Published in final edited form as:

J Stat Theory Pract. 2008 March 1; 2(1): 1–14.

A Reparametrization Approach for Dynamic Space-Time Models

Hyeyoung Lee^{*} and Sujit K. Ghosh[†]

Abstract

Researchers in diverse areas such as environmental and health sciences are increasingly working with data collected across space and time. The space-time processes that are generally used in practice are often complicated in the sense that the auto-dependence structure across space and time is non-trivial, often non-separable and non-stationary in space and time. Moreover, the dimension of such data sets across both space and time can be very large leading to computational difficulties due to numerical instabilities. Hence, space-time modeling is a challenging task and in particular parameter estimation based on complex models can be problematic due to the curse of dimensionality. We propose a novel reparametrization approach to fit dynamic space-time models which allows the use of a very general form for the spatial covariance function. Our modeling contribution is to present an unconstrained reparametrization method for a covariance function within dynamic space-time models. A major benefit of the proposed unconstrained reparametrization method is that we are able to implement the modeling of a very high dimensional covariance matrix that automatically maintains the positive definiteness constraint. We demonstrate the applicability of our proposed reparametrized dynamic space-time models for a large data set of total nitrate concentrations.

Keywords

Computational efficiency; Dynamic models; Reparametrization; Spatial models

1 Introduction

Researchers in diverse areas such as environmental and health sciences are increasingly faced with working with data that are observed over space and time. In recent years, there has been widespread attention in the statistical literature given to modeling space-time data (Gelfand et al., 1998; Kyriakidis and Journel, 1999; Cressie and Huang, 1999;; Gneiting, 2002; Stein, 2003; Huang and Hsu, 2004).

Dynamic linear models (DLM), or state-space models have been widely used to analyze multidimensional time series such as spatial time series. Many researchers have employed a DLM framework in environmental problems, in particular, those data that are observed over a long time series. A Bayesian approach has been commonly taken in the DLM framework for analyzing space-time data. Examples include Tonellato (1997), Sansó and Guenni (1999), Wikle et al. (1999), Stroud et al. (2001), Huerta et al. (2004), and Banerjee et al. (2005) among others. Huang and Cressie (1996) and Mardia et al. (1998) presented a non-Bayesian approach as well.

[†](corresponding author) S. K. Ghosh is a Professor in the Department of Statistics at North Carolina State University, Raleigh, NC, USA. ghosh@stat.ncsu.edu.

^{*}H. Lee is an employee of the Korea Institute of Patent Information, Seoul, Korea. hylee@kipi.or.kr

In this article, we take a Bayesian approach and use Markov chain Monte Carlo (MCMC) simulation to obtain the inference based on the posterior distribution of the parameters. Recent developments in MCMC computing allow fully Bayesian analyses of complex multilevel models for dynamic space-time data. Pole et al. (1994) and West and Harrison (1997) are good references on the DLM from a Bayesian point of view.

In general, space-time modeling is a challenging task which requires the manipulation of large data sets and considers both spatial and temporal correlations. Moreover, space-time processes are often complicated in that the dependence structure across space and time is non-trivial, often non-separable and non-stationary in space and/or time. Several modeling strategies have been proposed to address this problem. Wikle and Cressie (1999) developed a space-time Kalman-filter that achieves dimension reduction by decomposing the state-process into sets of basis functions and time series. To avoid a high computational load, they used an empirical Bayesian method for the estimation of model parameters rather than a fully Bayesian hierarchical approach. However, the additional variability in estimating parameters is ignored in such a method. Stroud et al. (2001) specified very simple random walk dynamics, while Xu and Wikle (2005) discussed efficient estimation approaches via the expectation-maximization (EM) algorithm for the parameter and covariance matrices with high dimensionality in dynamic space-time models. Thus, it appears that most of the methods available currently in the literature are limited either by restrictive assumptions on the covariance functions (e.g., stationary, isotropic, separable etc.) or if such restrictive assumptions are not made the methods suffer from curse of dimensionality due high dimensional matrix inversions. We propose a method that avoids such curse of dimensionality without making strong restrictive assumptions.

This article is organized as follows. First, we present a brief overview of the dynamic space-time models in Section 2.1 and related inference based on standard methods in Section 2.2. In Section 2.3, we propose a reparametrization approach to fit dynamic space-time models that is based on a very general class of spatial covariance function. These models are not limited by the stationarity and isotropy restrictions for the covariance functions. One of the main contributions of this article is to present an unconstrained reparametrization method to be used for a covariance function within the dynamic space-time modeling framework. Using this unconstrained reparametrization method, we are able to implement the models to fit high-dimensional data without many restrictive assumptions. In Section 3 we demonstrate the applicability of our proposed reparametrized dynamic space-time models for a large data set of total nitrate concentrations. Section 4 presents a general discussion and scope for future research.

