Appl. Statist. (2005)
54, *Part* 1, *pp.* 223–244

A Bayesian kriged Kalman model for short-term forecasting of air pollution levels

Sujit K. Sahu

University of Southampton, UK

and Kanti V. Mardia

University of Leeds, UK

[Received April 2003. Final revision January 2004]

Summary. Short-term forecasts of air pollution levels in big cities are now reported in newspapers and other media outlets. Studies indicate that even short-term exposure to high levels of an air pollutant called atmospheric particulate matter can lead to long-term health effects. Data are typically observed at fixed monitoring stations throughout a study region of interest at different time points. Statistical spatiotemporal models are appropriate for modelling these data. We consider short-term forecasting of these spatiotemporal processes by using a Bayesian kriged Kalman filtering model. The spatial prediction surface of the model is built by using the well-known method of kriging for optimum spatial prediction and the temporal effects are analysed by using the models underlying the Kalman filtering method. The full Bayesian model is implemented by using Markov chain Monte Carlo techniques which enable us to obtain the optimal Bayesian forecasts in time and space. A new cross-validation method based on the Mahalanobis distance between the forecasts and observed data is also developed to assess the forecasting performance of the model implemented.

Keywords: Bending energy; Gibbs sampler; Kalman filter; Kriging; Markov chain Monte Carlo methods; Spatial temporal modelling; State space model

1. Introduction

In recent years there has been a tremendous growth in the statistical models and techniques to analyse spatiotemporal data such as air pollution data. Spatiotemporal data arise in many other contexts, e.g. disease mapping and economic monitoring of real-estate prices. Often the primary interests in analysing such data are to smooth and predict time evolution of some response variables over a certain spatial domain.

Cressie (1994) and Goodall and Mardia (1994) obtained models for spatiotemporal data. Mardia *et al.* (1998) introduced a combined approach which they call kriged Kalman filter modelling. Recent references within this broad framework include Sansó and Guenni (1999, 2000), Stroud *et al.* (2001), Kyriakidis and Journel (1999), Wikle and Cressie (1999), Wikle *et al.* (1998), Brown *et al.* (2000), Allcroft and Glasbey (2003) and Kent and Mardia (2002).

Kent and Mardia (2002) provided a unified approach to spatiotemporal modelling through the use of drift and/or correlation in space and/or time to accommodate spatial continuity. For drift functions, they have emphasized the use of so-called principal kriging functions, and for

Address for correspondence: Sujit K. Sahu, School of Mathematics, Southampton Statistical Sciences Research Institute, University of Southampton, Highfield, Southampton, SO17 1BJ, UK.
E-mail: S.K.Sahu@maths.soton.ac.uk

 0035–9254/05/54223

correlations they have discussed the use of a first-order Markov structure in time combined with spatial blurring. Here we adopt one of their strategies but in a full Bayesian framework.

We work here with a process which is continuous in space and discrete in time. The underlying spatial drift is modelled by the principal kriging functions and the time component at observed sites is modelled by a vector random-walk process. The dynamic random-walk process models stochastic trend and the resulting Bayesian analysis essentially leads to Kalman filtering, which is a computational method to analyse dynamic time series data; see for example Mardia *et al.* (1998). In addition, the models proposed are presented in a hierarchical framework following Wikle *et al.* (1998). This allows the inclusion of a 'nugget' term in the spatial part of the model. The model is fitted and used for forecasting in a unified computational framework by using Markov chain Monte Carlo (MCMC) methods. The MCMC methods replace the task of Kalman filtering by using a random-walk model in time.

The plan of the remainder of the paper is as follows. In Section 2 we describe the data set that is used in this study. Section 3 describes the hierarchical Bayesian kriged Kalman filter model. Important computational details are discussed in Section 4. In Section 5 we return to the analysis of the data set that is described in Section 2. The paper ends with a discussion. The data that are analysed in the paper can be obtained from

`http://www.blackwellpublishing.com/rss`

2. New York City air pollution data

This paper is motivated by the need to develop coherent Bayesian computational methodology implementing flexible hierarchical models for short-term forecasting of spatiotemporal processes. In environmental monitoring and prediction problems it is often desired to predict the dependent variable, e.g. pollution level and rainfall, for 5 days or at most a week in advance.

The Environmental Protection Agency in the USA monitor atmospheric particulate matter that is less than 2.5 μm in size known as PM2.5. This PM2.5 measure is one of six primary air pollutants and is a mixture of fine particles and gaseous compounds such as sulphur dioxide (SO_2) and nitrogen oxides (NO_x). Interest in analysing fine particles such as PM2.5 comes from the fact that as those particles are less than 2.5 μm in diameter they are sufficiently small to enter the lungs and can cause various health problems. Short-term forecasting of PM2.5 levels is the focus of this paper.

