given an underlying ozone sampling model, our criteria will identify the true model as the best performing predictive model among several choices of fitting models.

We specify the true ozone model to be first-order autoregressive with 24-hour periodicity, conditioning on the current-hour temperature,

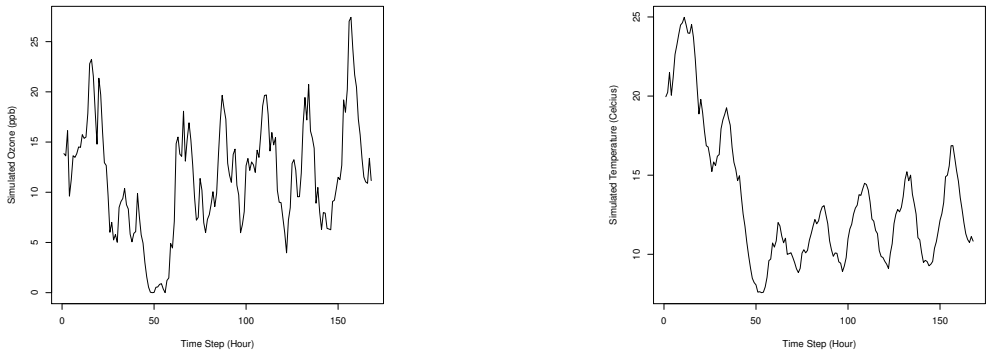
$$\begin{aligned} Z_t &= \alpha_0 + \alpha_1 X_t + a_z \cos(\frac{\pi(t \bmod 24)}{12}) + b_z \sin(\frac{\pi(t \bmod 24)}{12} \\ &+ \phi_{1_z}[Z_{t-1} - (\alpha_0 + \alpha_1 X_{t-1} + a_z \cos(\frac{\pi((t-1) \bmod 24)}{12}) + b_z \sin(\frac{\pi((t-1) \bmod 24)}{12})] \\ &+ e_t,\ e_t \sim N(0, \sigma_z^2) \end{aligned}$$

We specify the true temperature model to be marginally autoregressive with 24-hour periodicity,

$$\begin{aligned} X_t &= a_x \cos(\frac{\pi(t \bmod 24)}{12}) + b_x \sin(\frac{\pi(t \bmod 24)}{12} \\ &+ \phi_{1_x}[X_{t-1} - (a_x \cos(\frac{\pi((t-1) \bmod 24)}{12}) + b_x \sin(\frac{\pi((t-1) \bmod 24)}{12})] \\ &+ \eta_t,\ \eta_t \sim N(0, \sigma_x^2) \end{aligned}$$

In this simulation we also specify that ozone to be generated on square-root scale, and temperature to be generated on log scale. However, during any display of simulated data and evaluation of predictive performance, they will be transformed back to their actual scales. We first generate temperature data for a total of 168 hours, then use this information to generate the same amount of ozone data (Figure 6). The list of true parameters can be found in Table 1.

Here, we only demonstrate the identification of the ozone model, given full information on temperature. The reason is that the temperature model can be identified separately following a similar process. We consider three alternative ozone models, each capturing only part of



(a) (b)

Figure 6: Simulated ozone (a) and temperature (b) data

the underlying structure: (i) Alternative Model 1 does not take temperature into account, (ii) Alternative Model 2 disregards periodicity, and, (iii) Alternative Model 3 has no autoregressive structure at all.

$$\text{Alternative Model 1: } Z_t = a_z \cos(\frac{\pi(t \bmod 24)}{12}) + b_z \sin(\frac{\pi(t \bmod 24)}{12} + \phi_{1_z}[Z_{t-1} - (a_z \cos(\frac{\pi((t-1) \bmod 24)}{12}) + b_z \sin(\frac{\pi((t-1) \bmod 24)}{12})] + e_t,\ e_t \sim N(0, \sigma_z^2)$$

$$\text{Alternative Model 2: } Z_t = \alpha_0 + \alpha_1 X_t + \phi_{1_z}[Z_{t-1} - (\alpha_0 + \alpha_1 X_{t-1})] + e_t,\ e_t \sim N(0, \sigma_z^2)$$

$$\text{Alternative Model 3: } Z_t = \alpha_0 + \alpha_1 X_t + e_t,\ e_t \sim N(0, \sigma_z^2)$$

