Spatio-Temporal Analysis of Total Nitrate Concentrations Using Dynamic Statistical Models

Sujit K. Ghosh*, Prakash V. Bhave†,
Jerry M. Davis** and Hyeyoung Lee‡

Last updated on:September 30, 2009

Abstract

Atmospheric concentrations of total nitrate (TNO_3), defined here as gas-phase nitric acid plus particle-phase nitrate, are difficult to simulate in numerical air quality models due to the presence of a variety of formation pathways and loss mechanisms, some of which are highly uncertain. The goal of this study is to estimate the relative importance of these different pathways across the eastern United States by identifying empirical relationships that exist between TNO_3 concentrations and a set of covariates (ammonium, sulfate, ozone, wind speed, relative humidity, and precipitation) measured from January 1997 to July 2004. We develop two dynamic statistical models to quantify these relationships. A major advantage of these models over typical linear regression models is that their regression coefficients can vary temporally. Results show that TNO_3 is sensitive to ozone throughout the year, indicating an importance of daytime photochemical production of TNO_3, especially in the Southeast. Sensitivity of TNO_3 to residual ammonium ($NH_4^+ - 2SO_4^{2-}$) is most pronounced during winter, indicating a seasonal importance of gas/particle partitioning that is accentuated in the Midwest. Using a number of physical and chemical explanations, confidence is established in the spatial and temporal patterns of several such empirical relationships. In the future, these relationships may be used quantitatively to improve our mechanistic understanding of TNO_3 formation pathways and loss mechanisms in the atmosphere.

Keywords: chemical production, deposition, dynamic linear models, nitrate aerosol, spatial models.

---

*Sujit Ghosh is a Professor in the Department of Statistics at NC State University, †Prakash Bhave is a physical scientist in the Atmospheric Modeling and Analysis Division at the United States Environmental Protection Agency, **Jerry Davis is Professor Emeritus in the Department of Marine, Earth & Atmospheric Sciences at NC State University and ‡Hyeyoung Lee is a statistical modeler in a CRM division of Samsung Life Insurance.

1 Introduction

Nitrate is one of the major components of fine particulate matter ($PM_{2.5}$) across the United States (Malm et al., 2004), but it is one of the most difficult components to simulate accurately using numerical air quality models (Yu et al., 2005). Due to its semivolatile nature, nitrate partitions rapidly between gas-phase nitric acid (HNO_3) and fine particulate nitrate (NO_3^-). Therefore, accurate simulations of particulate nitrate require knowledge of the total nitrate (TNO_3) concentration as well as the partitioning behavior of TNO_3 into HNO_3 and NO_3^-. Thermodynamics of the inorganic aerosol system have been studied in detail over the past few decades (Ansari and Pandis, 1999; Zhang et al., 2000; references therein), making it possible to determine quite accurately the partitioning behavior of TNO_3. However, numerical simulation of ambient TNO_3 concentrations remains a significant challenge (Appel et al., 2008) because various atmospheric formation pathways and loss mechanisms exist and some of them are highly uncertain.

During the day, TNO_3 is produced predominantly by the following chemical reaction:

$$NO_2 + OH \rightarrow HNO_3 \qquad \text{(R1)}$$

At night, TNO_3 is produced by a series of reactions:

$$NO_2 + O_3 \rightarrow NO_3 + O_2 \qquad \text{(R2a)}$$

$$NO_2 + NO_3 \rightarrow N_2O_5 \qquad \text{(R2b)}$$

$$N_2O_5 + H_2O \rightarrow 2HNO_3 \qquad \text{(R2c)}$$

The N_2O_5 hydrolysis reaction (R2c) occurs in the gas phase and on particle surfaces, but its rate is highly variable and uncertain (see Brown et al., 2006; Davis et al., 2008.).

In general, TNO_3 may be removed from the atmosphere by wet deposition (i.e., rain out) and dry deposition. Wet deposition rates are strongly dependent on precipitation, whereas dry deposition depends on the partitioning of TNO3 between the gas and particle phases because the dry deposition velocity of HNO_3 is significantly greater than that of NO_3^-.

For effective air quality management, it is of interest to know the relative importance of each TNO_3 production and loss pathway at different times and locations. The only available method for accomplishing this objective involves the use of numerical air quality models (Gipson, 1999; Alexander et al., 2009), but those results are subject to the rather large uncertainties in several of the modeled processes. An alternate approach is to identify

empirical relationships that exist between TNO_3 concentrations and a set of observed variables that act as surrogates for the different TNO_3 formation and loss pathways. To quantify these empirical relationships, we employ the Reparameterized Dynamic Space Time Models (RDSTM) developed by Lee and Ghosh (2008). We use $\ln(TNO_3)$ as the response variable and consider the following variables as explanatory or predictor variables within a regression model framework: sulfate (SO_4^{2-}), ammonium (NH_4^+), ozone (O_3), temperature (T), relative humidity (RH), wind speed (WS), precipitation (P), solar radiation (SR), and dew point temperature (T_d). The RDSTM allows us to estimate dynamic relationships that may vary in time between $\ln(TNO_3)$ and the explanatory variables. In the future, this empirical information may be useful to diagnose and improve numerical air quality model predictions of ambient TNO_3.

Although we use the recently developed RDSTM framework for our data analysis, there are many other interesting and sophisticated models that could also be used to analyze data sets like ours. In recent years, there has been widespread attention in the statistical literature given to space-time data (Mardia and Goodall, 1993; Cressie, 1993; Mardia et al., 1998; Kyriakidis and Journel, 1999; Wikle and Cressie, 1999; Brown et al., 2000; Stroud et al., 2001; Kent and Mardia, 2002; Gelfand et al., 2005; Sahu and Mardia, 2005). In particular, environmental problems which are commonly temporally rich in data have motivated an extensive use of multivariate time series analysis techniques (Guttorp et al., 1994; Carroll et al., 1997). Often the primary interest in modeling space-time data is to predict the time evolution of a response variable over a given spatial domain. Typically, such predictions are made from data observed on a number of variables which themselves vary over time and space. Statistical models are employed in order to obtain accurate predictions of a response variable, such as concentrations of an air pollutant. Such models, if appropriately chosen, allow for accurate forecasting for near-future time periods and interpolation over the entire spatial region of interest.

Various approaches have been proposed to model space-time processes (Kyriakidis and Journel, 1999). One can consider the space-time problem from a multivariate geostatistical perspective, which requires that the space-time covariance functions be specified (Cressie and Huang, 1999; Gneiting, 2002; Schmidt and O'Hagan, 2003; Banerjee et al., 2004, Section 8.3). This approach has been limited in that the known class of valid space-time covariance functions is quite small, and such covariance functions are often not realistic for complicated dynamical processes. In addition, high dimensionality of these space-time models may prohibit practical implementation,

which can perhaps be avoided by our RDSTM framework.

Space-time processes can also be considered from a multiple time series perspective. That is, each spatial location is associated with a time series. Then, multivariate time series models can be used to analyze the space-time data (Gelfand et al., 1998, Kyriakidis and Journel, 1999; Shumway and Stoffer, 2000). Such methods have been difficult to implement in cases where dimensionality is high, that is, the number of spatial locations is large, e.g., for our nitrate data there were 33 locations in the eastern U.S.

Space-time models are often constructed by combining traditional time series techniques with methods from spatial statistics. In the time series context, popular approaches include ARIMA models (Box et al., 1994) for stationary data, and dynamic linear models (West and Harrison, 1997), which allow for nonstationary components such as temporal trends and seasonality. In the spatial setting, much of the literature revolves around isotropic second order stationary models (Cressie, 1993). A limitation of these methods is that such regularity assumptions may not be valid in practice, especially when the number of spatial locations is large or we observe volatilities over a long period of time.

Early attempts to develop space-time models assumed temporal stationarity. In an early Bayesian application, Handcock and Wallis (1994) considered the space-time modelling of winter temperature data observed over a region in the northern United States. They employed stationary Gaussian process models with an autoregressive model for the time series at each location and carried out separate spatial analyses to study global warming in each year. Carroll et al. (1997) again used stationary Gaussian processes, assuming a separable form for the space-time covariance function to study ground level ozone. Their model combines trend terms incorporating temperature and hourly or monthly effects, and an error model in which the correlation in the residuals is a nonlinear function of time and space, in particular the spatial structure is a function of the lag between observations.

Many researchers have developed space-time models that allow for nonstationary components. Guttorp et al. (1994) modeled the spatial covariances of hourly ozone levels using the Sampson and Guttorp (1992) nonparametric spatial covariance approach. They allowed the parameters of the model to vary as a function of time of day. Other approaches involving hierarchical Bayesian models include Wikle et al. (1999) and Waller et al. (1997). Wikle et al. (1999) analyzed monthly maximum atmospheric temperatures, and Waller et al. (1997) used generalized linear models to map lung cancer rates in Ohio. Two other notable contributions include Huerta et al. (2004) and Calder (2007). The latter uses a novel space-time approach

to a data set obtained from CASTNet.

Despite such a vast amount of literature on spatial and temporal models, our work based on the RDSTM framework has several distinguishable features: (i) there have been no other attempts to investigate spatio-temporal variability of TNO_3 in the statistics literature, though other pollutants (e.g., ozone, sulfate) have been examined in detail; (ii) compared to typical linear regression models, RDSTM as well as several of the space-time models referenced in the preceding paragraphs have the ability to estimate dynamic regression coefficients, which are critical for explaining atmospheric TNO_3 formation and loss because several dependencies vary seasonally; (iii) our covariate selection procedure is based on a combination of traditional statistical variable selection coupled with the ability to represent specific atmospheric formation and loss pathways rather than only their ability to explain the variability in the response variable. This expert-knowledge based scientific method to covariate selection increases the value of our results for the atmospheric research community.

This article is organized as follows. In Section 2, we describe the observational data and present some preliminary statistical analyses to identify the important predictors of TNO_3. In Section 3, we provide a brief description of the RDSTM. Results are presented in Section 4 and directions for future research are discussed in Section 5.

2 Data Description and Exploratory Analyses

2.1 Atmospheric Measurements

All of the data for this study are obtained from the U.S. EPA Clean Air Status and Trends Network (CASTNet) sites, which are located in rural areas. A complete description of this network can be found at the website: `http://www.epa.gov/castnet`. This study uses data from 33 sites in the eastern U.S. which are selected to overlap spatially with the major point sources of NO_X ($NO_2 + NO$) emissions (see Figure 9 in the Appendix).

One of main uses of the U.S. EPA CASTNet data is to evaluate the ability of deterministic air quality models to simulate the levels of various air pollutants in the atmosphere. CASTNet data have been used extensively to evaluate the U.S. EPA Community Multiscale Air Quality (CMAQ) model. Swall and Davis (2006) used a Bayesian statistical approach to evaluate the CMAQ model predictions of sulfate aerosol against CASTNet data. Eder and Yu (2006) and Appel et al. (2008) described full-year evaluations of the CMAQ model across several monitoring networks including the CASTNet.