The data set that we analyse here is the PM2.5 concentration data that were observed at 15 monitoring stations in the city of New York during the first 9 months of 2002. The data are observed once in every 3 days and during the first 9 months there were 91 equally spaced days. Out of these 1365 ($= 15 \times 91$) data points 126 were missing observations which we take to be missing completely at random.

Let $z(\mathbf{s}_i, t)$ denote the observed PM2.5 concentration level at site $\mathbf{s}_i$ and at time t where $i = 1, \ldots, n$ and $t = 1, \ldots, T$. Here we have $n = 15$ and $T = 91$.

Fig. 1 shows the locations of the sites numbered 1–15. The first three monitoring sites are in the Bronx area of the city, sites 4–6 are in Brooklyn, sites 7–10 are in Manhattan, sites 11–13 are in Queens and lastly sites 14 and 15 are in Staten Island. These five boroughs constitute the city of New York.

There are considerable spatiotemporal variations in these data. Fig. 2 provides the sitewise box plots of the data. The plot shows that sites 7 and 8 in the Manhattan area are more polluted than others. The concentration levels at sites 4–6 in Brooklyn are similar. However, the variations at

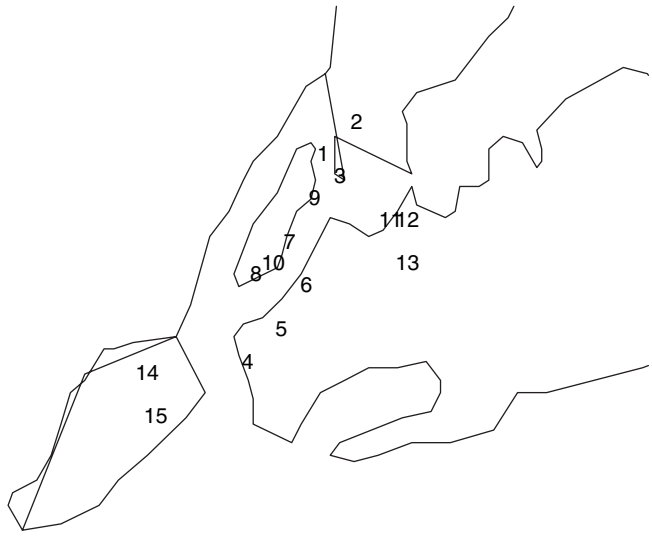


Fig. 1. 15 monitoring sites in New York City

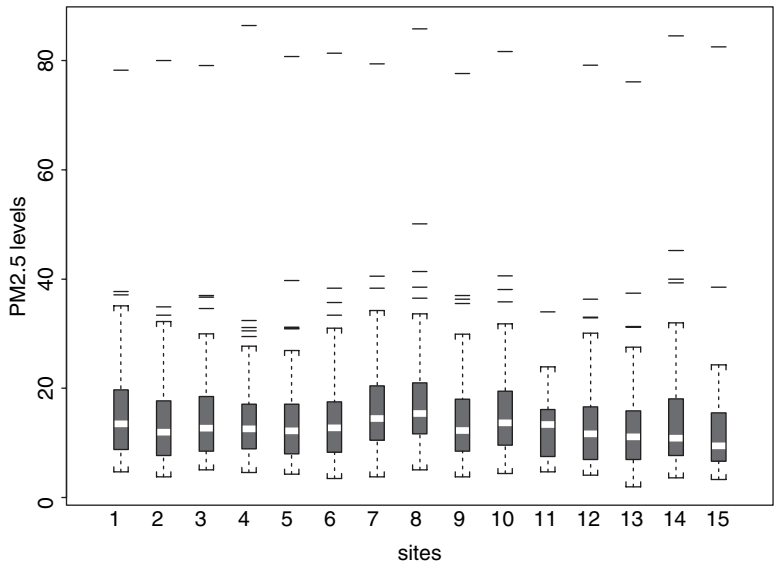


Fig. 2. Box plot of the data at 15 sites

sites 14 and 15 are not similar, although they are on the same island. More discussion regarding Fig. 2 is given later.

We formally investigate the spatial variation by using an empirical variogram of the data. We first remove the temporal trends by taking the first differences for each time series from the 15 sites, i.e. we obtain, $w(\mathbf{s}_i, t) = z(\mathbf{s}_i, t+1) - z(\mathbf{s}_i, t)$ for $t = 1, \ldots, T-1$ and $i = 1, \ldots, n$. The time series plots of the difference data (not shown) confirmed that there were no more temporal effects, but there were a few outliers. The variation in the resulting data $w(\mathbf{s}_i, t)$ (without the outliers) can be expected to have arisen from variation due to space.