Model comparison is done employing the criteria mentioned in section 3.4. Since we only have 168 hours of data, and each learning window is 120 hours long, we will only be able to produce 46 hours' worth of 8-hour average ozone forecasts if we start the first learning window at the first hour of simulation. The predictive performance of the models is summarized over these 46 forecasts (Table 2). As we can see, the true model outperforms the alternatives by having the lowest PMSE, CRPS, and the average interval length. These results support the use of our criteria for model selection.

4.2 The Village Green Data

In this section, we illustrate our strategy by modeling and forecasting ozone using the Village Green data from Durham, North Carolina.

We consider the following models for class I. Model 1 is an AR(1) model without intercept. Model 2 is an AR(2) model without intercept. Model 3 is an AR(1) model with periodic means as specified in section 3.3 and no intercept. We also include a bench mark model that sets the next-hour ozone to be the last observed value, and provides no inference. This is to establish

a baseline for our more complicated models to compare to.

Bench Mark: $Z_t = Z_{t-1}$

Model 1: $Z_t = \phi_{1_z} Z_{t-1} + e_t$

Model 2: $Z_t = \phi_{1_z} Z_{t-1} + \phi_{2_z} Z_{t-2} + e_t$

Model 3: $$Z_t - (a_z \cos(\frac{\pi(t \bmod 24)}{12}) + b_z \sin(\frac{\pi(t \bmod 24)}{12})$$
$$= \phi_{1_z}[Z_{t-1} - (a_z \cos(\frac{\pi((t-1) \bmod 24)}{12}) + b_z \sin(\frac{\pi((t-1) \bmod 24)}{12})] + e_t$$

Initial priors are chosen to be weak. For this reason, we also do not restrict distributions of autoregressive parameters to enforce stationarity. ϕ's are modeled to follow $N(0, 100)$. The prior is used for a and b as well. $e_t \sim N(0, \sigma_z^2)$, and σ_z^2 has the initial prior $IG(2, R)$, where R is the sample variance of residuals when AR(1) is fit for ozone using R's *lm* function. As is mentioned in section 3.6, these priors are updated as the fitting window moves forward and forecasting goes on.

Table 3 summarizes the performance of the class I models. We can see that Model 1 actually did worse than the bench mark model. This may be because the first-order autocorrelation of square-root ozone is so close to 1 that model 1 can be essentially concluded as a random walk model. However, modeling ϕ_1 and σ_z^2 may have introduced extra noise without offering more *illumination* compared to the "cheap update" strategy used by the bench mark model. On the other hand, Model 2 and Model 3 performed better than the bench mark, indicating the benefit of additional structures. In particular, Model 3 showed significant improvement in comparison to the other models, suggesting that periodicity is a useful feature to include for ozone.

For the class II models, since the first stage is already specified, we consider the following variations of second stage temperature models. They have structures identical to the above mentioned class I models, and their initial priors are set the same.

Bench Mark: $X_t = X_{t-1}$

Model 4: $X_t = \phi_{1_x} X_{t-1} + \eta_t$

Model 5: $X_t = \phi_{1_x} X_{t-1} + \phi_{2_x} X_{t-2} + \eta_t$

Model 6: $$X_t - (a_x \cos(\frac{\pi(t \bmod 24)}{12}) + b_x \sin(\frac{\pi(t \bmod 24)}{12})$$
$$= \phi_{1_x}[X_{t-1} - (a_x \cos(\frac{\pi((t-1) \bmod 24)}{12}) + b_x \sin(\frac{\pi((t-1) \bmod 24)}{12})] + \eta_t$$

Analyzing the second stage PMSE broken down in steps (Table 4), we find that Model 4 and Model 5 are out of consideration as their overall PMSEs almost double that of the bench mark model. The only acceptable one is Model 6. Its one-step ahead PMSE exceeded that of the bench mark, but this was compensated for by better predictive performances in two- and three-step forecasts. Again, the reason for the failures of Model 4 and Model 5 and the success of Model 6 may be the extremely strong first-order autocorrelation in log temperature, and the need for additional structure such as periodicity.