Zheng et al. (2007) compared two statistical methods (the dynamic linear model method and the generalized additive model method) to estimate ozone trends in the eastern US and to adjust for meteorological effects. To our knowledge, none of the previous analyses of CASTNet data sought to capture a dynamical relationship between TNO_3 and other measured variables.

The chemical species used in this study are nitric acid, particulate, nitrate, sulfate, ammonium, and ozone. Their respective formulas and units are HNO_3 ($\mu mol/m^3$), NO_3^- ($\mu mol/m^3$), SO_4^{2-} ($\mu mol/m^3$), NH_4^+ ($\mu mol/m^3$), and O_3 (ppb). Ozone data are available on an hourly basis, but the other chemical species are measured in weekly integrated samples beginning every Tuesday. The maximum hourly O_3 values on each day are averaged from Tuesday to Tuesday to get weekly values. Nitric acid and nitrate are summed to get TNO_3 ($\mu mol/m^3$). Residual ammonium ($\mu mol/m^3$) is calculated as $NH_4^+ - 2SO_4^{2-}$, and considered in the analysis because it provides an estimate of the amount of ammonium that is associated with fine particulate NO_3^-. The factor of two is based on the implicit assumption that the preferred form of particulate NH_4^+ is ammonium sulfate ($(NH_4)_2SO_4$) (Malm et al., 2004).

Meteorological variables are also observed at each of the CASTNet stations. In this study we consider temperature, relative humidity, dew point temperature, solar radiation, wind speed, and precipitation as covariates for the statistical analysis. The respective symbols and units used for these variables are T ($^\circ C$), RH (%), T_d ($^\circ C$), SR (W/m^2), WS (m/s), and P ($mm/week$). Dew point temperature is calculated from T and RH. The remaining meteorological variables are measured hourly. To conform to the weekly chemical measurements, precipitation data are summed over each week and the other meteorological variables are averaged to obtain weekly values.

The data used in this study were collected between January 1997 and July 2004, encompassing 394 weeks. The total number of weeks with available data for TNO_3 and all of the potential covariates varies from station to station with a maximum of 394 (all weeks observed) and a minimum of 361. On average across all sites, only 3% of the observations were missing.

2.2 Exploratory Data Analysis

An examination of the yearly median values indicates little variation from year to year in the chemical covariates. An exception to this is O_3 where the median values were higher in 1998 and 1999 than the other years. The median values for the meteorological covariates also show little year to year

Table 1: Spearman Rank Correlation Coefficients between Observed Values

	TNO_3	SO_4	NH_4	$ResNH_4$	O_3	SR	T	T_d	WS	RH	P
TNO_3	1.00										
SO_4	0.27	1.00									
NH_4	0.58	0.83	1.00								
$ResNH_4$	0.38	-0.58	-0.11	1.00							
O_3	0.09	0.66	0.41	-0.61	1.00						
SR	0.06	0.58	0.35	-0.53	0.89	1.00					
T	-0.09	0.69	0.40	-0.69	0.80	0.80	1.00				
T_d	-0.16	0.63	0.36	-0.63	0.64	0.63	0.91	1.00			
WS	0.50	-0.17	0.07	0.45	-0.07	-0.03	-0.22	-0.28	1.00		
RH	0.00	0.29	0.30	-0.11	-0.05	-0.09	0.16	0.28	0.00	1.00	
P	-0.14	-0.10	-0.01	-0.11	0.08	0.03	0.18	0.20	0.05	0.39	1.00

variation. Among the chemical variables, TNO_3 shows the greatest site-to-site variation, while among the meteorological variables, WS exhibits the largest spatial variation. All of the chemical species show a seasonal pattern with HNO_3, SO_4^{2-}, NH_4^+, and O_3 having higher values during summer than in winter. In contrast, NO_3^- and TNO_3 are lower in the summer and higher in the winter. Among meteorological variables, WS peaks in winter, while T, T_d, and SR, are highest during summer. Precipitation tends to peak in the summer due to convective activity. Relative humidity is lowest in the spring and reaches peaks in late summer and mid-winter, but these seasonal variations are not large (see Figures A.1-12 in the appendix of Lee, 2006).

Table 1 summarizes Spearman rank correlation coefficients between pairs of measured variables. The Spearman rank correlation is a nonparametric measure of the association between two variables based on the rank of the observed values of the two variables. It is known to be more robust than the Pearson correlation coefficient which measures the linear relationship between two variables (Steel et al., 1997). Table 1 shows that TNO_3 is only marginally correlated with a couple of the potential covariates, illustrating the difficulties in developing an empirical relationship. High correlations (0.80 or above) are found between O_3, SR, and T, and between T and T_d. These four variables exhibit moderate positive correlations with SO_4^{2-} (0.58 to 0.69) and negative correlations with residual ammonium $ResidNH_4$ (-0.69 to -0.53). Whereas SO_4^{2-} exhibits a high correlation with NH_4^+ (0.83), its correlation with P (-0.10) is the lowest (in magnitude) in Table 1. Knowledge of these correlations assist with the covariate selection process described in the following section.

A frequency histogram of all available TNO_3 data reveals a distribution

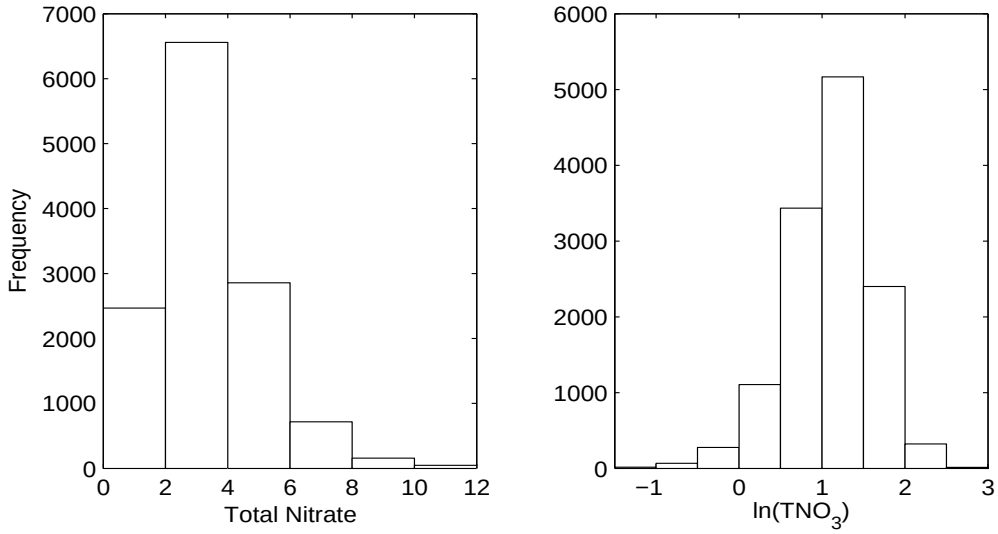

Figure 1: Frequency histograms of TNO_3 and $\ln(TNO_3)$

that is right-skewed, whereas that of $\ln(TNO_3)$ more closely resembles a symmetric and bell-shaped distribution (see Figure 1). Although we use the terms "$\ln(TNO_3)$" and "total nitrate concentrations" in this text, $\ln(TNO_3)$ is the response variable in both of the statistical models discussed below. We also (linearly) standardize the chemical species and meteorological variables such that the covariates have an empirical mean of 0 and a variance of 1.

2.3 Covariate Selection

Before we apply the RDSTM to CASTNet data, covariates must be selected carefully, avoiding statistical pitfalls such as multicollinearity, to maximize our chances of obtaining results that are physically and chemically interpretable. Conventional methods for covariate selection were initially considered, such as the stepwise approach and best-subsets. These methods seek to determine the set of covariates that optimally explain the variability in the response variable. The best-subsets procedure (Miller, 2002) based on minimizing the mean residual sum of squares selected the following variables as covariates: SO_3, NH_3, ResidNH_4, O_3, T_d, WS, T and P, whereas the LASSO procedure (Tibshirani, 1996) based on a L_1-penalization approach selected all of the above variables except T as covariates.

However, the objective of this study is slightly different from typical model applications. As discussed in Section 1, we seek to develop empirical

relationships between ln(TNO_3) and a set of covariates that act as surrogates for the different formation and loss pathways. Selecting covariates which meet this objective is a challenge because very few measured variables can uniquely represent a single pathway, due to the various interdependent processes in the atmosphere. For example, the SO_3 variable is closely related with numerous pathways that affect TNO_3 concentrations. Both SO_3 and TNO_3 originate from similar emission sources (power plants are the main source of sulfur oxides and nitrogen oxides in rural areas) and oxidation processes (e.g., reaction with OH), both build up during stagnant meteorological conditions, both are diluted during periods of high winds, and both are removed efficiently during precipitation events. Therefore, any model that includes SO_3 as a covariate would deduce that the variability in TNO_3 is dominated by the variability in SO_3. Such a result would not enhance our understanding of the relative importance of different TNO_3 formation and loss pathways in the atmosphere. Thus, we opt against using the covariates selected by either the best-subsets or LASSO approaches.

Instead, we apply our knowledge of the atmospheric processes to select five covariates that meet the objectives of this study: ResidNH_4, O_3, WS, RH, and P. ResidNH_4 is an indicator of the gas/particle partitioning behavior of TNO_3. Nitrate will partition preferentially to the particle phase (NO_3^-) when ResidNH_4 is large and it will partition to the gas phase (HNO_3) when ResidNH_4 is small. Thus, ResidNH_4 serves as a surrogate for TNO_3 removal by dry deposition (see Section 1). Ozone serves as a surrogate for the OH radical which plays a major role in the daytime production of TNO_3 (see equation (R1)). WS affects all pollutant concentrations through dilution and also impacts the dry deposition velocities of HNO_3 and NO_3^-. RH may play an important chemical role both at night and during the day. High daytime RH favors partitioning of TNO_3 to the particle phase. At night, high RH enhances the formation of TNO_3 via N_2O_5 hydrolysis (see equation (R2c)). Precipitation acts as an atmospheric scavenging agent for TNO_3 so it is a surrogate for TNO_3 removal by wet deposition. While making the above selections, the information in Table 1 is used to avoid multicollinearity problems. For example, we eliminate SR, T, and T_d because they are all highly correlated with O_3.