2 Dynamic Space-Time Models (DSTM)

Dynamic linear model (West and Harrison, 1997) is probably one of the most widely known and used subclass in dynamic models. The term *dynamic* is related to changes in time series processes due to the passage of time. DLM can be seen as a generalization of traditional regression models that allows changes in parameter values over time. DLM provide a very flexible framework that permits smooth and abrupt changes in the time series generating process. We adapt a DLM framework to space-time data and call the model a dynamic space-time model (DSTM). A DSTM framework views the data as arising from spatial time series.

2.1 Overview

Suppose the process $\{Z(s, t); s \in \mathbb{R}^2, t \in [0, \infty)\}$ is observed on a finite number of sites labeled as $s_1, \dots, s_n$ at each time t , where $t = t_1, t_2, \dots, t_m$. Assuming that the t_j 's are observed over equal lags, we will relabel $t_j = j$ for $j = 1, \dots, m$. Consider the $n \times 1$ vector time

series $\mathbf{Z}_t = (Z(s_1, t), \dots, Z(s_n, t))^T$ at time t . For each t , DSTM is usually formed by an observation equation and an evolution equation. An observation equation describes the relationship between the observation ($\mathbf{Z}_t$) and the regressors (X_t) that takes the form of a multivariate regression process,

$$\mathbf{Z}_t = X_t \boldsymbol{\beta}_t + \mathbf{v}_t, \quad \mathbf{v}_t \sim N_n(\mathbf{0}, \sum_t^v) \quad (2.1)$$

where X_t is an $n \times p$ observed design matrix 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,

$$\boldsymbol{\beta}_t = G_t \boldsymbol{\beta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim N_p(\mathbf{0}, \sum_t^\omega) \quad (2.2)$$

where G_t is a $p \times p$ evolution matrix. There are several ways to model the G_t 's. The most common assumption is that the G_t 's are structurally known, possibly up to some finite number of parameters. In this article, we do not make any structural assumption about the G_t 's, but we assume that $G_t = G$ for all t and that G follows a matrix-valued normal

distribution with mean G_0 and variance-covariance parameters Ω_0 and $\sum_0^G$. That is, $G \sim MN_{p \times p}(G_0, \Omega_0, \sum_0^G)$. We also assume that the $\mathbf{v}_t$ and $\boldsymbol{\omega}_t$ error vectors are independent and have multivariate normal distributions with mean $\mathbf{0}$ and variance-covariance matrices $\sum_t^v$ and $\sum_t^\omega$, respectively. The model is completed with a normal prior for the initial state, $\boldsymbol{\beta}_1 \sim N(\boldsymbol{\beta}_0, \sum_0^\omega)$, where $\boldsymbol{\beta}_0$ is assumed known.

Equivalently, the model can be written using hierarchical specifications as follows:

$$\begin{aligned} \mathbf{Z}_t | \boldsymbol{\beta}_t &\sim N_n(X_t \boldsymbol{\beta}_t, \sum_t^v), \\ \boldsymbol{\beta}_t | \boldsymbol{\beta}_{t-1}, G &\sim N_p(G \boldsymbol{\beta}_{t-1}, \sum_t^\omega), \\ G &\sim MN_{p \times p}(G_0, \Omega_0, \sum_0^G), \\ \boldsymbol{\beta}_1 &\sim N(\boldsymbol{\beta}_0, \sum_0^\omega), \end{aligned}$$

where the matrix valued normal distribution $MN_{p \times p}(G_0, \Omega_0, \sum_0^G)$ has the probability density function given by

$$p(G) = (2\pi)^{-p^2/2} |\Omega_0|^{-p/2} \left| \sum_0^G \right|^{-p/2} \exp \left(-\frac{1}{2} \text{tr} \left[\Omega_0^{-1} (G - G_0) \sum_0^{G^{-1}} (G - G_0)^T \right] \right).$$

This Bayesian hierarchical approach not only helps organize our thinking about the model, but also fully accounts for all sources of uncertainty without making substantial structural assumptions such as spatial stationarity, isotropy, etc.

In order to keep our illustrations simple, first we consider the following simplified DSTM.

Specifically, first we assume that $\sum_t^v = \sum^v$ and $\sum_t^\omega = \sum^\omega$ in equations (2.1) and (2.2). That

is, the conditional variance-covariance matrices Σ^v and Σ^ω do not change over time (and hence are static). Later we discuss how to relax these assumptions. However notice that the marginal variances, $\text{Var}[\mathbf{Z}_t]$'s do change over time. Specifically,

$$\begin{aligned}\text{Var}[\mathbf{Z}_t] &= \mathbf{X}_t \text{Var}[\boldsymbol{\beta}_t] \mathbf{X}_t^T + \Sigma^v \\ \text{Var}[\boldsymbol{\beta}_t] &= \mathbf{G} \text{Var}[\boldsymbol{\beta}_{t-1}] \mathbf{G}^T + \Sigma^\omega.\end{aligned}$$