Recall the first stage model: $Z_t = \alpha_0 + \alpha_1 X_t + e_t$. Setting α's initial priors to be $N(0, 100)$, and $e_t \sim N(0, \sigma_z^2)$, where $\sigma_z^2 \sim IG(2, R)$, we integrate it with the second stage models to obtain predictive performance of the full models (Table 5).

The overall performance of the class II models is poor compared to that of the class I models. Furthermore, the best performing second stage model ended up producing the worst performing full model. This likely results from first stage misspecification – autoregressors are

better predictors of ozone than temperature, if only one of the two can be included. We do not eliminate the plausibility of Model 6 as a second stage model when both autoregressors and temperature are used as covariates in the first stage regression (that is, in class III).

We consider the following models for class III. Based on what we have learned from previous classes, Model 7 has the first order autoregressive term and periodicity terms in addition to current-hour temperature in its first stage. Its second stage is essentially the best performing temperature model (Model 6) from class II. The dual-periodic structure assumes that ozone is conditionally dependent on periodicity given temperature. However, ozone and temperature are highly correlated, and the periodicity in ozone may be induced by that in temperature. In other words, ozone may be conditionally independent on periodicity given temperature. To account for this possibility, we attempt model 8, which has no periodic structure in its first stage. Parametrization of the initial priors agrees with those from the previous two classes. Although the class III models performed better than those in class II, neither of them was able to beat Model 3 in class I (Table 6). Therefore, the first stage specification, particularly with the use of temperature, may still not be appropriately identified.

$$
\begin{aligned}
\text{Model 7: } Z_t &= \alpha_0 + \alpha_1 X_t + a_z \cos(\frac{\pi(t \bmod 24)}{12}) + b_z \sin(\frac{\pi(t \bmod 24)}{12} + \phi_{1_z} Z_{t-1} + e_t \\
X_t &= a_x \cos(\frac{\pi(t \bmod 24)}{12}) + b_x \sin(\frac{\pi(t \bmod 24)}{12} \\
&\quad + \phi_{1_x}[X_{t-1} - (a_x \cos(\frac{\pi((t-1) \bmod 24)}{12}) + b_x \sin(\frac{\pi((t-1) \bmod 24)}{12})] + \eta_t \\
\text{Model 8: } Z_t &= \alpha_0 + \alpha_1 X_t + \phi_{1_z} Z_{t-1} + e_t \\
X_t &= a_x \cos(\frac{\pi(t \bmod 24)}{12}) + b_x \sin(\frac{\pi(t \bmod 24)}{12} \\
&\quad + \phi_{1_x}[X_{t-1} - (a_x \cos(\frac{\pi((t-1) \bmod 24)}{12}) + b_x \sin(\frac{\pi((t-1) \bmod 24)}{12})] + \eta_t
\end{aligned}
$$

We consider the following models for class IV. Model 9 combines the best performing ozone model (Model 3) in class I and the best temperature model (Model 6) in class II. To parallel with class III, we attempt Model 10 without periodic structure in the first stage. Again, parametrization of the initial priors agrees with those from the previous classes.

$$
\begin{aligned}
\text{Model 9: } Z_t &= \alpha_0 + \alpha_1 X_t + a_z \cos(\frac{\pi(t \bmod 24)}{12}) + b_z \sin(\frac{\pi(t \bmod 24)}{12} \\
&\quad + \beta_1[Z_{t-1} - (\alpha_0 + \alpha_1 X_{t-1} + a_z \cos(\frac{\pi((t-1) \bmod 24)}{12}) + b_z \sin(\frac{\pi((t-1) \bmod 24)}{12})] \\
&\quad + e_t \\
X_t &= a_x \cos(\frac{\pi(t \bmod 24)}{12}) + b_x \sin(\frac{\pi(t \bmod 24)}{12} \\
&\quad + \phi_1[X_{t-1} - (a_x \cos(\frac{\pi((t-1) \bmod 24)}{12}) + b_x \sin(\frac{\pi((t-1) \bmod 24)}{12})] + \eta_t \\
\text{Model 10: } Z_t &= \alpha_0 + \alpha_1 X_t + \beta_1[Z_{t-1} - (\alpha_0 + \alpha_1 X_{t-1})] + \beta_2[Z_{t-2} - (\alpha_0 + \alpha_1 X_{t-2})] + e_t \\
X_t &= a_x \cos(\frac{\pi(t \bmod 24)}{12}) + b_x \sin(\frac{\pi(t \bmod 24)}{12} \\
&\quad + \phi_1[X_{t-1} - (a_x \cos(\frac{\pi((t-1) \bmod 24)}{12}) + b_x \sin(\frac{\pi((t-1) \bmod 24)}{12})] + \eta_t
\end{aligned}
$$