3 Reparametrized Dynamic Space-Time Models (RDSTM)

3.1 The Full RDSTM

The approach taken here is to view the data as arising from a vector-valued time series where each component of the vector corresponds to a spatial location. From this perspective, it might be more appropriate to call our model a reparametrized multivariate time series model instead of a spatio-temporal model. We adapt to the framework of dynamic linear models (DLM, West and Harrison, 1997), which describe the temporal evolution of the spatial vector in a latent space. Suppose the response variable $Z(s,t)$ is observed at a finite number of sites labeled as $s_1, ..., s_n$ at each time t, where $t = 1, 2, \cdots, m$. Consider the $n \times 1$ vector time series $\boldsymbol{Z}_t = (Z(s_1,t), ..., Z(s_n,t))^T$ at time t. For each t, the DLM is usually characterized by an observation equation and an evolution equation. An observation equation describes the relationship between the vector of observations ($\boldsymbol{Z}_t$) and the matrix of regressors ($\boldsymbol{X}_t$) that takes the form of a multivariate regression model evolving over time

$$\boldsymbol{Z}_t = \boldsymbol{X}_t\boldsymbol{\beta_t} + \boldsymbol{\nu_t}, \quad \boldsymbol{\nu}_t \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_t^{\nu}) \tag{1}$$

where $\boldsymbol{X}_t$ is an $n \times p$ observed design matrix (containing a row of 1's for the intercept) and $\boldsymbol{\beta}_t$ is a $p \times 1$ vector of regression coefficients or state parameters. An evolution equation describes the dynamics of the vector of regression coefficients or state parameters $\boldsymbol{\beta}_t$ through time

$$\boldsymbol{\beta}_t = \boldsymbol{G}_t\boldsymbol{\beta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_t^{\omega}) \tag{2}$$

where $\boldsymbol{G}_t$ is a $p \times p$ evolution matrix. There are several ways to model the $\boldsymbol{G}_t$'s. The most common assumption is that the $\boldsymbol{G}_t$'s are structurally known, possibly up to some finite number of parameters. In our study, we do not make any structural assumption about the $\boldsymbol{G}_t$'s but we assume that $\boldsymbol{G}_t = \boldsymbol{G}$ for all t and that $\boldsymbol{G}$ follows a matrix-valued normal distribution with mean $\boldsymbol{G}_0$ and variance-covariance parameters $\boldsymbol{\Omega}_0$ and $\boldsymbol{\Sigma}_0^G$ (Nagar and Gupta, 2000, chapter 2). We also assume that the $\boldsymbol{\nu}_t$ and $\boldsymbol{\omega}_t$ error vectors are independent and have multivariate normal distributions with mean $\boldsymbol{0}$ and variance-covariance matrices $\boldsymbol{\Sigma}_t^{\nu} = \boldsymbol{\Sigma}^{\nu}$ and $\boldsymbol{\Sigma}_t^{\omega} = \boldsymbol{\Sigma}^{\omega}$, respectively, for all t. The model is completed with a normal prior for the initial state, $\boldsymbol{\beta}_1 \sim N(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0^{\omega})$, where $\boldsymbol{\beta}_0$ is known. Inverse-Wishart distributions (Nagar and Gupta, 2000, chapter 3) can be used as priors for $\boldsymbol{\Sigma}^{\nu}$ and $\boldsymbol{\Sigma}^{\omega}$.

The updating scheme in the dynamic space-time model (which involves several matrix operations) may not be easy to implement when the $\boldsymbol{G}$ matrix is completely unknown. Further, these types of multivariate updating schemes can be very unstable and time consuming when the dimensions are very large and some intermittent observations are missing. In order to avoid such numerical instabilities and to accelerate model fitting, we use an equivalent univariate scheme for the aforementioned DLM using a novel reparametrization method (Lee and Ghosh, 2008).

Suppose Z_{it} denotes $\ln(\text{TNO}_3)_{it}$ and X_{itk} the k^{th} covariate (centered and scaled) at site i and time t, where $i = 1, \cdots, n = 33$, $t = 1, \cdots, m = 394$, and $k = 1, \cdots, p = 6$. Notice that $\boldsymbol{Z}_t = (Z_{1t}, \cdots, Z_{nt})^T$ and $\{\boldsymbol{X}_t\}_{n\times p} = ((X_{itk}))_{1\leq i\leq n, 1\leq k\leq p}$. Recall that $X_{it1} = 1$ for all i and t, i.e., the first column represents the intercept. Then following the work of Lee and Ghosh (2008), the RDSTM consists of the observation equation that can be written as

$$Z_{it} = \sum_{k=1}^{p} \beta_{kt} X_{itk} + \sum_{i'=1}^{i-1} \phi_{ii'} Z_{i't} + \nu_{it}, \tag{3a}$$

$$Z_{1t} = \sum_{k=1}^{p} \beta_{kt} X_{1tk} + \nu_{1t}, \quad \nu_{1t} \sim N(0, \sigma^2_{\nu 1}) \tag{3b}$$

where $\nu_{it} \sim N(0, \sigma^2_{\nu i})$ for $i = 2, \cdots, n$ and $t = 1, \cdots, m$. In (3) above, notice that $\phi_{ii'}$ denotes the entries of the lower triangular matrix of the Cholesky decomposition $\boldsymbol{T}\boldsymbol{\Sigma}^{\nu}\boldsymbol{T}^T = \boldsymbol{D}$ of the positive definite matrix $\boldsymbol{\Sigma}^{\nu}$ where $\boldsymbol{D}$ is the diagonal matrix of eigenvalues ($\sigma^2_{\nu i}$'s) of $\boldsymbol{\Sigma}^{\nu}$ and $\boldsymbol{T}$ is the unique lower triangular matrix with $\phi_{ii'}$'s as its lower triangular entries and all the diagonal entries being equal to unity. The evolution equation can now be written as

$$\beta_{kt} = \sum_{k'=1}^{p} \beta_{k't-1} g_{kk'} + \sum_{k'=1}^{k-1} \psi_{kk'} \beta_{k't} + \omega_{kt}, \tag{4a}$$

$$\beta_{1t} = \sum_{k'=1}^{p} \beta_{k't-1} g_{1k'} + \omega_{1t}, \quad \omega_{1t} \sim N(0, \sigma^2_{\omega 1}) \tag{4b}$$

where $\omega_{kt} \sim N(0, \sigma^2_{\omega k})$ for $k = 2, \cdots, p$, $t = 2, \cdots, m$ and the initial state equation can be written as,

$$\beta_{k1} = \beta_{k0} + \sum_{k'=1}^{k-1} \psi_{kk'} \beta_{k'1} + \omega_{k1}, \tag{4c}$$

where $\omega_{k1} \sim N(0, \sigma^2_{\omega k})$ for $k = 2, \cdots, p$. Here again, $\psi_{kk'}$ denotes the entries of the lower triangular matrix of the Cholesky decomposition of the positive definite matrix $\mathbf{\Sigma}^{\omega}$, and $\sigma^2_{\omega k}$ denotes the kth eigenvalue of $\mathbf{\Sigma}^{\omega}$.The model is completed with

$$\beta_{11} = \beta_{10} + \omega_{11}, \quad \omega_{11} \sim N(0, \sigma^2_{\omega 1}). \tag{4d}$$

Notice that the multivariate observation model in (1) corresponds to the system of univariate observation models (3a,b) and similarly, the multivariate evolution model (2) corresponds to the system of univariate evolution models (4a,b,c,d). Using these univariate reparametrized regression models we avoid numerical instabilities due to high dimensionality that could occur in a multivariate scheme. Also, this allows missing data to be imputed from their full conditional distributions. In addition, by allowing the $\phi_{ii'}$'s to be completely unstructured, the RDSTM does not require simplifying assumptions like stationarity, isotropy, etc. for the spatial covariance function.

Our results from the RDSTM are obtained numerically using a Markov chain Monte Carlo (MCMC) procedure via the `WinBUGS` software available at: `http://www.mrc-bsu.cam.ac.uk/bugs/`. As our data involves missing observations, the proposed RDSTM performs univariate imputations using Gibbs sampling as opposed to multivariate imputations. Gibbs sampling provides a natural solution by imputing values for the missing data at each iteration by sampling from their full conditional distribution given the available data. Regression coefficients are then updated conditionally on the imputed values. We assume that each of the standardized covariates, when missing, follows a standard normal distribution (i.e., $X^{miss}_{itk} \sim N(0, 1)$). We analyze the data using vague priors (i.e., proper priors with large variance) on parameters to have minimal impacts on the posterior inference. We assign independent zero-mean normal distributions with variance 10^3, denoted by $N(0, 10^3)$, as priors to $\phi_{ii'}$, $\psi_{kk'}$ and $g_{kk'}$, and independent $Ga(10^{-3}, 10^{-3})$ priors to $1/\sigma^2_{\nu i}$ and $1/\sigma^2_{\omega k}$, where $Ga(a, b)$ denotes a Gamma distribution with mean a/b and variance a/b^2. It can be shown that the above series of univariate priors corresponds to a special case of the inverse-Wishart prior for the covariance matrices (Nagar and Gupta, 2000). Notice that the $\sigma_{\nu i}$'s and $\sigma_{\omega k}$'s in our full RDSTM represent the eigenvalues of the original conditional covariance matrices $\mathbf{\Sigma}^{\nu}$ and $\mathbf{\Sigma}^{\omega}$, respectively, and hence the flat priors (e.g., $Ga(0.001, 0.001)$) do not necessarily lead to improper posteriors. In order to check the convergence issue numerically we re-ran our code with a few other values of a and b and found the parameter estimates virtually

the same as reported in the next section using the above mentioned default values. Though we could have used other weakly informative priors for the parameters, we would not expect the results to differ substantially.