Using the above recursive relations we can easily see that $\text{Var}[\mathbf{Z}_t]$ does evolve over time even when Σ^v is assumed to be constant over time. Then our simplified DSTM can be written as,

$$\mathbf{Z}_t = \mathbf{X}_t \boldsymbol{\beta}_t + \mathbf{v}_t, \quad \mathbf{v}_t \sim N_n(\mathbf{0}, \Sigma^v), \quad (2.3)$$

$$\boldsymbol{\beta}_t = \mathbf{G} \boldsymbol{\beta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim N_p(\mathbf{0}, \Sigma^\omega), \quad (2.4)$$

where $\boldsymbol{\beta}_1 \sim N(\boldsymbol{\beta}_0, \Sigma^\omega)$. Here, Σ^v and Σ^ω are unstructured variance-covariance matrices, and for the ease of exposition, we initially assume that $\mathbf{Z}_t$ and $\mathbf{X}_t$ are observed at all time points t . Later we discuss how to relax this assumption if some observations are missing.

The Bayesian model is completed with the specification of a prior distribution for the parameters. These include the data-model covariance matrix Σ^v and the error covariance matrix Σ^ω . Prior specifications for Σ^v and Σ^ω can be tricky as these matrices are usually high-dimensional (e.g., Σ^v is $n \times n$), and they need to be positive definite (pd). It is customary to use the inverse Wishart distributions to model such covariance matrices. Note that our models require no restrictive assumptions such as stationarity and isotropy as Σ^v is left unstructured. However, if such assumptions are deemed necessary, we can easily incorporate them in our modeling framework.

2.2 Updating Equations

Suppose that the time at origin $t=0$ represents the current time, and that the existing information available is denoted by the D_0 , the initial information set. Generally, we have the information set at time t , D_t . As time evolves, so does the information. Observing the values of $\mathbf{Z}_t$ at time t implies that D_t includes both the previous information set D_{t-1} and the observations $\mathbf{Z}_t$, that is, $D_t = \{\mathbf{Z}_t, D_{t-1}\}$.

Traditionally, inference for dynamic models is made sequentially by obtaining the prior predictive and updated distributions for the state parameters $\boldsymbol{\beta}_t$ for each time t . The prior predictive distributions are respectively obtained by

$$\begin{aligned}p(\boldsymbol{\beta}_t | D_{t-1}) &= \int p(\boldsymbol{\beta}_t | \boldsymbol{\beta}_{t-1}) p(\boldsymbol{\beta}_{t-1} | D_{t-1}) d\boldsymbol{\beta}_{t-1} \\ p(\mathbf{Z}_t | D_{t-1}) &= \int p(\mathbf{Z}_t | \boldsymbol{\beta}_t) p(\boldsymbol{\beta}_t | D_{t-1}) d\boldsymbol{\beta}_t\end{aligned}$$

and the updated distribution is obtained by Bayes' theorem as

$$p(\beta_t|D_t) \propto p(\mathbf{Z}_t|\beta_t)p(\beta_t|D_{t-1})$$

where D_t represents all the available information up to time t .

The above updating scheme in the dynamic space-time model may not be easy to implement when the G matrix is completely unknown and the matrices Σ^v and Σ^ω are completely unstructured, even if we are able to derive the analytical form of the full conditional distributions required for the above updating equations. The main limitation of these types of multivariate updating schemes are the problems associated with high dimensional matrix inversions. Moreover, these types of multivariate updating schemes are not generally applicable when some observations are missing or censored, as it is generally very hard to sample from truncated multivariate distributions. In order to avoid such numerical instabilities and to accelerate model fitting, we propose a method via reparametrization of the model which leads to an univariate scheme for the aforementioned DSTM. As by-products of using this reparametrization method, we show that it is possible to obtain several extensions of the usual DSTM (e.g., relaxing the Gaussian assumption, allowing nonstationary and nonseparable covariances etc.).

2.3 A Reparametrization Method

In this section, we describe a reparametrization method for the covariance matrices Σ^v and Σ^ω . Modeling a covariance matrix Σ is difficult because (i) it is a high dimensional parameter and (ii) it is restricted to be positive definite. Pourahmadi (1999) introduced an unconstrained parametrization procedure to model a temporal covariance matrix. Chen and Dunson (2003) and Daniels and Zhao (2003) have also used similar parametrizations to model covariance matrices in longitudinal studies. The Cholesky decomposition of the inverse of a covariance matrix is used to associate a unique unit lower triangular and a unique diagonal matrix with each covariance matrix. The entries of the lower triangular matrix and the log of the diagonal matrix are completely unconstrained and have interpretations similar to regression coefficients and prediction variances, respectively, when regressing a measurement on its predecessors. Using the Cholesky decomposition and the ensuing unconstrained and statistically meaningful reparametrization, Daniels and Pourahmadi (2002) provide a convenient and intuitive framework for developing conditionally conjugate prior distributions for covariance matrices, and show their connections to the generalized inverse Wishart priors. However, to the best of our knowledge this type of reparametrization of covariance matrices has been used only to model temporal processes, taking advantage of the natural ordering of time. We extend these methodologies to spatial and temporal processes and show several extensions. It is well-known that there is no unique ordering on space (of dimension two and higher). For our application we “ordered” the sites, $s_1 < s_2 < \dots < s_n$, in *lexicographical order* using the latitude and longitude of the sites. In other words, $s_1 = (x_1, y_1) < s_2 = (x_2, y_2)$ if and only if $x_1 < x_2$ or $(x_1 = x_2 \text{ and } y_1 < y_2)$, which is in fact a total order. Other ordering such as *product order* (which is a partial order) might also be used.