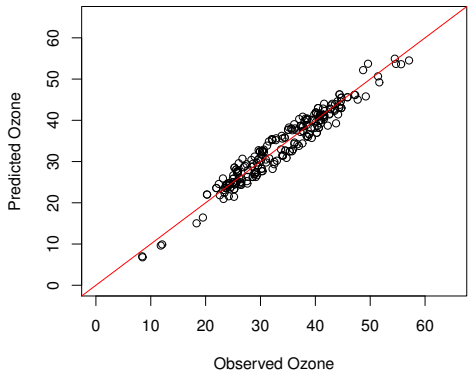


Figure 7: Predicted by observed 8-hour average ozones of Model 9, for 240 windows

A summary of their predictive performance shows that model 9 emerges as the best model in all four classes (Table 7). This is understandable because the inclusion of temperature in addition to periodicity in the first stage model helped explain more variability in ozone (particularly in comparison to Model 3). Moreover, temperature forecasts were provided by the most accurate second stage model among the ones discussed above. A summary of posterior means and 95% credible intervals for parameters in one forecast can be found in Table 8. The scatter plot of predicted ozone vs. observed ozone shows that points cluster closely around the 45^o line with no apparent pattern (Figure 7). The posterior predictive distributions of 8-hour average ozone for a few selected hours show that true values fall inside their 90% credible intervals (Figure 8). These provide further evidence in support of Model 9 which we therefore offer as a recommendation for real-time forecasting.