3.2 A Simplified RDSTM

Difficulties encountered while interpreting results of the full RDSTM (see Section 4.1.2) motivated the development of a simplified model. In particular, we restrict our attention to $\boldsymbol{\Sigma}_t^{\nu} = \sigma_t^2 \boldsymbol{I}_n$, where I_n denotes an identity matrix (i.e., a diagonal matrix with diagonal entries as unity). Notice that the variance term (σ_t^2) is now allowed to change with time t. In other words, we assume that the TNO_3 observations (conditionally on the $\boldsymbol{\beta}_t$'s) are spatially independent but we use a site-specific intercept, $\boldsymbol{\alpha}$, to capture spatial differences in the mean concentrations. Thus a reduced version of the RDSTM can now be written as

$$\boldsymbol{Z}_t = \boldsymbol{\alpha} + \boldsymbol{X}_t \boldsymbol{\beta_t} + \boldsymbol{\nu_t}, \quad \boldsymbol{\nu}_t \sim N(\boldsymbol{0}, \sigma_t^2 \boldsymbol{I}_n), \tag{5}$$

where $\boldsymbol{X}_t$ is an $n \times (p-1)$ observed design matrix (the intercept term is omitted) and $\boldsymbol{\beta}_t$ is a $(p-1) \times 1$ vector of regression coefficients or state parameters. Notice, that with a little abuse of notation, we are re-using $\boldsymbol{X}_t$ and $\boldsymbol{\beta}_t$ for this simplified model although their dimensions are not identical to those in the full model. The evolution equation (2) remains unchanged except we now express the equation as follows:

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_0 + \boldsymbol{G}_t(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}_0) + \boldsymbol{\omega}_t, \quad \text{for } t = 2, 3, \ldots \text{ and } \boldsymbol{\omega}_t \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}^{\omega}), \tag{6}$$

where $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 + \boldsymbol{\omega}_1$. The time-varying variance paremeters σ_t^2 are modeled using an exchangeable process, $1/\sigma_t^2 \sim Ga(a_\nu, b_\nu)$, and an inverse-Wishart prior is used for $\boldsymbol{\Sigma}^{\omega}$. For the reduced RDSTM, the σ_t^2's represent time-varying variances of the measurement error process, but these are very well estimated by borrowing information across various sites at a given time point t. Hence, a flat prior does not lead to nearly improper posteriors. For comparison with equation (3a,b), the observation equation in the simplified model can be written as

$$Z_{it} \;\; = \;\; \alpha_i + \sum_{k=1}^{p-1} \beta_{kt} X_{itk} + \nu_{it}, \tag{7}$$

where $i = 1, \cdots, n$, $t = 1, \cdots, m$, and α_i is the site-specific intercept term. It is noted that the simplified RDSTM still captures the (unconditional)

spatial correlations among the response variable, and that

$$\begin{aligned} Cov[\mathbf{Z}_t, \mathbf{Z}_{t'}] &= E[Cov[\mathbf{Z}_t, \mathbf{Z}_{t'}|\boldsymbol{\beta}_t, \boldsymbol{\beta}_{t'}]] + Cov[E[\mathbf{Z}_t|\boldsymbol{\beta}_t], E[\mathbf{Z}_{t'}|\boldsymbol{\beta}_{t'}]] \\ &= Cov[\boldsymbol{\nu}_t, \boldsymbol{\nu}_{t'}] + \mathbf{X}_t Cov[\boldsymbol{\beta}_t, \boldsymbol{\beta}_{t'}]\mathbf{X}_{t'}^T, \end{aligned}$$

is not necessarily a diagonal matrix.

For both versions of RDSTM, we obtain 10,000 iterates using a single chain from the MCMC sampler. The first 5000 iterates are discarded as a part of the Markov chain burn-in period, and all the posterior summaries reported below are based on Monte Carlo estimates from the remaining 5000 iterates. The number of burn-in samples and final MCMC sample sizes are chosen using trace plots for the parameters. Trace plots of the sampled values versus the iteration number are examined for evidence of when the simulation appears to have stabilized to a stationary distribution. The WinBUGS code developed for this study are available on the journal website as a part of supplemental materials and can be adapted easily for applications to other data sets. The full model run takes about 6 hours while the reduced model takes about 5.2 hours on a Pentium 4 CPU 2.4GHz PC with 1.00 GB of RAM.

The output from both models are compared based on their predictive performances. In Section 4.2.1, we obtain site-specific predictive values of $\ln(\text{TNO}_3)$ using each of the RDSTM versions and compare with the observed values.

4 Results

4.1 Full RDSTM

4.1.1 Statistical Significance and Stationarity

For every covariate, k, and week, t, the RDSTM provides posterior estimates (e.g., posterior medians) of the dynamic regression coefficient, β_{kt}. The left half of Figure 2 illustrates the seasonal variation of β for each covariate. Weekly β values from the full RDSTM are binned by month-of-year to obtain the boxplot distributions shown. Regression coefficients for ResidNH_4 exhibit a distinct seasonal pattern, peaking during the cooler months and reaching a minimum from July – September. Conversely, the coefficients for O_3 and WS are highest during the summer months. Coefficients for RH and P do not exhibit a strong seasonal pattern. It is important to note that a typical linear regression model (LRM) would yield coefficients that are fixed

in time, thus missing the potentially important dynamic relationships shown in Figure 2 between $\ln(TNO_3)$ and the covariates.

To assess the statistical significance of these time-varying relationships, we check whether or not the 95% equal-tail credible interval (obtained by computing 2.5% and 97.5% posterior percentiles) for each regression coefficient overlaps with zero. Alternatively, one may also use Bayes factors to assess the significance of these time-varying regression coefficients, but we find significance testing using posterior intervals much simpler computationally for our high-dimensional models. Also, Bayes factors are known to be sensitive to priors and are not even well-defined when improper priors are used. We count two separate numbers in each month for all of the covariates. The first count provides the number of weeks which have significant positive coefficients (i.e., lower limit of the 95% interval is positive) and the other counts weeks with significant negative coefficients (i.e., the upper limit is negative). These counts are summarized in the right half of Figure 2. The dynamic regression coefficients for ResidNH_4, O_3, and WS, are found to be statistically significant during 57%, 54%, and 40% of all weeks in the study period, respectively.

To gain a better understanding of the stochastic aspects of our results, we check whether the covariance function obtained from the RDSTM (using $\phi_{ii'}$ and $\sigma^2_{\nu i}$) is stationary in nature. To this end, we first compute the posterior median of each entry of the conditional covariance matrix $\mathbf{\Sigma}^{\nu}$ and then we obtain the box plots of these posterior medians binned by inter-site distance and direction to obtain the correlograms in Figure 3. These provide a measure of spatial autocorrelation as a function of inter-site distance. In the spatial analysis, generally, correlations at short inter-site distances (under 100 km) are important to identify the characteristics of the covariance function (e.g., stationary, nonstationary, isotropic, or anisotropic). The CASTNet sites are somewhat sparse and unfortunately only a few site pairs are separated by less than 100 km. Hence, our data may be inadequate for this spatial analysis.

In Figure 3, the X-axis represents the inter-site distance in kilometers and the Y-axis represents the posterior median of the conditional correlation between residual $\ln(TNO_3)$ (after adjusting for predictors) at two different locations. We also computed three different directional correlograms to see if such residual concentrations appear to be isotropic. The first correlogram plot shows that the correlation has a low median value (0.1) at a distance 200km, and it decreases to zero as the inter-site distance increases. The underlying spatial variations seem to be stationary because the estimated conditional correlations computed at fixed distance appear to depend only

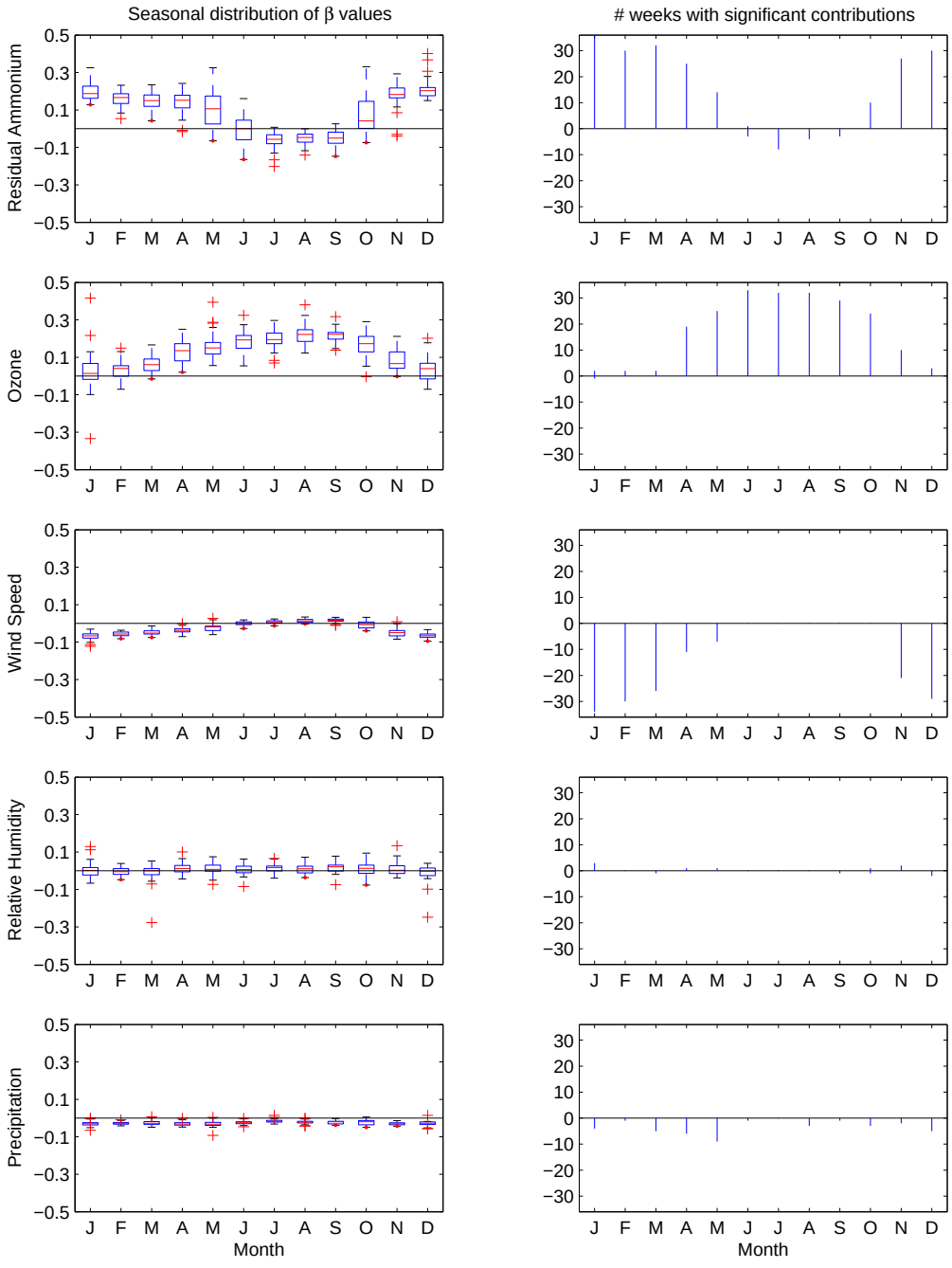

Figure 2: Monthly summary of dynamic regression coefficients from the full RDSTM

on distances between two locations. Also the underlying spatial variations seem isotropic because three different directional correlograms look similar, which means the conditional correlations are not changing significantly with direction.

Considering a variety of stationary isotropic structures for the covariance model of the RDSTM, we find that the exponential covariance model with parameter values of $\tau^2 \approx 0.029$ (nugget), $\sigma^2 \approx 0.008$ (partial sill), and $r \approx 20$ (range) has the smallest Frobenius distance (0.031) from the estimated conditional covariance matrix of RDSTM. Recall that a Frobenius distance between two square matrices is defined as the square root of the sum of the squared differences between their elements (Golub and van Loan, 1996, p.55). Hence the exponential covariance model seems appropriate to fit the conditional covariance ($\mathbf{\Sigma}^{\nu}$) of RDSTM. The advantage of using an unstructured covariance model is that we do not have to make *a priori* structural (parametric) assumptions (e.g., stationary, isotropic etc.) about the covariance. Instead, we can fit our unstructured model, which has the capability of capturing many features of the covariance, and let our data decide which structure is most appropriate. Moreover, unlike most spatial-temporal analyses in the literature, spatial interpolation is not a goal in this study as spatial smoothing is not even possible since the covariates are not available at unmonitored sites.