For our DSTM framework, we define two lower triangular matrices T^v and T^ω and two diagonal matrices D^v and D^ω such that $T^v \Sigma^v T^{vT} = D^v$ and $T^\omega \Sigma^\omega T^{\omega T} = D^\omega$. Such decompositions of positive definite matrices are unique. More precisely, let T^v and T^ω be the lower triangular matrices with 1's as their diagonal entries and $-\phi_{ii'}$, $i > i'$ and $-\psi_{kk'}$, $k > k'$ as their lower triangular entries, respectively. Also, let D^v and D^ω be diagonal matrices with entries $\sigma_1^{v2}, \dots, \sigma_n^{v2}$ and $\sigma_1^{\omega2}, \dots, \sigma_p^{\omega2}$, respectively. We now re-express the equations (2.3) and (2.4) using the entries of their lower triangular and diagonal matrices.

Let $Z_{it} = Z(s_i, t)$, $i = 1, \dots, n$, $t = 1, \dots, m$ and $X_{itk} = X_k(s_i, t)$, $k = 1, \dots, p$. Notice that $\mathbf{Z}_t = (Z_{1t}, \dots, Z_{nt})^T$ and $\{X_t\}_{n \times p} = ((X_{itk}))_{1 \leq i \leq n, 1 \leq k \leq p}$, which appear in (2.3). Then we can represent the DSTM defined by (2.3) and (2.4) as follows. The observation equation can be written as,

$$Z_{it} = \sum_{k=1}^p \beta_{kt} X_{itk} + \sum_{i'=1}^{i-1} \phi_{ii'} Z_{i't} + v_{it}, \quad (2.5)$$

$$Z_{1t} = \sum_{k=1}^p \beta_{kt} X_{1tk} + v_{1t}, \quad (2.6)$$

where $i = 2, \dots, n$, $t = 1, \dots, m$, $E[v_{it}] = 0$, $E[v_{it}^2] = \sigma_i^2$, and $E[v_{it}v_{i't}] = 0$, $i \neq i'$. Notice that $\mathbf{v}_t = (v_{1t}, \dots, v_{nt})^T$ as in (2.3). The evolution equation can now be written as,

$$\beta_{kt} = \sum_{k'=1}^p \beta_{k't-1} g_{kk'} + \sum_{k'=1}^{k-1} \psi_{kk'} \beta_{k't} + \omega_{kt}, \quad (2.7)$$

$$\beta_{1t} = \sum_{k'=1}^p \beta_{k't-1} g_{1k'} + \omega_{1t}, \quad (2.8)$$

where $k = 2, \dots, p$, $t = 2, \dots, m$, $E[\omega_{kt}] = 0$, $E[\omega_{kt}^2] = \sigma_k^2$, and $E[\omega_{kt}\omega_{k't}] = 0$, $k \neq k'$, and initial state equation can be written as,

$$\beta_{k1} = \beta_{k0} + \sum_{k'=1}^{k-1} \psi_{kk'} \beta_{k'1} + \omega_{k1}, \quad (2.9)$$

where $k = 2, \dots, p$. Notice that $\boldsymbol{\omega}_t = (\omega_{1t}, \dots, \omega_{pt})^T$ as in (2.4). The model is completed with

$$\beta_{11} = \beta_{10} + \omega_{11}. \quad (2.10)$$

Here, notice that no structural constraints are required for the elements of T^v , T^ω , D^v , and D^ω (i.e., $\phi_{ii'}, \psi_{kk'} \in \mathbb{R}$ and $\sigma_i^2, \sigma_k^2 \in (0, \infty)$). In particular, for our application we may specify a prior distribution for these parameters as,

$$\begin{aligned}
\phi_{ii'} &\sim N(\phi_0, \sigma_\phi^2), \quad 1 \leq i' < i \leq n, \\
\psi_{kk'} &\sim N(\psi_0, \sigma_\psi^2), \quad 1 \leq k' < k \leq p, \\
\sigma_{v_i}^2 &\sim IG(a_v, b_v), \quad i=1, \dots, n, \\
\sigma_{\omega_k}^2 &\sim IG(a_\omega, b_\omega), \quad k=1, \dots, p, \\
g_{kk'} &\sim N(g_0, \sigma_g^2), \quad k, k' \in 1, \dots, p,
\end{aligned}$$