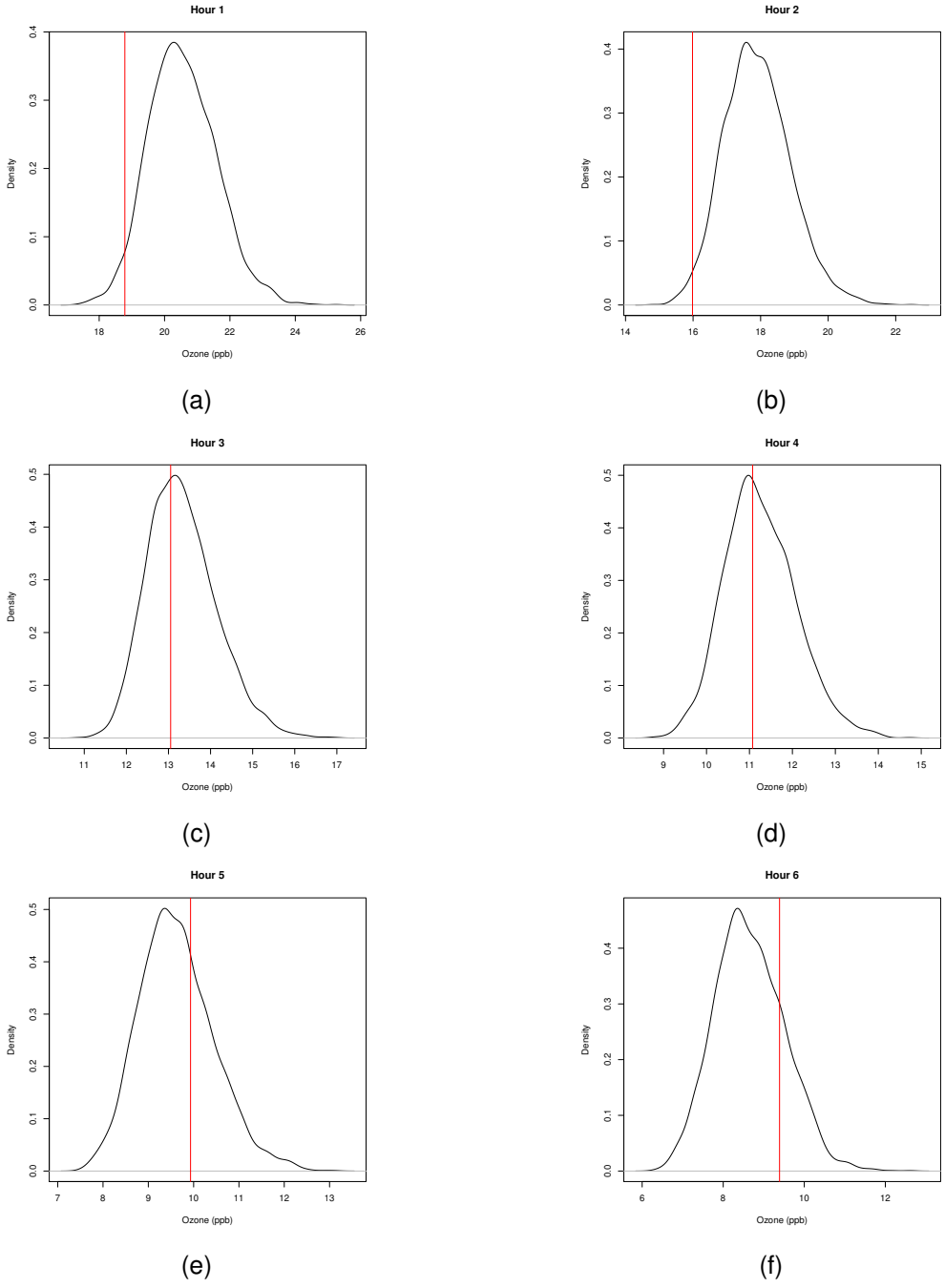


Figure 8: Posterior predictive distributions overlayed with true values (indicated by vertical lines) of Model 9, for a few selected hours

In addition, we experimented with heterogeneous variance in Model 9. As is suggested in section 3.4, we let $e_t \sim N(0, \sigma_z^2 X_t)$ in Model 9.1, and $e_t \sim N(0, \sigma_z^2 X_t^2)$ in Model 9.2. Their predictive performance is summarized in Table 9, from which we can tell that richer modeling of variance did not gain in terms of forecasting accuracy. Further more, it widened average interval lengths for both models, indicating increased uncertainty. The likely explanation is that random X's are assigned to the variances of the Z's when forecasting, adding too much noise for good prediction.

Finally, we can apply our recommended model to another Village Green site. In particular, we considered the Washington, D.C. site, choosing the identical time window to that above for

the Durham site. With regard to temperature, in this time window the DC site is similar to the Durham site. However, as an area with higher population density, we anticipate higher ozone levels. Fitting Model 9 to the temperature and ozone collected at the DC station during the same time period produced a PMSE of 5.39, a CRPS of 1.31, and an average interval length of 8.24. These numbers may be compared with those for Model 9 in Table 7. The less accurate prediction and greater uncertainty may be attributed to higher levels and higher variability in the DC ozone data compared to that in Durham.

5 Summary

In this paper, we compared a wide variety of models with different hierarchical structures and autoregressive structures. We also explored modeling of periodicity and heterogeneity, all with a focus on real-time forecasting, *not* retrospective modeling. According to PMSE, CRPS, and average interval length, we identified the two-stage bivariate autoregressive model with periodic means structures and homogeneous variance (Model 9) to be the best forecasting model for 8-hour average ozone at the Village Green Durham station. By comparing across the three classes of models, we found that a first-order autoregression with periodicity characterizes ozone distributions while incorporating accurate information on temperature further improves ozone forecasting. Furthermore, this model easily meets our goal of real time forecasting, with each hourly forecast taking only 9 seconds on average. In fact, this opens the door to forecasting more highly resolved in time, e.g., minute by minute updated real-time forecasting should that prove to be of interest.