4.1.2 Physical and Chemical Interpretation

The RDSTM results at individual sites are examined in an effort to estimate the relative importance of different physical and chemical pathways influencing the ambient TNO_3 concentrations. As an example, model results from the Ann Arbor, MI site are illustrated in Figure 4. To visualize the results easily, each vertical bar represents a four-week average of RDSTM outputs rather than the individual weekly results. The solid black line represents the intercept term, β_{0t}, which is treated as a dynamic variable in the full RDSTM. The contributions of each covariate to $\ln(TNO_3)$ during each time step are shown as red (O_3), orange (RH), yellow (WS), green ($ResidNH_4$), and blue (P) patches. These contributions are computed by the product $\beta_{kt}X_{itk}$, in equation 3. Colored patches are plotted above the black line if the product is postive, and below the black line otherwise. Similarly, the contributions of spatial terms are plotted in purple. Those are computed from equation 3a as $\sum_{i'=1}^{i-1} \phi_{ii'} Z_{i't}$. The model error, ν_{it}, is shown as gray patches.

The small amount of gray area relative to the sum of all colorful patches

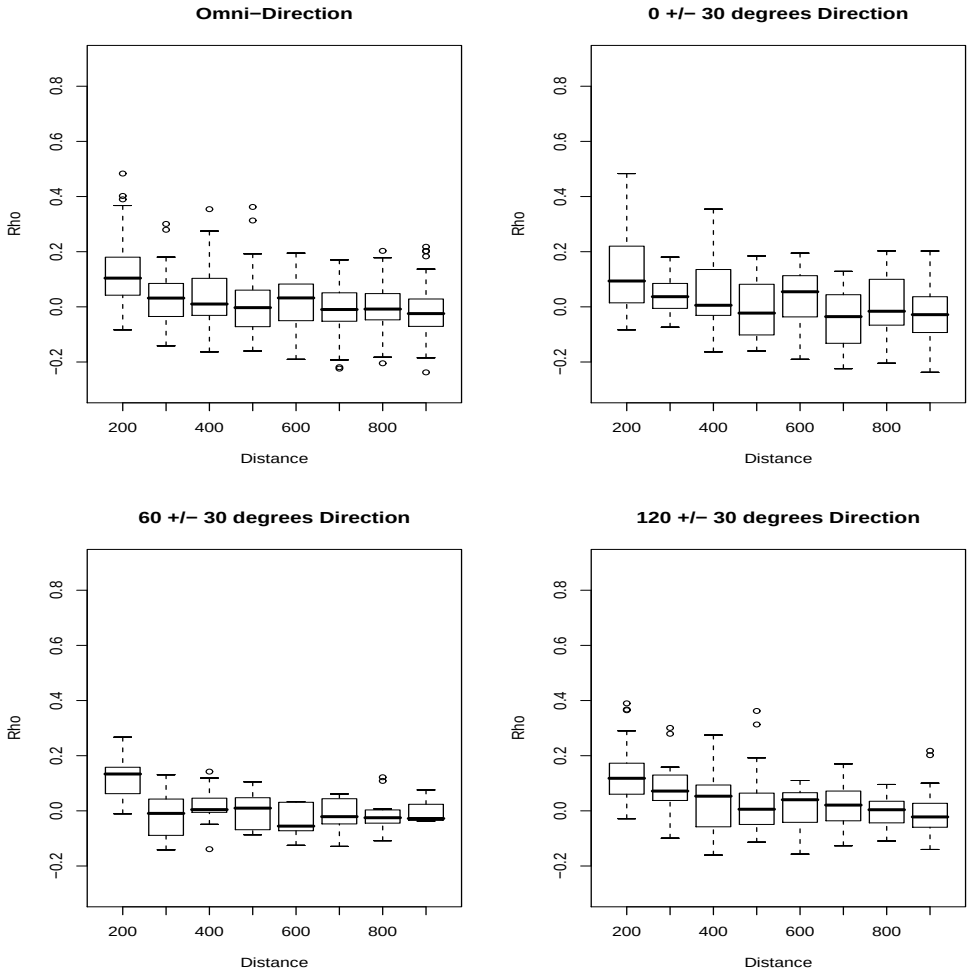

Figure 3: Fitted spatial correlogram in several directions

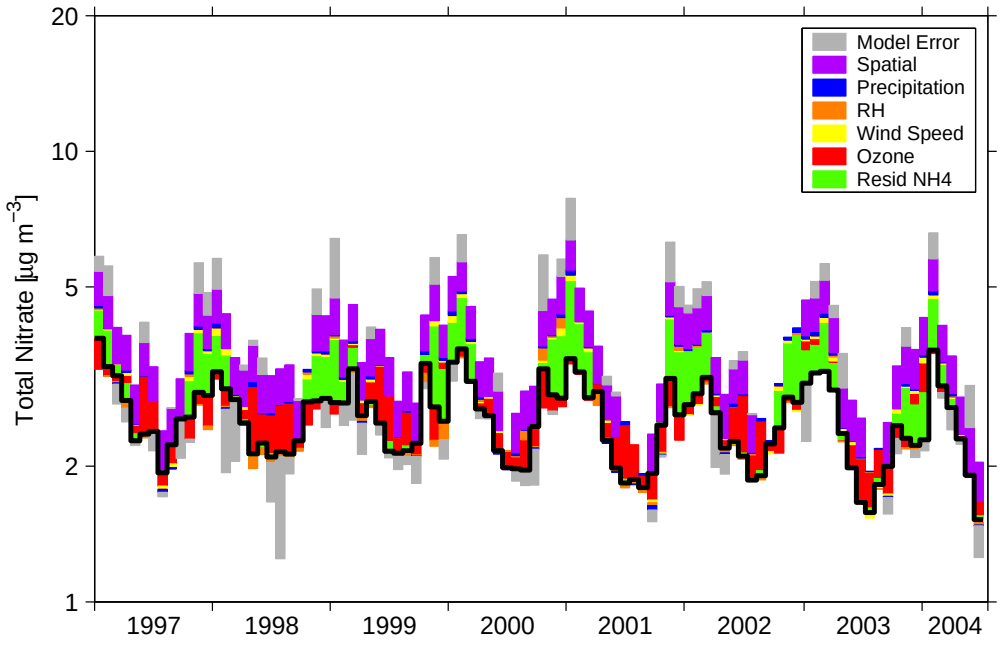

Figure 4: Covariate and spatial contributions to total nitrate at the Ann Arbor, MI site using the full RDSTM

in Figure 4 indicates that most of the variability in $\ln(TNO_3)$ at the Ann Arbor site is explained by the full RDSTM. However, two aspects of the full RDSTM formulation encumber our interpretation of Figure 4. First, the intercept term exhibits a pronounced seasonal cycle that peaks in the winter and reaches a minimum value every summer. This implies that much of the seasonal variability in $\ln(TNO_3)$ is not explained by any of the covariates and, therefore, is not being attributed to specific atmospheric processes. From the perspective of air pollution control, it is of little value to know that TNO_3 peaks every winter if the pathways producing that pollution remain unknown. Second, the spatial term is quite large and swamps the summed contribution from all observable covariates at many of the CASTNet sites. This result is also of limited value because it does not enhance our knowledge of the complex source and sink processes which control TNO_3.

Given our primary objective of determining the influence of each TNO_3 production and loss pathway in the atmosphere, it is critical that we maximize the amount of variability in $\ln(TNO_3)$ that can be tied to those pathways rather than to any latent variables. Analysis of the full RDSTM results motivated our development of the simplified RDSTM. As described in Section 3.2, the intercept term is fixed in time in the simplified RDSTM. The matrix of conditional spatial dependencies, $\phi_{ii'}$, is removed, but the intercept is allowed to vary with location to capture spatial differences in the temporally-averaged $\ln(TNO_3)$ concentrations.

4.2 Simplified RDSTM

4.2.1 Comparison to Full RDSTM

To establish confidence in the simplified RDSTM, we compare its predictive performance to that of the full RDSTM in Figure 5. Some degradation in model performance is expected because the simplified RDSTM contains neither a matrix of spatial terms nor a time-varying intercept. The left plot illustrates the performance of the full RDSTM, which matches 99.1% of the observed values within a factor of 2, 96.7% within a factor of 1.5, and 83.8% within factor of 1.25. The analogous performance statistics from the simplified RDSTM are only slightly lower: 98.9%, 93.3% and 74.7%. We therefore conclude that the simplified model is a reasonable substitute for the full RDSTM. It should be noted that the RDSTM performance statistics are substantially better than those of numerical air quality models for TNO_3 (Appel et al., 2008), so the empirical relationships drawn from either RDSTM will provide unique insights. The only data points which are poorly

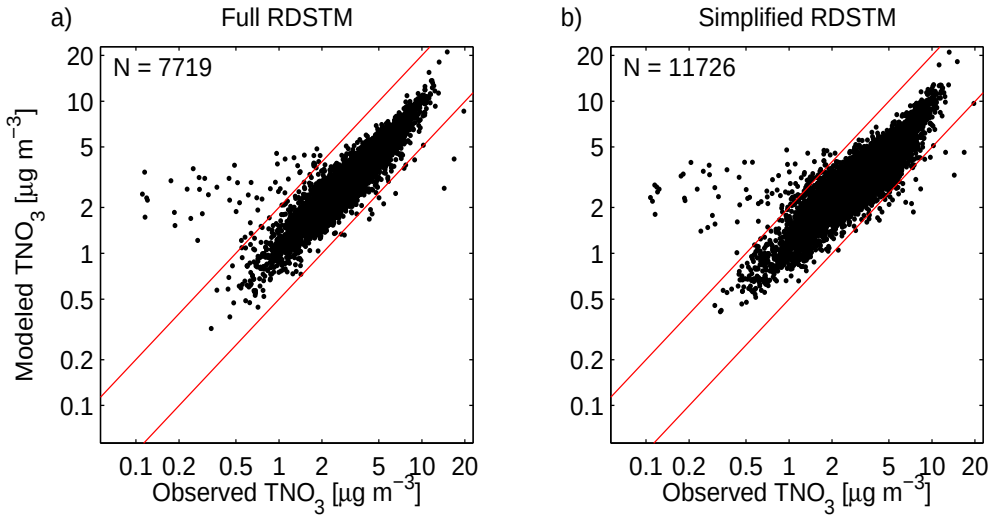

Figure 5: Comparison of performance between the full and simplified models. Red lines encompass the points where model predictions match the observed total nitrate concentrations within a factor of 2.

predicted by the RDSTM are those with very low $\ln(TNO_3)$ concentrations (i.e., less than 1 μg m^{-3}). From an air quality management perspective, model performance during such pristine conditions is of minor importance.