where $\phi_0, \sigma_\phi^2, \psi_0, \sigma_\psi^2, a_v, b_v, a_\omega, b_\omega, g_0$ and σ_g^2 are all known values that can be used to quantify the prior information if available, otherwise we can use values that would generate a set of vague priors. Here $N(a, b)$ denotes a normal distribution with mean a and variance b , and $IG(a, b)$ denotes an inverse gamma distribution with mean $\frac{b}{a-1}$ for $a > 1$ and variance $\frac{b^2}{(a-1)^2(a-2)}$ for $a > 2$. For instance, we choose these values in such a way that they will have minimal impact on the posterior inference of the parameters. Other prior distributions can also be adapted for our framework very easily.

The advantages of using our proposed reparametrized DSTM (RDSTM) can be summarized as follows:

- i. *Numerical stability*: RDSTM avoid numerical instabilities caused by the multivariate updating scheme for DLM when the dimensions are very large.
- ii. *Routine handling of missing data*: RDSTM allow missing data to be imputed from its full conditional distribution.
- iii. *Implementation using WinBUGS*: RDSTM can be implemented using WinBUGS, while the general DLM cannot.

We can extend the above framework to allow $\phi_{ii'}$ and $\psi_{kk'}$ to vary with time t , thus allowing $\sum_t^v$ to change with time t . Also by letting $\xi_i(t) = \log \sigma_i^{v^2}(t)$ and $\eta_k(t) = \log \sigma_k^{\omega^2}(t)$, we can replace equations (2.5)–(2.10) with time varying coefficients. However for such extensions we need to use autoregressive models for these time varying coefficients. Although these extensions still keep all the updating equations univariate within the MCMC methodology, but it would require much more time to process a single cycle of the Gibbs sampler, as each such cycle will consist of a large number of univariate samples. Nevertheless, the method will not suffer from numerical instabilities and the curse of dimensionality as compared to multivariate updating schemes.

In next section, we demonstrate the applicability of our RDSTM to total nitrate concentration data. We also discuss the practical benefits of the estimates obtained from RDSTM in the context of atmospheric sciences, in addition to computational efficiency.

3 Application to Total Nitrate Concentration Data

We provide a brief background and description of the data before we present the results based on fitting our proposed RDSTM to data on total nitrate concentrations. In this article, we have used the CASNET data set as an illustration to show the applicability of the proposed reparametrization method. More details and related statistical analysis will be presented in a separate article elsewhere.

3.1 Background

The release of different types of hazardous emissions has been instrumental in polluting the atmosphere over the last few decades, and hence atmospheric deposition has become a

major topic of concern in environmental studies. The U.S. Environmental Protection Agency (EPA) established the Clean Air Status and Trends Network (CASTNET) to monitor air pollutant emissions and pollutant deposition (National Research Council, 2004). One of the goals of the scientists working with environmental data is to improve the accuracy of total nitrate predictions and to find the relationship of total nitrate to other predictor variables. To this end, one would like to explore the empirical relationship that may exist between the observed values of total nitrate and other observed variables. To estimate this empirical relationship in the observed data, we employ RDSTM proposed in Section 2. We use total nitrate as the response variable and consider the following variables as explanatory or predictor variables within a regression model: sulfate (SO_4), ammonium (NH_4), ozone (O_3), relative humidity (RH), wind speed (WS), and precipitation (P). We expect these variables to be reasonably good predictors of total nitrate. RDSTM allows us to estimate dynamic relationships that vary with weeks or months between total nitrate and the chemical species and meteorological variables.

3.2 The CASTNET Data

All of the data for this study were obtained from the U.S. EPA CASTNET sites. A complete description of this network can be found at: <http://www.epa.gov/castnet>. Figure 1 shows the locations of the stations used in this study. We used $n = 33$ stations in the eastern U. S. These sites were selected on the basis of the extent of NO_X ($NO_2 + NO$) emissions, where NO represents nitrous oxide. Note that all the CASTNET sites are located in rural locations.

Our data were collected between January, 1997 and July, 2004, and a total of $m = 394$ weeks are available during this period. As we observe a long time series of total nitrate (and other predictor variables) over a fixed domain of monitoring stations it is reasonable to consider DSTM which are ideal for modeling spatial times data. The chemical species used in this study were nitric acid (HNO_3), ($\mu mol/m^3$), nitrate (NO_3^-) ($\mu mol/m^3$), sulfate (SO_4) ($\mu mol/m^3$), ammonium (NH_4) ($\mu mol/m^3$) and ozone (O_3) (ppb). Nitric acid and nitrate were summed to get total nitrate ($\mu mol/m^3$). Residual ammonium was also used, which is calculated as $(NH_4 - 2 \times SO_4)$. In the statistical analysis one could use ammonium, sulfate and residual ammonium. However, since residual ammonium is a known linear combination of the other two, only one of the other two should be used. By using sulfate and residual ammonium, one is prevented from having two ammonium variables in the statistical analysis, which also makes sense thermodynamically.