Finally, future work envisions extending the current temporal model to handle space-time variation. For example, we would like to forecast 8-hour average ozone over a grid of locations across the continental United States. In order to think about this, let s_i denote an observed location, and s_{0j} denote a new location (on the grid). We have $Y_t(s_i) = \frac{1}{8}\sum_{k=-4}^{3} Z_{t+k}(s_i)$. Now we are not only interested in forecasting $\tilde{y}_t = (y_t(s_1), y_t(s_2), ..., y_t(s_n))'$, but also in forecasting $\tilde{y}_t' = (y_t(s_{01}), y_t(s_{02}, ..., y_t(s_{0m}))'$. An extension of the foregoing with real-time potential is to model ozone at the observed locations using the recommended AR model, then make hourly forecasts at new locations using a Gaussian Process (GP) model. Below we show a tentative model for ozone, drawing upon our work here:

$$
\begin{aligned}
Z_t(s_i) &= \alpha_0(s_i) + \alpha_1 X_t(s_i) + a\cos(\frac{\pi(t \bmod 24)}{12}) + b\sin(\frac{\pi(t \bmod 24)}{12} \\
&\quad + \beta_1[Z_{t-1}(s_i) - (\alpha_0(s_i) + \alpha_1 X_{t-1}(s_i) + a\cos(\frac{\pi((t-1) \bmod 24)}{12}) \\
&\quad + b\sin(\frac{\pi((t-1) \bmod 24)}{12})] + e_t(s_i) \\
e_t(s_i) &\sim N(0, \sigma_x^2) \\
\tilde{\alpha}_0(s_i) &\sim MVN(\alpha_0 \mathbb{1}_n, \sigma_\alpha^2 R_{\psi_\alpha}) \\
(R_{\psi_\alpha})_{ij} &= exp(-\psi_\alpha ||s_i - s_j||)
\end{aligned}
$$

In addition, we show a tentative model for temperature, again drawing upon our work here:

$$
\begin{aligned}
X_t(s_i) &= a\cos(\frac{\pi(t \bmod 24)}{12}) + b\sin(\frac{\pi(t \bmod 24)}{12} \\
&\quad + \phi_1[X_{t-1}(s_i) - (a\cos(\frac{\pi((t-1) \bmod 24)}{12}) + b\sin(\frac{\pi((t-1) \bmod 24)}{12})] + \eta_t(s_i) \\
\eta_t(s_i) &\sim MVN(0, \sigma_x^2 R_{\psi_x}) \\
(R_{\psi_\eta})_{ij} &= exp(-\psi_\eta ||s_i - s_j||)
\end{aligned}
$$

After obtaining posterior samples of $\boldsymbol{\theta_x} = (a, b, \phi_1, \sigma_x^2, \psi_\eta)$ from the temperature model, we can draw posterior predictive samples of $\tilde{X}_{t+1}, \tilde{X}_{t+2}, \tilde{X}_{t+3}$. Combining those with posterior samples of $\boldsymbol{\theta_z} = (\alpha_1, a, b, \sigma_z^2, \sigma_\alpha^2, \psi_\alpha)$ from the ozone model, we can acquire posterior predictive samples of $\tilde{Z}_{t+1}, \tilde{Z}_{t+2}, \tilde{Z}_{t+3}$, then aggregate them to produce samples of $\tilde{y}_t$.

However, the substantial increase in the amount of data and increased computational demand due to the addition of spatial structure will likely render real-time forecasting infeasible. One way to enhance computational efficiency is to implement the kriging using $\tilde{y}_t$, i.e., sample $y_t(s_0 j) \sim y_t(s_0 j)|\tilde{y}_t$ instead of $y_t(s_0 j) \sim y_t(s_0 j)|(\tilde{Z}_l, \tilde{X}_l), l = 1, ..., t$. Additional strategies to improve computing speed include modeling sub-regions of the U.S. in parallel as well as programming in a lower level language than R, such as C++.