To understand the behaviour of an isotropic and stationary process $W(\mathbf{s}, t)$ we use the variogram defined by

$$2\,\gamma(d) = E[\{W(\mathbf{s}_1, t) - W(\mathbf{s}_2, t)\}^2]$$

where d is the distance between the spatial locations $\mathbf{s}_1$ and $\mathbf{s}_2$. Traditionally variograms are calculated by grouping the possible values of d into bins, and by computing one value by taking the sample average of $\{w(\mathbf{s}_1, t) - w(\mathbf{s}_2, t)\}^2$ values for which the distance d between $\mathbf{s}_1$ and $\mathbf{s}_2$ lies within a given bin. Here, we adopt a slightly different procedure. The estimate of $\gamma(d)$ for an observed distance of d is given by

$$\hat{\gamma}(d) = \frac{1}{2(T-1)} \sum_{t=1}^{T-1} \{w(\mathbf{s}_1, t) - w(\mathbf{s}_2, t)\}^2,$$

assuming that there are no missing observations. We remove the missing observations from the above sum and adjust the denominator accordingly.

We use the geodetic distance between two locations with given latitudes θ_1 and θ_2 and longitudes ϕ_1 and ϕ_2 (converted to radians). The geodetic distance d is the distance at the surface of the Earth considered as a sphere of radius $R = 6371$ km. We use the formula

$$d = 6371 \cos^{-1}(B) \text{ (km)},$$

where

$$B = \sin(\theta_1)\sin(\theta_2) + \cos(\theta_1)\cos(\theta_2)\cos(\phi_1 - \phi_2).$$

Fig. 3 provides a plot of the estimated variogram $\hat{\gamma}(d)$ against d. Site 14 has been omitted from this plot because it contained two outlying extreme observations on June 10th and 13th, and these high observations distorted the general linear trend that is seen in the variogram plot. We shall return to this issue later in Section 5.2.

The variogram plot (Fig. 3) shows strong linear spatial variations. The full curve in Fig. 3 is the empirical LOESS fit (S-PLUS function `loess`) to the estimated variogram. The variogram plot does not show a clear finite range and a finite sill. However, a finite range and a finite sill can be seen if the five extreme variogram values for distance values above 30 were ignored. The

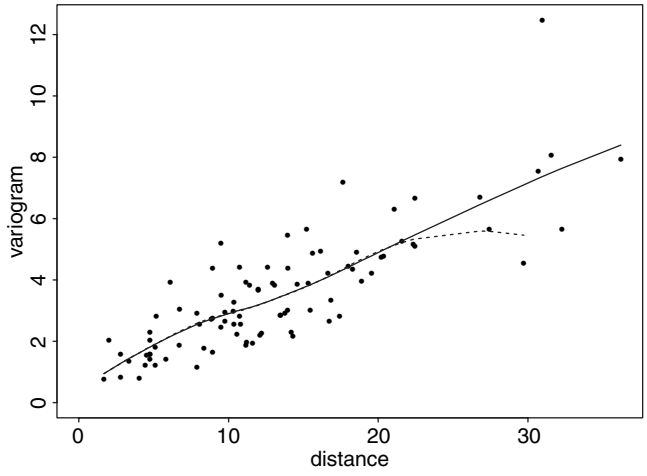


Fig. 3. Variogram of the differenced data after removing site 14: ———, empirical LOESS fit; · · · · · · ·, empirical LOESS fit after removing five extreme points corresponding to distance values more than 30

dotted curve in the plot is the LOESS fit to the variogram after removing these five extreme values. The underlying theoretical variogram corresponding to the dotted curve does indicate the presence of a finite sill and a finite range.

Note, however, that the plotted variograms are to be treated as exploratory tools where the main objective is to show spatial variations in the data. These exploratory and empirical variograms should not be confused with the Matérn family (Matérn, 1986) of covariance functions that are assumed in Section 3.1 for the latent variables which appear in a lower level hierarchy of model building. The latent variables there are not same as the differenced data points $w(\mathbf{s}, t)$ here. See Section 5.2 for more discussion regarding this.

The box plots in Fig. 2 also indicate that there is a very large observed value at each site. Further investigation (Fig. 4) shows that the large observation at each site was for July 7th which was the first day of monitoring after the July 4th firework celebrations. These large observations are also seen to be positively skewed (see the box plot of the data for July 7th plotted in Fig. 5). In Fig. 5 the box plot of the data for July 4th is also presented for comparisons. This plot shows negative skewness for the PM2.5 concentration data on July 4th. Perhaps this is to be expected for pollution data on a regular day since high levels of concentration can only be expected to occur at a few sites. In any case this sort of differences in observed variations will affect the spatial predictions; see Section 5.2 where we report the spatial predictions for both July 4th and July 7th.