Meteorological variables are observed on site at each of the CASTNET stations. In this study we used relative humidity (RH) (%), wind speed (WS) (m/s) and precipitation (P) ($mm/week$). All of the meteorological variables were measured hourly. With the exception of ozone, which was measured hourly, the chemical species were averaged over a week from Tuesday to Tuesday. To conform to this weekly pattern, RH and WS variables were averaged over the same period and the daily maximum O_3 values were averaged over each week. Precipitation was summed over the same period. A more detailed study of this data has been conducted by these authors along with collaborators with expertise in gas and aqueous phase chemical processes and meteorology. The results of such data analysis will be presented elsewhere.

3.3 Model Fitting

Our results from the RDSTM (equations (2.5)–(2.10)) were obtained numerically using a Markov chain Monte Carlo (MCMC) procedure via the WinBUGS software. Usually a complete data set (i.e., no missing data) is required to fit the multivariate version of DSTM, however we have missing values as well in the data. Our proposed RDSTM overcomes this limitation of a regular DSTM and performs imputations using Gibbs sampling. Gibbs

sampling provides a natural solution by imputing values for the missing data at each iteration, sampling from their conditional posterior predictive distribution given the available data. Regression coefficients are then updated conditional on the imputed values. We assumed that each of the standardized covariates, when missing, follows a standard normal distribution, that is, $X_{itk}^{miss} \sim N(0, 1)$.

We analyzed the data using vague priors (i.e., proper priors with large variance) on parameters to have minimal impacts on the posterior inference. We assigned independent $N(0, 10^3)$ priors to $\phi_{ii'}$, $\psi_{kk'}$ and $g_{kk'}$, and independent $G(10^3, 10^3)$ priors to $1/\sigma_{vi}^2$ and $1/\sigma_{wk}^2$. Sensitivity analysis was also performed to see the impacts of prior specifications, and it was observed that the posterior inference was insensitive to such prior specifications as long as relatively vague prior distributions were used. We obtained 10,000 iterates using a single chain from the MCMC sampler. The first 5000 iterates were discarded as a part of the Markov chain burn-in period, and all the posterior summaries reported were based on Monte Carlo estimates from the remaining 5000 iterates. The number of burn-in iterations required and the final MCMC sample sizes were chosen using trace plots for parameters by diagnosing their convergence performances. We examined trace plots of the sampled values versus iteration number to look for evidence of when the simulation appears to have stabilized to a stationary distribution.

3.4 Results

For every covariate, the RDSTM provides posterior estimates of the dynamic regression coefficient (β_t) that changes with each week from January, 1997 to July, 2004, $t = 1, \dots, 394$. This is in sharp contrast to the (static) linear regression coefficients. It is of interest to know how the covariates are dynamically related to total nitrate. RDSTM can provide the dynamic nature of the regression coefficients (see Figure 2), which provides a more informative relation between the total nitrate and other predictors.

We first checked how many weeks in each month have a significant positive/negative effect on total nitrate. For illustrative purpose, we used 95% equal-tail credible intervals to see whether the dynamic regression coefficients β_t are significant in the sense that these intervals do not contain the number zero. More formal methods, such as the use of Bayes factors and posterior predictive p-values can also be used for this purpose. To this end, we counted two separate numbers of significant weeks in each month for all the covariates. The first count provides the number of weeks which have significant positive coefficients (i.e., the lower limit of 95% interval is positive), and the other the significant negative coefficients (i.e., the upper limit is negative). The left-side plot in Figure 2 summarizes this result for SO_4 . The right-side plot in Figure 2 presents the box plots for the posterior medians of significant coefficients for SO_4 in each month. Here, a red '*' indicates the regression coefficient from linear regression models. These plots explain how strongly SO_4 is related to total nitrate across months and thus provides a better interpretation of the regression coefficients. We also computed the mean and the standard deviation of these posterior medians to summarize our findings numerically. This is given in Table 1 for SO_4 . Here, N_{sig} represents the number of significant weeks, and N_{tot} represents the total number of weeks. Using these dynamic regression coefficients, we can find which covariate has the biggest effect on the response variable, total nitrate, at what time of the year. In this article, we only show the results for SO_4 . More detailed analysis that includes results based on other explanatory variables will be reported elsewhere.