One practical advantage of the simplified model over the full RDSTM is that the former requires far less observational data to make a prediction of $\ln(TNO_3)$. Due to the spatial matrix $\phi_{ii'}$ in the full RDSTM, equation 3a can be evaluated only on weeks when TNO_3 observations are available at all 33 sites. Using the simplified RDSTM, we can predict $\ln(TNO_3)$ at any time and location when observations of the 5 covariates are available. Evidence of this advantage is seen in Figure 5, in which the right-hand plot contains 50% more data points than the left plot.

4.2.2 Covariate Contributions

Having established confidence in $\ln(TNO_3)$ predicted from the simplified RDSTM, we may proceed with a physical and chemical interpretation of the model results. As discussed earlier, our objective is to use the model outputs to infer the relative importance of each TNO_3 formation and loss pathway as a function of time and location. A convenient way to compute

the relative contributions from each covariate is

$$\text{Contrib}_k = \frac{\beta_{kt} X_{itk}}{\sum_{k=1}^{p-1} |\beta_{kt} X_{itk}| + |\nu_{it}|}, \quad (8)$$

following the notation from equation 7. By definition, the contributions calculated in this manner range between -1 and +1. The site-specific intercept term, α_i, is purposefully omitted from equation 8 to allow comparisons of the covariate contributions across all sites. Results are summarized by month in Figure 6. Comparison across months is facilitated by the fact that the denominator of equation 8 exhibits no discernable seasonal cycle.

Figure 6a indicates that ResidNH$_4$ makes a positive contribution to TNO$_3$ during the winter months and has little or no impact on TNO$_3$ during the summer. These results are supported by the following explanation. It can be seen from the definition of ResidNH$_4$ (NH_4^+ minus $2SO_4^{2-}$) that the numerical value of this covariate can be zero, positive, or negative. When the particle-phase NH_4^+ is exactly enough to neutralize all of the SO_4^{2-}, ResidNH$_4$ is zero. If there is insufficient NH_4^+ to neutralize all of the SO_4^{2-}, ResidNH$_4$ is negative and the particles are acidic. In the eastern U.S., this often occurs during summer when SO_4^{2-} levels are high (Ferek et al., 1983). During those months, we expect most of the TNO$_3$ to remain in the gas phase and, therefore, be insensitive to the magnitude of ResidNH$_4$. The RDSTM results are in agreement with this expectation, showing little or no contribution from the ResidNH$_4$ covariate between April and October (see Figure 6a). When the ambient NH_4^+ concentration is in excess of the amount required to neutralize SO_4^{2-}, ResidNH$_4$ is positive. In the eastern U.S., positive values of ResidNH$_4$ are most often observed during the winter when SO_4^{2-} concentrations are lowest. Under these conditions, TNO$_3$ partitions favorably to the particle phase and the dry deposition rate of TNO$_3$ decreases substantially (Dennis et al., 2008). Thus, we expect the ambient TNO$_3$ concentrations to be enhanced during the winter months due to slower deposition. Again, the RDSTM results are in agreement with this expectation showing positive contributions from the ResidNH$_4$ covariate between November and March.

The contribution of the O$_3$ covariate to TNO$_3$ is positive during most months, with the largest relative contributions in June and August (see Figure 6b). The long summer days coupled with high mid-day solar elevation angles provide ample solar radiant energy for the photolysis of NO$_2$, which is necessary for the formation of O$_3$. A number of other reasons also contribute to the seasonal cycle of O$_3$ concentrations in the eastern U.S., where peak values are observed every summer. The eventual photolysis of O$_3$ leads to

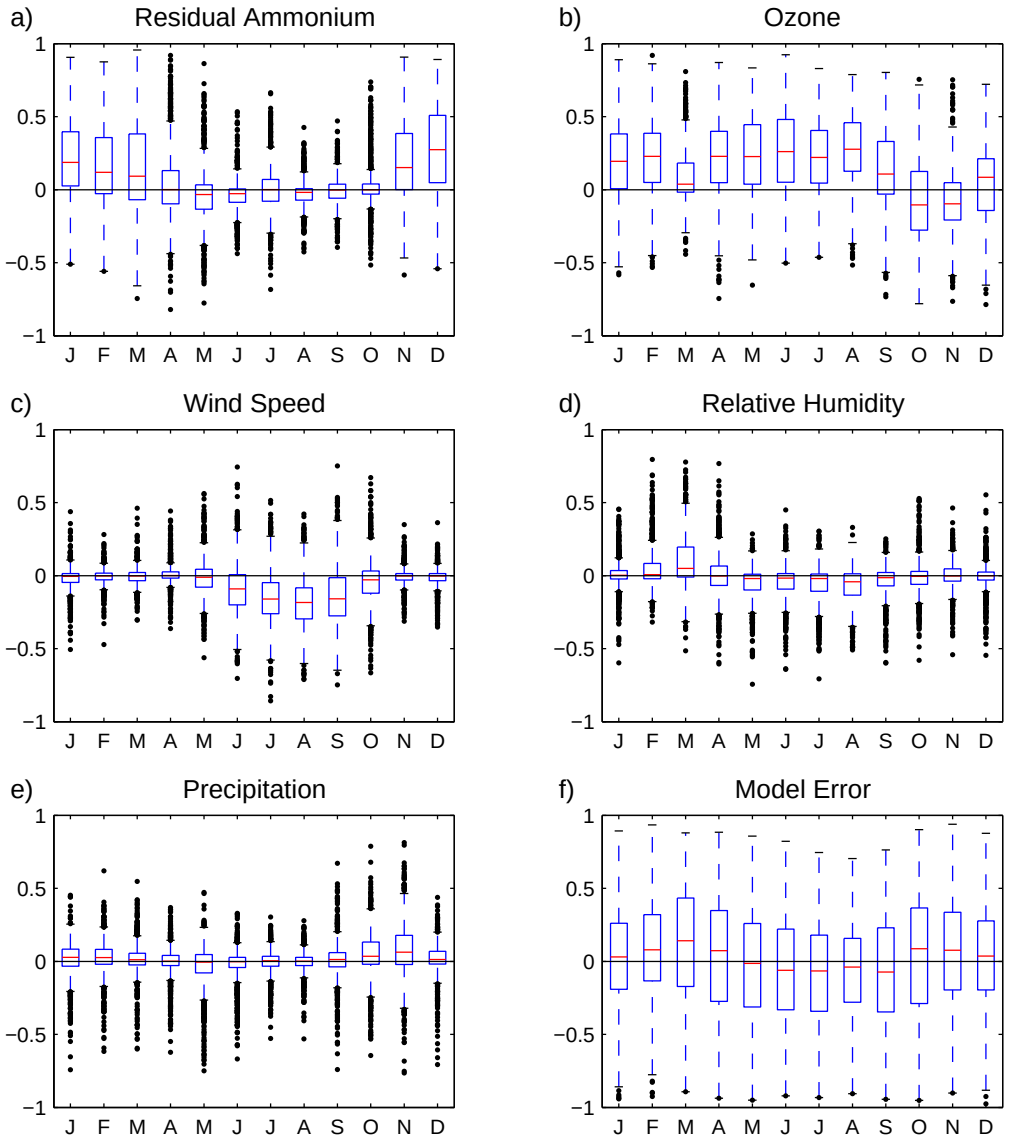

Figure 6: Relative contribution of each covariate and the model error to $\ln(TNO_3)$, as defined in equations 8 and 9

the formation of OH, which is critical to the daytime oxidation of NO_2 that produces TNO_3 (see R1). Given the dearth of routine OH measurements, O_3 is used in the present study as a surrogate for daytime TNO_3 *production*. It is therefore encouraging to see positive contributions from the O_3 covariate during most months. The few negative contributions during October and November result from positive β values during those months multiplied with negative values of the O_3 covariate. The negative covariate values result from the standard normalization of RDSTM inputs discussed in Section 2.2. It is surprising that the positive contributions from O_3 in January and February are nearly as large as the summertime O_3 contributions and are comparable in magnitude to the $ResidNH_4$ contributions during the same months. The high wintertime O_3 contributions are explored on a site-by-site basis in the following section.

Simplified RDSTM results for the meteorological covariates are also quite interesting. As noted in Section 2.3, WS impacts the dry deposition velocities of HNO_3 and NO_3^-. The WS contribution is negligible from November through April, whereas a clear negative contribution is seen between June and September (see Figure 6c). During summer, TNO_3 is primarily in the form of gas-phase HNO_3. The deposition velocity of HNO_3 is high and very sensitive to wind speed. In contrast, TNO_3 during winter is mostly in the particle phase. Particulate nitrate has a relatively low deposition velocity so one would not expect wintertime $\ln(TNO_3)$ concentrations to be sensitive to the small, WS-dependent changes in the NO_3^- deposition rate.

The low covariate contributions from RH are a bit surprising. It is well documented that partitioning of TNO_3 to the particle phase is favored under high RH conditions (Stelson and Seinfeld, 1982), which in turn should decrease the TNO_3 deposition rate. Moreover at night, high RH is expected to promote TNO_3 formation via N_2O_5 hydrolysis in the gas phase (Mentel et al., 1996) and on particle surfaces (Kane et al., 2001). For these reasons, one might expect a positive contribution from the RH covariate during all months. However, recent field measurements by Brown et al. (2006) imply that the nighttime formation rate of $\ln(TNO_3)$ is highly variable and may depend on a number of factors in addition to RH. Our result, showing a lack of $\ln(TNO_3)$ sensitivity to ambient RH (see Figure 6d), is in line with that recent finding and further indicates that RH is not as influential as $ResidNH_4$ in dictating the gas/particle partitioning behavior of TNO_3.

Precipitation is seen to have a small effect on $\ln(TNO_3)$ when averaged by month across all sites. The large number of data points falling below the box and whiskers during all months (see Figure 6e) is indicative of the fact that precipitation is temporally variable and, during many individual time

periods, this covariate has a strong negative effect on $\ln(TNO_3)$. This result is anticipated because precipitation serves as our surrogate for wet removal of TNO_3. Furthermore, the simplified RDSTM yields negative values of β for the precipitation covariate during all months. This means that if precipitation were to increase during any month while all other covariates are held constant, $\ln(TNO_3)$ would decrease. This result is intuitive and supports the credibility of the RDSTM for estimating the relative importance of different TNO_3 sources and sinks.

Figure 6f also shows the relative contribution of model error, calculated as

$$\text{Contrib}_{err} = \frac{\nu_{it}}{\sum_{k=1}^{p-1} |\beta_{kt} X_{itk}| + |\nu_{it}|}. \quad (9)$$

During each month, the contribution of model error has a median value close to zero and its distribution is centered evenly about zero. This illustrates that the simplified RDSTM results are temporally unbiased.