These very large observations make the data non-stationary in time and will cause problems in modelling using traditional regression-based methods. The short-term forecasting models that we propose here are non-stationary and are seen to be adequate for the entire data; see Section 5. Moreover, our modelling approach here does not require explicit modelling of the large observations (e.g. using an indicator covariate for the days with large observations).

Some exploratory linear models fitted to both the raw data sets and their transformations suggested that it is better to model the square-root transformation of the data which encouraged normality. Smith *et al.* (2003) also reported similar findings. Henceforth, we model the square root of the data. However, we make the predictions on the original scale for ease of communicating to practitioners.

The Environmental Protection Agency use mostly linear regression models to forecast the PM2.5 levels. Models based on classification and regression trees are also used sometimes; see for example Dye *et al*. (2002). Some explanatory variables, e.g. precipitation, temperature, wind speed and holidays, are used in their models. However, there are several limitations in their approach. The main drawbacks arise because regression models cannot be used satisfactorily for data which are correlated in space and time. The explanatory variables can be used in our analysis as well perhaps to enhance model fitting, but we do not include those here because some of the explanatory variables are themselves to be predicted first to obtain forecasts of PM2.5.

Smith *et al.* (2003) analysed PM2.5 data for North Carolina, South Carolina and Georgia by using specific models for spatial and temporal effects. They used weekly dummy variables to model the time effect and incorporated a spatial trend model using thin plate splines. See for example Mardia and Goodall (1993) for more on thin plate splines. Moreover, they have included covariates, e.g. land use, in their model to discriminate between concentration levels in the vast area that is covered by the three states.

3. The kriged Kalman filter model

The general model that we propose here is for spatiotemporal data recorded at n sites $\mathbf{s}_i$, $i = 1, \ldots, n$, over a period of T equally spaced time points. Let $\mathbf{Z}_t = (Z(\mathbf{s}_1, t), \ldots, Z(\mathbf{s}_n, t))'$ denote the n-dimensional observation vector at time point t, $t = 1, \ldots, T$.

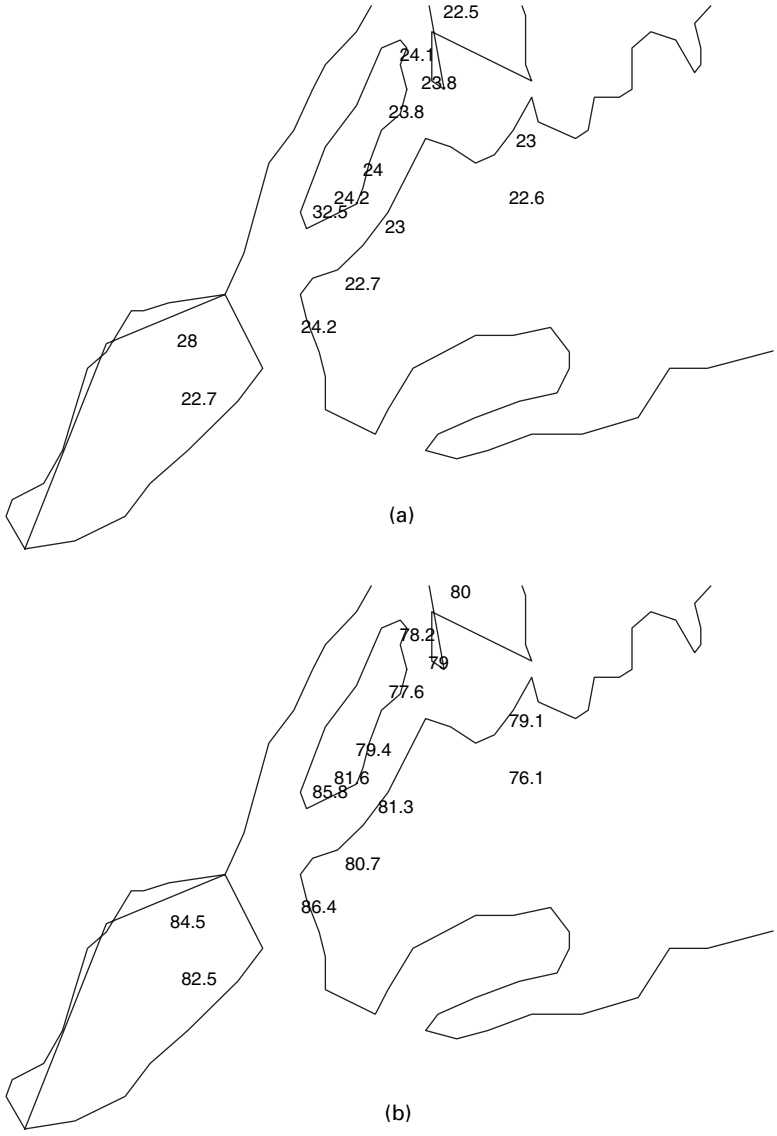


Fig. 4. Raw data for (a) July 4th and (b) July 7th

Often, the first step in modelling spatiotemporal data is to assume a hierarchical model