In the left plot of Figure 2, sulfate seems to have a positive relationship with total nitrate uniformly over all months. This follows from the fact that in rural areas power plants are the main source of both SO_X (SO_3 plus SO_2) and NO_X (NO plus NO_2). Also, both pollutants

would build up during stagnant meteorological conditions and be diluted during periods of high winds. In addition to the counts of significant weeks, the plot on the right-side of Figure 2 shows that the relationship and associated uncertainties between total nitrate and sulfate that varies with month. Sulfate seems to have a stronger relationship with total nitrate along with higher uncertainty during winter months. During the winter sulfate levels are low, and hence we have high levels of residual ammonium. This results in an increase of total nitrate because the residual ammonium reacts with nitrate. For linear regression models, the regression coefficients for June through September are far below those of RDSTM.

Table 1 shows the mean and the standard deviation of the posterior medians of the significant regression coefficients. The mean ranges from 0.169 (July) to 0.646 (January) and such differences are significant as evident from non-overlapping box plots (e.g., compare the box plots of June-September to the rest of the box plots). The above analysis and related results clearly show the advantages of the RDSTM over more traditional static models in terms of being much more informative in extracting the relationship between the response (e.g. total nitrate) and other explanatory variables (e.g., SO_4). As the models are not limited by stationarity and other restrictive assumptions, our estimates are robust against such model misspecifications.

4 Discussions and Future Research

The illustrative application in the previous section clearly demonstrates that our proposed reparametrized method can be used fit relatively high-dimensional ($33 \times 394=13,002$) data without making any restrictive assumptions about the spatial covariance functions. It also demonstrates that data irregularities like missing or censored observations can be handled easily. As a part of future research, the results from RDSTM could be used to diagnose the inherent problems numerical models like Community Multiscale Air Quality (CMAQ) might have with the simulations of total nitrate. As a first step in this effort, we have applied RDSTM to atmospheric measurements and obtained a dynamic relationship between total nitrate and the covariates over time. In the future, RDSTM may be applied to obtain predictive values of total nitrate based on atmospheric data at grid locations used in CMAQ. By using the methods similar to those developed by Fuentes and Raftery (2005), we can compare the RDSTM based results using the CASNET data with the RDSTM based results using the CMAQ simulated data, and make adjustments to CMAQ, if needed.

The model proposed by Wikle and Cressie (1999) closely resembles our DSTM. They introduced a dimension reduced approach to space-time Kalman filtering (STKF). The model can be written as,

$$\begin{aligned} Z_t &= \Phi a_t + v_t^*, & v_t^* &\sim N(\mathbf{0}, R+V), \\ a_t &= H a_{t-1} + \eta_t^*, & \eta_t^* &\sim N(\mathbf{0}, JQJ^T). \end{aligned}$$

This model has the same form as our DSTM except that F does not change over time and the variance matrices $R + V$ and JQJ^T are also assumed to be known (or estimated externally). Wikle and Cressie (1999) used an empirical Bayesian technique for computing efficiency, because a fully Bayesian hierarchical approach requires highly intensive computational resources for Kalman filtering. But doing so, some statistical precision (and hence efficiency) is lost which can be obtained with the fully Bayesian approach. However, our RD-STM achieve both computing efficiency by using the reparametrized univariate scheme, and statistical efficiency by using the fully hierarchical approach. Further Wikle and Cressie

used standard method of moments to estimate R , V , Q , and B for computational efficiency, and then used these estimated values as known quantities within the Kalman filter. This again leads to underestimation of uncertainty. Moreover, the use of maximum likelihood type procedures are not generally possible to implement due to numerical instabilities. As we mentioned earlier, our RDSTM approach is not limited by such approximations due to high dimensionality. In addition, it is not clear if the positive definiteness is ensured for the estimated covariance matrices obtained by such moment based methods. In contrast, our RDSTM guarantees that the covariance matrices are positive definite and the estimates are robust to spatial covariance misspecifications. Finally, it is well-known that a fully hierarchical method of estimation provides the true measure of uncertainty as compared to two-stage methods that uses plugged-in estimates from the first stage.

As a final remark our proposed RDSTM also can be extended to develop a very flexible class of space-time models that are not subject to structural restrictions. Several possible extensions are possible and we list a few of those as follows:

- i. *Dynamic modeling of covariance functions*: we can allow Σ^v and Σ^w to depend on t , that is, ϕ_{ii} 's ψ_{kk} 's, σ_i^{v2} 's, and σ_k^{w2} 's, depend on t and thus making these parameters dynamic as well.
- ii. *Relaxing the Gaussian assumption on errors*: if we suspect the presence of outliers, we can assume other than Gaussian distributions, such as t -distributions for v_t 's and w_t 's.
- iii. *Extension of the Markovian assumption*: we can assume θ_t is generated by a higher order Markovian process rather than the first-order processes used in the evolution equations in (7) and (8).