Finally, it is worthwhile to highlight a couple of the unique results shown in Figure 6 that could not be obtained using a standard LRM in which the regression coefficients are fixed in time. For example, a typical LRM would likely yield positive contributions from the $ResidNH_4$ covariate during winter and negative contributions during summer due to the seasonal sign change in the $ResidNH_4$ concentrations discussed above. In contrast, the RDSTM produces a more physically meaningful result showing essentially no effect of $ResidNH_4$ during summer. Also, a standard LRM would probably show lower contributions (i.e., more negative) from WS during winter than summer due to the larger magnitude of surface wind speeds in winter. The RDSTM produces a very different result showing no sensitivity to WS during winter.

4.2.3 Site-Specific Analyses

Figures 7 and 8 illustrate the time series of RDSTM results at two sites. The format of these plots is analogous to that of Figure 4, where the solid line represents the intercept term, α_i, and the colored patches represent the absolute contributions from different covariates, $\beta_{kt} X_{itk}$. For visualization purposes, model results are averaged into 4-week intervals. Results at the Ann Arbor, MI site (Figure 7) are similar to those at most other midwestern locations. The two most important covariates are $ResidNH_4$ and O_3, and their contributions dominate during winter and summer, respectively. Wind speed makes a small negative contribution each summer. Results at

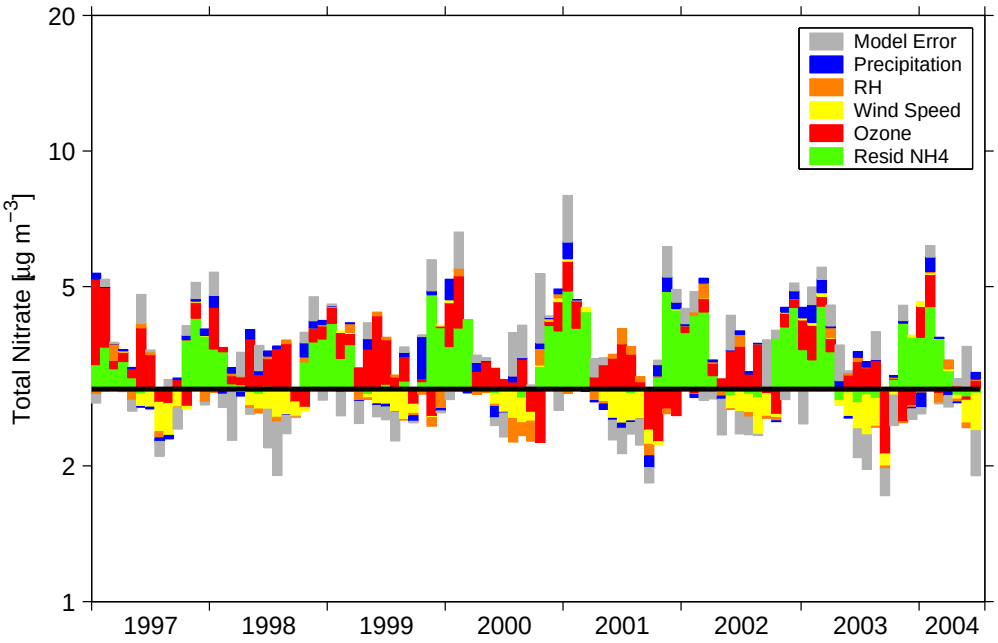

Figure 7: Covariate contributions to total nitrate at the Ann Arbor, MI site using the simplified RDSTM

the Georgia Station, GA site (Figure 8) are representative of those at other southeastern CASTNet locations. At these sites, the O_3 covariate makes a substantial positive contribution throughout the year while the contributions from ResidNH_4 and WS are smaller in magnitude and their sign is more variable. The high wintertime O_3 contributions discussed above in association with Figure 6b are in fact driven by results at the southeastern sites. In the winter, day length and the elevation angle of the noon sun increase as one moves south. As a result, climatological maps show an increase in ground level solar radiation incident on a horizontal surface. It follows that O_3 levels and, hence, the daytime production of TNO_3 should be highest at the southern sites. This is what we find in the RDSTM results (compare red patches in Figures 7 and 8). Low ResidNH_4 contributions are due to lower ammonia emissions in central Georgia as compared to the Midwest (Gilliland et al., 2006).

In order to further test the validity of our simplified RDSTM, we used out-of-sample spatial and temporal cross-validation measures. For example,

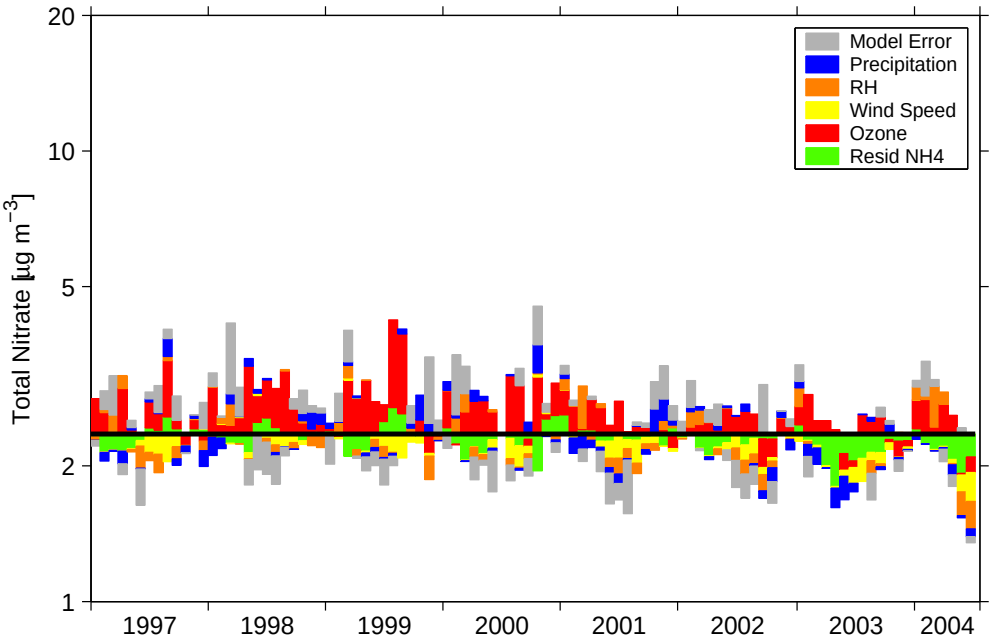

Figure 8: Covariate contributions to total nitrate at the Georgia Station, GA site using the simplified RDSTM

we can withhold observations from a selected number of sites and weeks and then use the remaining observations to fit the simplified RDSTM and obtain the posterior predictive distributions of the withheld observations. Upon finding that the spatial (conditional) correlations are rather weak in our data set (see Section 4.1.1), we used only temporal cross-validation to test the goodness of fit by withholding data for the most recent observed year (i.e., 2004) at some selected sites (CKT130, OXF122 and ANA115). Our simplified RDSTM model performed very well by capturing an overall 95.7% of the withheld observations by using 95% posterior predictive intervals across these three selected sites for the year 2004. For sites CKT130, OXF122 and ANA115, the 95% posterior predictive intervals based on the simplified RDSTM captured about 96%, 93% and 98% of the withheld observations, respectively. These additional results indicate that the simplified RDSTM model is well calibrated and can possibly be used for near-future predictions of TNO_3 concentrations.

5 Discussion and Future Research

In this study, the statistical model developed by Lee and Ghosh (2008) has been refined and simplified to increase its utility for atmospheric scientists and air quality managers. A quantitative understanding of the sources and sinks of TNO_3 is extremely important in the United States, where multi-million dollar policy decisions are made based on the knowledge of such physical and chemical processes (NRC, 2004). Rather than drawing attention to the values of the regression coefficients, which is a focal point in many statistical analyses, we focus on the absolute contribution of each covariate to $\ln(TNO_3)$ (i.e., $\beta_{kt} \times X_{itk}$). In doing so, a few challenges have been encountered which may motivate future research. Transformation of the response variable from TNO_3 to $\ln(TNO_3)$ is warranted for numerical reasons (see Section 2.2), but it hinders the ability to compute covariate contributions in the native units of TNO_3 ($\mu mol/m^3$) which are desired by atmospheric scientists. In addition, the (linear) standardization of covariates to have an empirical mean of 0 and a variance of 1 facilitates data imputation and the fitting of extreme values in the observational data set, but the resultant sign of $\beta_{kt} \times X_{itk}$ is physically counterintuitive in some cases. For example, precipitation is found to make a positive contribution to $\ln(TNO_3)$ at many times and locations (see Figure 6e) in spite of the fact that the β values for precipitation are most often negative. Future efforts directed at the above issues could further increase the appeal of our model results to air quality managers.

In Section 4.2, statistical metrics were used to establish confidence in the simplified RDSTM for reproducing the observed $\ln(TNO_3)$ concentrations. More importantly from an air quality management perspective, some confidence is established in the seasonal and spatial patterns of each covariate's contribution to $\ln(TNO_3)$ using a variety of physical and chemical explanations. For example, the RDSTM helped confirm results from numerical air quality models which suggest that TNO_3 concentrations are dominated by daytime production during summer months and by gas/particle partitioning during winter months. The strong negative influence of WS on summertime TNO_3 concentrations confirms the importance of HNO_3 deposition, and suggests that errors in the performance of numerical models during summer months may be tied to dry deposition velocity formulas. Finally, the lack of dependence on RH provides useful input in a current controversy over whether or not the nighttime hydrolysis of N_2O_5 exhibits a RH dependence (Davis et al., 2008). In general, the RDSTM results indicate that the ResidNH_4, O_3, and WS covariates have the greatest impact

on TNO_3 concentrations. The monthly contributions of these three covariates to $\ln(TNO_3)$ match qualitatively with expectations based on the known production and loss pathways. The RH and P covariates have a smaller net effect on ambient TNO_3, which is also an informative result.

Though we have focused largely on qualitative aspects of the model results, we must not overlook that the RDSTM provides a quantitative and robust empirical relationship between atmospheric TNO_3 concentrations and a set of observable covariates. In the future, this quantitative relationship may prove useful in improving our mechanistic understanding of TNO_3 formation and loss processes. For example, the simplified RDSTM described in this study can be applied to the outputs of a numerical air quality model simulation to determine the time-varying relationship between the simulated TNO_3 concentrations and the numerically simulated values of ResidNH_4, O_3, WS, RH and P. Then, that quantitative relationship can be compared with the empirical relationship derived in the present study to assess whether the numerical models are capturing the correct relationships between TNO_3 and the selected surrogate variables across time and space. Such a comparison should provide unique quantitative insights that may lead to an improved ability to simulate the atmospheric concentrations of TNO_3.