$$\mathbf{Z}_t = \mathbf{Y}_t + \varepsilon_t \tag{1}$$

where $\mathbf{Y}_t = (Y(\mathbf{s}_1, t), \ldots, Y(\mathbf{s}_n, t))'$ is an unobserved but scientifically meaningful process (signal) and ε_t is a white noise process. Thus we assume that the components of ε_t are independent and identically distributed normal random variables with mean 0 and unknown variance σ_ε^2. In geostatistics, these error terms are often known as a nugget effect. A certain specific correlation structure for ε can also be considered. However, we assume specific structures in the next level of model hierarchy. The prior distribution for $\tau_\varepsilon^2 = 1/\sigma_\varepsilon^2$ is assumed to be the gamma distribution with shape parameter a and rate parameter b. We assume that $a = b = 0.001$ so that the gamma

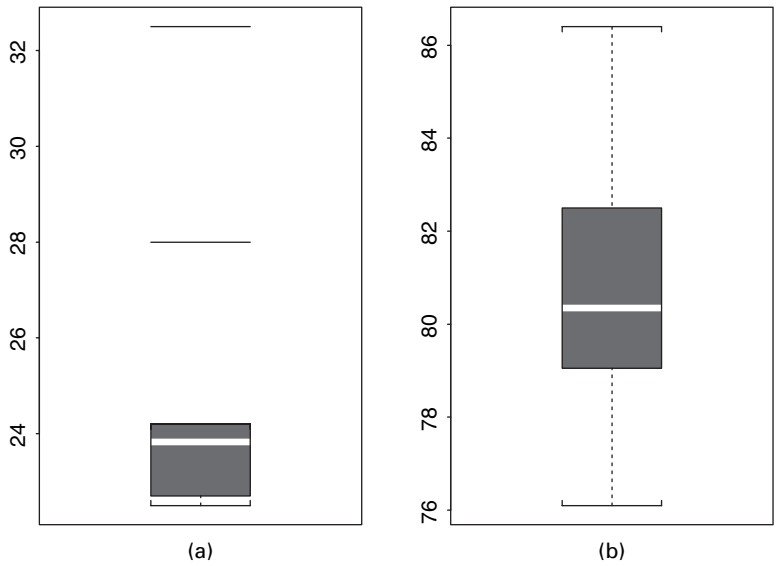


Fig. 5. Box plots of the data for (a) July 4th and (b) July 7th

distribution has mean 1 and variance 1000. The resulting prior distribution has the desirable property that it is proper but diffuse.

The space–time process $\mathbf{Y}_t$ is thought to be the sum of parametric systematic components $\boldsymbol{\theta}_t$ and an isotropic time homogeneous spatial process denoted by $\boldsymbol{\gamma}_t$. Thus we assume that

$$\mathbf{Y}_t = \boldsymbol{\theta}_t + \boldsymbol{\gamma}_t \tag{2}$$

where the error term $\boldsymbol{\gamma}_t$ is assumed to be zero mean Gaussian with covariance matrix Σ_γ which has elements

$$\sigma(\mathbf{s}_i, \mathbf{s}_j) = \mathrm{cov}\{Y(\mathbf{s}_i, t), Y(\mathbf{s}_j, t)\} \tag{3}$$

for $i, j = 1, \ldots, n$. The quantity $\sigma(\mathbf{s}_i, \mathbf{s}_j)$ is the covariance function of the spatial process to be specified later. The components of $\boldsymbol{\theta}_t$ are unspecified as well and will be discussed in the following subsections.

The modelling hierarchies (1) and (2) are used when it is desired to predict the smooth process $Y(\mathbf{s}, t)$ rather than the observed noisy process $Z(\mathbf{s}, t)$; see for example Wikle and Cressie (1999). They also pointed out that it is not desirable to coalesce the two equations into

$$\mathbf{Z}_t = \boldsymbol{\theta}_t + \boldsymbol{\gamma}_t + \boldsymbol{\varepsilon}_t. \tag{4}$$

This equation also defines an inefficient and often unidentifiable parameterization; see for example Gelfand *et al.* (1995) for other examples.