Tables

Parameter	Value	Parameter	Value
Z_1	$\log 13.85$	X_1	$\sqrt{20}$
α_0	-1.75	ϕ_{1_x}	1
α_1	2.05	a_x	-0.19
ϕ_{1_z}	0.91	b_x	-0.06
a_z	-0.19	σ_x^2	0.002
b_z	-0.76		
σ_z^2	0.08		

Table 1: True parameters of ozone (left) and temperature (right) models for simulated data

Model	PMSE	CRPS	Average Interval Length
True	0.07	0.20	2.41
Alternative 1	1.39	0.66	4.39
Alternative 2	0.41	0.36	3.02
Alternative 3	0.46	0.40	3.43

Table 2: Summary of predictive performance of models for simulated data

Model	PMSE	CRPS	Average Interval Length
Bench mark	3.88	-	-
1	3.92	1.16	7.73
2	3.45	1.11	8.93
3	2.16	0.85	7.20

Table 3: Summary of predictive performance of class I models

Model	One-Step	Two-Step	Three-Step	Overall
Bench Mark	1.83	5.72	10.77	6.11
4	4.45	11.05	18.96	11.49
5	3.36	11.33	22.37	12.35
6	2.46	5.33	7.67	5.15

Table 4: Summary of second stage PMSE of class II models, by step

Model	PMSE	CRPS	Average Interval Length
Bench Mark	14.29	2.38	7.18
4	14.56	2.35	8.04
5	15.82	2.41	8.39
6	16.99	2.6	7.70

Table 5: Summary of predictive performance of class II models

Model	PMSE	CRPS	Average Interval Length
7	3.11	0.97	6.48
8	4.05	1.15	6.58

Table 6: Summary of predictive performance of class III models

Model	PMSE	CRPS	Average Interval Length
9	2.08	0.82	6.63
10	2.48	0.94	8

Table 7: Summary of predictive performance of class IV models

Parameter	Posterior Mean	95% CI
α_0	-1.96	(-9.41, 6.16)
α_1	2.11	(0.36, 3.80)
ϕ_{1_z}	0.91	(0.78, 1.01)
a_z	-0.17	(-0.60, 0.27)
b_z	-0.76	(-1.09, -0.43)
σ_z^2	0.08	(0.06, 0.11)
a_x	-0.19	(-0.23, -0.15)
b_x	-0.06	(-0.10, -0.02)
ϕ_{1_x}	1.000	(0.997, 1.002)
η_t^2	0.002	(0.001, 0.002)

Table 8: Summary of posterior parameter distributions of Model 9, for one forecast

Model	PMSE	CRPS	Average Interval Length
9.1	2.10	0.86	7.79
9.2	2.27	0.99	10.73

Table 9: Predictive performance of heterogeneous variations of Model 9

References

Achar, J.A., Rodrigues, E.R., Tzintzun, G. (2011). Using stochastic volatility models to analyse weekly ozone averages in Mexico City. *Environmental and Ecological Statistics*, 18, 271-290.

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

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

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

Camalier, L., Cox, W., Dolwick, P. (2007). The effects of meteorology on ozone in urban areas and their use in assessing ozone trends. *Atmospheric Environment*, 41, 7127-7137. doi:10.1016/j.atmosenv.2007.04.061

Duan, J., Tan, J., Yang, L., Wu, S., Hao, J. (2008). Concentration, sources and ozone formation potential of volatile organic compounds (VOCs) during ozone episode in Beijing. *Atmospheric Research*, 88, 25-35. doi:10.1016/j.atmosres.2007.09.004

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

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

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

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

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

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

US EPA (2009). Assessment of the impacts of global change on regional US air quality: A synthesis of climate change impacts on ground-level ozone. Technical Report MSU-CSE-00-2, U.S. Environmental Protection Agency, Washington DC.

US EPA (2013). Integrated science assessment of ozone and related photochemical oxidants. Technical Report EPA/600/R-10/076F, U.S. Environmental Protection Agency, Washington DC.

Wei, W. W. S. (1994). Time series analysis. Reading: Addison-Wesley publ.

Wilson, A., Rappold, A. G., Neas, L. M, and Reich, B. J. (2014). Modeling the effect of temperature on ozone-related mortality. *The Annals of Applied Statistics*, 8, 1728-1749.