Acknowledgements

We thank Steven Howard at EPA for processing and formatting the CASTNet data for input to the RDSTM, and Kristen Foley, Jon Pleim, and Jenise Swall at EPA for helpful comments on this paper. We thank the Editor, the Associate Editor and the three Referees for their insightful comments and helpful suggestions on this manuscript. We believe our paper has been substantially improved as a result. The United States Environmental Protection Agency through its Office of Research and Development partially funded and collaborated in the research described here. It has been subjected to Agency review and approved for publication.

References

Alexander, B.; Hastings, M.G.; Allman, D.J.; Dachs, J.; Thornton, J.A.; Kunasek, S.A. (2009). Quantifying atmospheric nitrate formation pathways based on a global model of the oxygen isotopic composition ($\Delta^{17}O$) of atmospheric nitrate, Atmospheric Chemistry and Physics, 9, 5043-5056.

Ansari, A. S.; Pandis, S. N. (1999). An analysis of four models predicting the partitioning of semivolatile inorganic aerosol components. *Aerosol Sci. Technol.*, **31**, 129-153.

Appel, K.W.; Bhave, P.V.; Gilliland, A.B.; Sarwar, G.; Roselle, S.J. (2008). Evaluation of the Community Multiscale Air Quality (CMAQ) Model Version 4.5: Sensitivities Impacting Model Performance; Part II Particulate Matter, *Atmospheric Environment*, **42**, 6057-6066.

Banerjee, S., Carlin, B. P., and Gelfand, A. E.(2004). *Hierarchical modeling and analysis for spatial data*, Chapman & Hall/CRC.

Box, G., Gwilym M. Jenkins, G. M. and Reinsel, G. (1994). *Time Series Analysis: Forecasting & Control*, Prentice Hall, New York.

Brown, P. E., Karesen, K. F., Roberts, G. O., and Tonellato S. (2000). Blur-generated non-separable space-time models. *Journal of the Royal Statistical Society, Series B*, **62**, 847-860.

Brown, S. S.; Ryerson, T. B.; Wollny, A. G.; Brock, C. A.; Peltier, R.; Sullivan, A. P.; Weber, R. J.; Dube, W. P.; Trainer, M.; Meagher, J. F.; Fehsenfeld, F. C.; Ravishankara, A. R. (2006). Variability in nocturnal nitrogen oxide processing and its role in regional air quality. *Science*, **311**, 67-70.

Calder, C.A. (2007). Dynamic Factor Process Convolution Models for Multivariate Space-Time Data with Application to Air Quality Assessment. *Environmental and Ecological Statistics*, **14**, 229-247.

Carroll, R., Chen, R., George, E., Li, T., Newton, H., Schmiediche, H., and Wang, N. (1997). Ozone exposure and population density in harris county, Texas. *Journal of the American Statistical Association*, **92**, 392-415.

Cressie, N. A. C. (1993). *Statistics for spatial data*. Wiley: New York, revised edition.

Cressie, N. A. C. and Huang, H.-C. (1999). Classes of nonseparable, spatiotemporal stationary covariance functions. *Journal of the American Statistical Association*, **94**, 1330-1340.

Davis, J.M.; Bhave, P.V.; Foley, K.M. (2008). Parameterization of N_2O_5 Reaction Probabilities on the Surface of Particles Containing Ammo-

nium, Sulfate, and Nitrate, *Atmospheric Chemistry and Physics*, **8**, 5295-5311.

Dennis, R. L.; Bhave, P. V.; Pinder, R. W. (2008). Observable indicators of the sensitivity of $PM_{2.5}$ nitrate to emission reductions - Part II: Sensitivity to errors in total ammonia and total nitrate of the CMAQ-predicted non-linear effect of SO_2 emission reductions. *Atmospheric Environment*, **42**, 1287-1300.

Eder, B. and Yu, S. (2006). A performance evaluation of the 2004 release of Models-3 CMAQ. *Atmospheric Environment*, **40**, 4811-4824.

Ferek, R. J.; Lazrus, A. L.; Haagenson, P. L.; Winchester, J. W. (1983). Strong and weak acidity of aerosols collected over the northeastern United States. *Environ. Sci. Technol.*, 17, 315-324.

Gelfand, A. E., Banerjee, S. and Gamerman, D. (2005). Spatial processes modelling for univariate and multivariate dynamic spatial data. *Environmetrics*, **16**, 465479.

Gelfand, A. E., Ghosh, S. K., Knight, J. R., and Sirmans, C. F. (1998). Spatio-Temporal Modeling of Residential Sales Data. *Journal of Business & Economic Statistics*, **16**, 312-321.

Gilliland, A. B.; Appel, K. W.; Pinder, R. W.; Dennis, R. L. (2006). Seasonal NH3 emissions for the continental united states: Inverse model estimation and evaluation. *Atmos. Environ.*, **40**, 4986-4998.

Gipson, G. L. (1999). Process Analysis. Chapter 16 in Science Algorithms of the EPA Models-3 Community Multiscale Air Quality (CMAQ) Modeling System (D. W. Byun and J. K. S. Ching, eds.), EPA/600/R-99/030, Research Triangle Park, NC.

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association*, **97**, 590-600.

Golub, G. H. and Van Loan, C. F. (1996). *Matrix Computations, 3rd ed.* Baltimore, MD: Johns Hopkins.

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

Handcock, M. S. and Wallis, J. R. (1994). An approach to statistical spatial temporal modeling of meteorological fields (with discussion), *Journal of the American Statistical Association*, **89**, 368-390.

Huerta, G., Sanso, B., Stroud, J. (2004). A Spatiotemporal Model for Mexico City Ozone Levels. *Journal of the Royal Statistical Soceity, Series C -Applied Statistics*, **53**, 231-248.

Kane, S. M.; Caloz, F.; Leu, M. T. (2001). Heterogeneous uptake of gaseous N_2O_5 by $(NH_4)(2)SO4$, NH_4HSO_4, and H_2SO_4 aerosols. *Journal of Physical Chemistry A*, **105**, 6465-6470.

Kent, J. T. and Mardia, K. V. (2002). Modelling strategies for spatial-temporal data. In Spatial Cluster Modelling (eds. A. Lawson and D. Denison), 214-226, London, Chapman and Hall.

Kyriakidis, P. C. and Journel, A. G. (1999). Geostatistical space-time models: a review. *Mathematical Geology*, **31**(6), 651-684.

Lee, H. (2006). *Reparametrized Dynamic Space-Time Models and Spatial Model Selection*, North Carlona State University (unpublished dissertation)
`http://www.lib.ncsu.edu/theses/available/etd-04282006-115620/`

Lee, H., and Ghosh, S. K. (2008). A Reparametrization Approach for Dynamic Space-Time Models, *Journal of Statistical Theory and Practice*, **2**, 1-14.

Malm, W. C.; Schichtel, B. A.; Pitchford, M. L.; Ashbaugh, L. L.; Eldred, R. A. (2004). Spatial and monthly trends in speciated fine particle concentration in the United States. *J. Geophys. Res.*, **109**, D03306.

Miller, A. (2002). *Subset Selection in Regression*, CRC Press, New York.

Mardia, K. V. and Goodall, C. R. (1993). Spatial temporal analysis of multivariate environmental monitoring data. In Multivariate Environmental Statistics (G. P. Patil and C. R. Rao, eds.) 347-386. North-Holland, Amsterdam.

Mardia, K. V., Goodall, C., Redfern, E. J., and Alonso, F. J. (1998). The kriged Kalman filter (with discussion). *Test*, **7**, 217-285.

Mentel, T. F.; Bleilebens, D.; Wahner, A. (1996). A study of nighttime nitrogen oxide oxidation in a large reaction chamber - The fate of NO_2

N_2O_5, HNO_3, and O_3 at different humidities. *Atmos. Environ.*, **30**, 4007-4020.

Nagar, D. K. and Gupta, A. K. (2000). *Matrix Variate Distributions*, CRC Press, New York.

National Research Council (Committee on Air Quality Management in the United States) (2004). Air Quality Management in the United States. National Academies Press, Washington, D.C.

Sahu, S. and Mardia, K. V. (2005). A Bayesian kriged Kalman model for short-term forecasting of air pollution levels. *Applied Statistics*, **54**, 223-244.

Sampson, P. D. and Guttrop, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure, *Journal of the American Statistical Association*, **87**, 108-119.

Schmidt, A. M. and O'Hagan, A. (2003). Bayesian inference for nonstationary spatial covariance structures via spatial deformations, *Journal of the Royal Statistical Society, Series B*, **65**, 743-758,

Shumway, R.H. and Stoffer, D. S. (2000). *Time Series Analysis and Its Applications*, Springer, New York.

Steel, R.G.D., Torrie, J.H., and Dickey, D.A. (1997). *Principles and Procedures of Statistics: A Biometric Approach.* New York, McGraw-Hill.

Stelson, A. W.; Seinfeld, J. H. (1982). Relative humidity and temperature dependence of the ammonium nitrate dissociation constant. *Atmos. Environ.*, **16**, 983-992.

Stroud, J. R., Müller, P., and Sanso, B. (2001). Dynamic models for spatiotemporal data. *Journal of Royal Statistical Society, Series B*, **63**, 673-689.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statististical Soceity, series B*, **58**, 267-288.

West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models*. New York: Springer.

Waller, L., Carlin, B. P., Xia, H. and Gelfand, A. E. (1997). Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association*, **92**, 607-617.

Wikle, C. and Cressie, N. (1999). A dimension reduced approach to space-time kalman filtering. *Biometrika*, **86**, 815-829.

Wikle, C., Berliner, M., and Cressie, N. (1999). Hierarchical Bayesian space-time models. *Environmental and Ecological Statistics*, **5**, 117-154.

Yu, S.; Dennis, R.; Roselle, S.; Nenes, A.; Walker, J.; Eder, B.; Schere, K.; Swall, J.; Robarge, W. (2005). An assessment of the ability of three-dimensional air quality models with current thermodynamic equilibrium models to predict aerosol NO_3^-. *J. Geophys. Res.*, **110**, D07S13.

Zhang, Y.; Seigneur, C.; Seinfeld, J. H.; Jacobson, M.; Clegg, S. L.; Binkowski, F. S. (2000). A comparative review of inorganic aerosol thermodynamic equilibrium modules: similarities, differences, and their likely causes. *Atmos. Environ.*, **34**, 117-137.

Zheng, J., Swall, J.L., Cox, W. M. and Davis, J.M. (2007). Interannual variation in meteorologically adjusted ozone levels in the eastern United States: A comparison of tow approaches. *Atmospheric Environment*, **41**, 705-716.

Figure 9: CASNET sites that were used to obtain data for this study

Appendix

The following figure shows the location of the sites that were used in our analyais.