3.1. Models for the spatial covariance

We assume that the covariance function belongs to the Matérn family (Matérn, 1986)

$$\sigma(\mathbf{s}_i, \mathbf{s}_j) = \sigma_\gamma^2 \frac{1}{2^{\kappa-1}\,\Gamma(\kappa)} \lambda d_{ij}\, K_\kappa(\lambda d_{ij}), \qquad \lambda > 0, \quad \kappa \geqslant 1, \tag{5}$$

where d_{ij} is the geodetic distance between sites $\mathbf{s}_i$ and $\mathbf{s}_j$, $K_\kappa(\cdot)$ is the modified Bessel function of the second kind and of order κ; see for example Berger *et al.* (2001). For our illustration we take

$\kappa = 1$ and consider several values for λ. We choose the particular λ by using a predictive model choice criterion. We estimate σ_γ^2 by using MCMC methods. There are many possible parametric and semiparametric models for covariance of isotropic spatial processes; see for example Ecker and Gelfand (1997) where a Bayesian model choice study has been presented.

In our Bayesian set-up, a prior distribution for σ_γ^2 must be specified. Here we assume that $\tau_\gamma^2 = 1/\sigma_\gamma^2$ follows the gamma prior distribution with parameters a and b. We take $a = b = 0.001$ so that the gamma distribution has mean 1 and variance 1000. Our choice avoids the default improper prior distribution, namely

$$\pi(\sigma_\gamma^2) = 1/\sigma_\gamma^2, \qquad \sigma_\gamma^2 > 0,$$

because this may lead to improper posterior distributions which would be difficult to verify in practice; see for example Berger *et al.* (2001) and Gelfand and Sahu (1999).

3.2. Principal kriging functions

The systematic component θ_t is assumed to evolve as a stochastic time-varying linear combination of some optimal spatial functions. These are taken to be the principal kriging functions following Kent and Mardia (2002) and Mardia *et al.* (1998). Given a certain known covariance function, the unbiased linear prediction of the spatial process is called 'kriging'. The principal kriging functions are used as the optimal spatial functions on which the dynamic temporal effects take place. Thus the first term of $\boldsymbol{\theta}_t$ is given by

$$H\boldsymbol{\alpha}_t = \left(\sum_{j=1}^{p} h_{\mathbf{s}_1 j}\alpha_{tj}, \ldots, \sum_{j=1}^{p} h_{\mathbf{s}_n j}\alpha_{tj} \right)'$$

where the matrix H is $n \times p$ with ijth element $h_{\mathbf{s}_i j}$, for $i = 1, \ldots, n$, $j = 1, \ldots, p$ and $\boldsymbol{\alpha}_t = (\alpha_{t1}, \ldots, \alpha_{tp})'$. The choice of p is discussed at the end of this section. The columns of H are determined by principal fields in kriging space and $\boldsymbol{\alpha}_t$ is a temporal state vector which varies in time.

The matrix H quantifies the spatial component in the model; when multiplied by the dynamic time component $\boldsymbol{\alpha}_t$, it provides a time-varying linear combination of the spatial regression surface that is described by the columns of H. The columns consist of two sets of spatial trend fields. The first set of q columns corresponds to the constant, linear and quadratic functions of co-ordinate dimensions, say. For example, if $q = 3$ and $d = 2$ the first column can be chosen to be $\mathbf{1}$ corresponding to the constant trend field and the entries in the other two columns can be taken as the X- and Y-co-ordinates of the locations where data have been observed. This $n \times q$ matrix is denoted by F in the following discussion.

The remaining $p - q$ fields are chosen as the spatial directions relative to an assumed covariance structure. The directions are obtained as follows. Assume, for developing the principal functions, that the data are collected for only one time point; thus the suffix t is suppressed in the following discussion. Let Σ_γ and $F'\Sigma_\gamma^{-1}F$ be non-singular matrices. Assume that $\mathbf{Z} = (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))'$ follows the multivariate normal distribution with mean and variance given by

$$E(\mathbf{Z}) = F\boldsymbol{\mu},$$

$$\mathrm{cov}(\mathbf{Z}) = \Sigma_\gamma,$$

which is a simplified version of the full model that is considered in this paper. Under this model (and a flat prior on the parameter $\boldsymbol{\mu}$) the predictive mean for a site $\mathbf{s}$ is

$$E\{Z(\mathbf{s})|\Sigma_\gamma, \mathbf{z}\} = \mathbf{f}(\mathbf{s})'A\mathbf{z} + \boldsymbol{\sigma}(\mathbf{s})'B\mathbf{z} \tag{6}$$

where $\mathbf{f}(\mathbf{s})$ is the $(q \times 1)$-vector of the trend field at the site $\mathbf{s}$; $\mathbf{z}$ is the realization and $\boldsymbol{\sigma}(\mathbf{s}) = (\sigma(\mathbf{s}, \mathbf{s}_1), \ldots, \sigma(\mathbf{s}, \mathbf{s}_n))'$;

$$A = (F'\Sigma_\gamma^{-1}F)^{-1}F'\Sigma_\gamma^{-1},$$

$$B = \Sigma_\gamma^{-1} - \Sigma_\gamma^{-1}FA.$$

Methods are available for singular Σ_γ which are required for thin plate splines; see Kent and Mardia (1994). If the site $\mathbf{s}$ coincides with any particular $\mathbf{s}_i$, $i = 1, \ldots, n$, then it is easy to see that the above predictive mean reduces to $z(\mathbf{s})$ as expected.