Acknowledgments

We would like to thank Dr. Prakash Bhawe from the NOAA Atmospheric Sciences Modeling Division, RTP, NC for his many insights on the chemical aspects of this work, and Dr. Jerry M. Davis from the NCSU Department of Marine, Earth & Atmospheric Sciences for his meteorological expertise. We would also like to thank these same two individuals and Dr. Alice Gilliland and Mr. Steven Howard from the NOAA Atmospheric Sciences Modeling Division, RTP, NC for providing the data used in this research. We would also like to thank Dr. Dave Holland at EPA for the help and support provided by him during this research. The research work of both authors was partially funded by an EPA cooperative grant (Co-Op 533246).

References

- Banerjee S, Gamerman D, Gelfand AE. Spatial process modelling for univariate and multivariate dynamic spatial data. *Environmetrics*. 2005; 16:465–479.
- Chen Z, Dunson DB. Random effects selection in linear mixed models. *Biometrics*. 2003; 59:762–769. [PubMed: 14969453]
- Cressie NAC, Huang HC. Classes of nonseparable, spatiotemporal stationary covariance functions. *Journal of the American Statistical Association*. 1999; 94:1330–1340.
- Daniels MJ, Pourahmadi M. Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika*. 2002; 89:553–566.
- Daniels MJ, Zhao Y. Modeling the random effects covariance matrix in longitudinal data. *Statistics in Medicine*. 2003; 22:1631–1647. [PubMed: 12720301]
- Fuentes M, Raftery AE. Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. *Biometrics*. 2005; 66:36–45. [PubMed: 15737076]
- Gelfand AE, Ghosh SK, Knight JR, Sirmans CF. Spatio-Temporal Modeling of Residential Sales Data. *Journal of Business & Economic Statistics*. 1998; 16:312–321.

- Gneiting T. Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association*. 2002; 97:590–600.
- Huang HC, Cressie NAC. Spatio-temporal prediction of snow water equivalent using the Kalman filter. *Computational Statistics and Data Analysis*. 1996; 22:159–175.
- Huang HC, Hsu NJ. Modeling transport effects on ground-level ozone using a non-stationary space-time model. *Environmetrics*. 2004; 15:251–268.
- Huerta G, Sanso B, Stroud JR. A spatio-temporal model for Mexico city ozone levels. *Journal of the Royal Statistical Society, Series C*. 2004; 53:231–248.
- Kyriakidis PC, Journel AG. Geostatistical space-time models: a review. *Mathematical Geology*. 1999; 31(6):651–684.
- Mardia KV, Goodall C, Redfern EJ, Alonso FJ. The kriged Kalman filter (with discussion). *Test*. 1998; 7:217–285.
- National Research Council. *Air Quality Management in the United States*. The National Academic Press; 2004.
- Pole, A.; West, M.; Harrison, PJ. *Applied Bayesian forecasting and times series analysis*. Chapman and Hall; New York: 1994.
- Pourahmadi M. Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika*. 1999; 86:677–690.
- Sansó B, Guenni L. Venezuelan rainfall data analyzed using a Bayesian space-time model. *Applied Statistics*. 1999; 48:345–362.
- Stein ML. Space-time covariance functions. *Journal of the American Statistical Association*. 2003; 100:310–321.
- Stroud JR, Müller P, Sanso B. Dynamic models for spatio-temporal data. *Journal of Royal Statistical Society, Series B*. 2001; 63:673–689.
- Tonellato, S. *Bayesian dynamic linear models for spatial time series*. Dipartimento di Statistica, Università CaFoscari di Venezia; Venice, Italy: 1997. Technical report (Rapporto di ricerca 5/1997)
- West, M.; Harrison, PJ. *Bayesian Forecasting and Dynamic Models*. 2. Springer; New York: 1997.
- Wikle C, Berliner M, Cressie N. Hierarchical Bayesian space-time models. *Environmental and Ecological Statistics*. 1999; 5:117–154.
- Wikle C, Cressie N. A dimension reduced approach to space-time kalman filtering. *Biometrika*. 1999; 86:815–829.
- Xu K, Wikle C. Estimation of Parameterized Spatio-Temporal Dynamic Models. *Ecological and Environmental Statistics*. 2005 to appear.

Figure 1.
Locations of Stations (indicated by red dots).

Figure 2.
Frequencies and Box Plots for Posterior Medians of Significant Weeks in Each Month for SO_4 .

Table 1

Mean and standard deviations of Posterior Medians of Significant Regression Coefficients for SO_4 .

Month	N_{sig}	N_{tot}	Mean	S.E.
January	34	34	0.646	0.322
February	33	33	0.441	0.116
March	34	35	0.496	0.167
April	35	35	0.425	0.134
May	30	34	0.354	0.224
June	31	35	0.194	0.106
July	32	35	0.169	0.085
August	27	31	0.202	0.136
September	30	30	0.290	0.137
October	31	31	0.405	0.194
November	29	29	0.502	0.212
December	32	32	0.528	0.236