The matrix B is known as the *bending energy matrix*; see for example Bookstein (1989) who motivated its use from the study of thin plate splines. Consider the spectral decomposition of B,

$$B = UEU',$$

$$B\mathbf{u}_i = e_i\mathbf{u}_i,$$

where $U = (\mathbf{u}_1, \ldots, \mathbf{u}_n)$ and $E = \text{diag}(e_1, \ldots, e_n)$, and assume without loss of generality that the eigenvalues are in non-decreasing order, $e_1 = \ldots = e_q = 0 < e_{q+1} \leqslant \ldots \leqslant e_n$. It is easy to verify that B satisfies $BF = 0$. Thus the columns of F can be thought of as the eigenvectors that are associated with the null eigenvalues, $e_1, \ldots, e_q$.

Any observation vector $\mathbf{z}$ can be represented as a linear combination of the eigenvectors $\mathbf{u}_i$ since the latter set forms a basis. Indeed, suppose that $\mathbf{z} = \sum_{i=1}^n c_i\mathbf{u}_i$ for suitable constants c_i. Now the predictive mean (6) reduces to

$$\mathbf{f}(\mathbf{s})'A \sum_{i=1}^{n} c_i\mathbf{u}_i + \sum_{i=q+1}^{n} c_i e_i\, \boldsymbol{\sigma}(\mathbf{s})'\mathbf{u}_i.$$

Thus the predictive mean is a linear combination of the q trend fields $f_1(\mathbf{s}), \ldots, f_q(\mathbf{s})$ and the $n - q$ *principal kriging functions* $e_i\, \boldsymbol{\sigma}(\mathbf{s})'\mathbf{u}_i$. These functions span the space of all kriging solutions with observations at the n given sites, the specified trend fields and the covariogram. We shall use the terms principal kriging functions and principal fields interchangeably henceforth.

The smaller eigenvalues of B are associated with large scale spatial variation (global features) and the larger eigenvalues describe local spatial variation. This can be inferred from the fact that the global trend fields that are described by the columns of F are the eigenvectors corresponding to the zero eigenvalues of B. See also Kent and Mardia (2002) and Mardia *et al.* (1998) for more details in this regard. In practice, for model reduction, we may choose to work with $p - q < n - q$ principal functions. Thus, when the values at the observed sites are to be predicted, we choose the $p - q$ columns of H to be $e_i\Sigma_\gamma\mathbf{u}_i$, $i = q + 1, \ldots, p$. Hence the matrix H is taken as

$$H = (F, e_{q+1}\Sigma_\gamma\mathbf{u}_{q+1}, \ldots, e_p\Sigma_\gamma\mathbf{u}_p). \tag{7}$$

In what follows we shall illustrate the choice of p and q in particular examples, including the case $p = q$ for which no principal kriging functions are taken in the model. The model with only polynomials (without the principal fields) are often used in the literature; see for example the spatiotemporal model that was adopted by Sansó and Guenni (1999). Principal kriging functions have some advantages over only polynomial-type trend functions, which

is the case for $p = q$. They grow less quickly than polynomials outside the domain of the data.

3.3. Dynamic temporal trend models

Motivated by our example, here we concentrate on smoothing and short-term forecasting in the temporal domain. A standard procedure in such cases is to adopt a random-walk state space type of formulation for temporal components; see for example Stroud *et al.* (2001) and Gelfand *et al.* (2004). We thus assume that

$$\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1} + \boldsymbol{\eta}_t, \tag{8}$$

where the p-dimensional error term $\boldsymbol{\eta}_t$ is assumed to be normally distributed with mean 0 and covariance matrix Σ_η. To complete the modelling hierarchies we suppose that $\boldsymbol{\alpha}_0 \sim N(0, C_\alpha I)$ and with a large value of C_α. Here I denotes the identity matrix of appropriate order. See West and Harrison (1997) for more on dynamic time series models.

We assume that $Q_\eta = \Sigma_\eta^{-1}$ has the Wishart prior distribution, i.e.

$$Q_\eta \sim W_p(2a_\eta, 2b_\eta)$$

where $2a_\eta$ is the assumed prior degrees of freedom (greater than or equal to p) and b_η is a known positive definite matrix, to be specified later. We say that $\mathbf{X}$ has the Wishart distribution $W_p(m, R)$ if its density is proportional to

$$|R|^{m/2}|x|^{(m-p-1)/2} \exp\{-\tfrac{1}{2}\operatorname{tr}(Rx)\}$$

if x is a $p \times p$ positive definite matrix; see for example Mardia *et al.* (1979), page 85. (Here $\operatorname{tr}(A)$ is the trace of a matrix A.) To obtain diffuse but proper prior distributions we choose $a_\eta = p/2$. This assumption makes the prior distributions worth the same number of observations as the corresponding dimensions and is often used in a multivariate Bayesian modelling framework. The matrix $2b_\eta$ is chosen to be 0.01 times the identity matrix. This again comes from the requirement of assuming diffuse prior distributions.

An alternative to the assumption of stochastic trend is to consider deterministic polynomial trend models. For example, we can assume that $\boldsymbol{\alpha}_t = (1, t, t^2, \ldots, t^{p-1})$. This polynomial trend model is not as flexible as the stochastic trend model (8). Hence we do not consider the polynomial trend model at all, and we always work with the stochastic trend model (8).

A referee has commented that from a fluid dynamics perspective this random-walk model cannot be fully justified for atmospheric systems. In fact, Mardia *et al.* (1998) have taken the state equation (8) of the form

$$\boldsymbol{\alpha}_t = P\boldsymbol{\alpha}_{t-1} + \boldsymbol{\eta}_t$$

with unknown transition matrix P. There are some identifiability problems with this approach as discussed by Kent and Mardia (2002) for a general P and a general covariance matrix for $\boldsymbol{\eta}_t$. They showed that it is sufficient to assume that the largest eigenvalue of P is less than 1 in absolute value and the matrix H is of full rank.

In our Bayesian set-up the identifiability problems can be resolved by assuming proper prior distributions for both P and $\boldsymbol{\alpha}_t$. However, we model with the choice $P = I$ which is motivated by the need to develop models for short-term forecasting. Moreover, this choice avoids insurmountable problems in MCMC convergence (which we have encountered) arising from the weak identifiability of the parameters under sufficiently diffuse prior distributions.

4. Computations

4.1. The joint posterior distribution

To obtain the joint posterior distribution we recall that

$$Z(\mathbf{s}_i,t)|Y(\mathbf{s}_i,t)\sim N\{Y(\mathbf{s}_i,t),\sigma_\varepsilon^2\},\qquad i=1,\ldots,n,\quad t=1,\ldots,T,$$

where

$$\mathbf{Y}_t=(Y(\mathbf{s}_1,t),Y(\mathbf{s}_2,t),\ldots,Y(\mathbf{s}_n,t))'\sim N(\boldsymbol{\theta}_t,\Sigma_\gamma),\qquad t=1,\ldots,T,$$

independently. Further, we have assumed that

$$\theta(\mathbf{s},t)=\sum_{j=1}^{p}h_{\mathbf{s}j}\alpha_{tj},\tag{9}$$

and $\boldsymbol{\alpha}_t\sim N(\boldsymbol{\alpha}_{t-1},\Sigma_\eta)$ for $t=1,\ldots,T$ and $\boldsymbol{\alpha}_0\sim N(0,C_\alpha I)$.

Let $\boldsymbol{\xi}$ denote the following exhaustive set of parameters:

(a) the error precision parameters, $\tau_\gamma^2=1/\sigma_\gamma^2$ and $\tau_\varepsilon^2=1/\sigma_\varepsilon^2$, and
(b) the latent process $\mathbf{Y}_t$, $t=1,\ldots,T$,
(c) the dynamic parameters, $\boldsymbol{\alpha}_t$, $t=1,\ldots,T$, and their precision matrix $Q_\eta=\Sigma_\eta^{-1}$, and
(d) the missing data, $Z^*(\mathbf{s},t)$ for all $\mathbf{s}$ and t for which $Z(\mathbf{s},t)$ is missing.

The log-likelihood function for the hierarchical model is given by

$$\log\{f(\mathbf{z}_1,\ldots,\mathbf{z}_T|\boldsymbol{\xi})\}\propto\frac{nT}{2}\log(\tau_\varepsilon^2)-\frac{\tau_\varepsilon^2}{2}\sum_{t=1}^{T}(\mathbf{z}_t-\mathbf{y}_t)'(\mathbf{z}_t-\mathbf{y}_t)-\frac{T}{2}\log|\Sigma_\gamma|$$
$$-\frac{1}{2}\sum_{t=1}^{T}(\mathbf{y}_t-\boldsymbol{\theta}_t)'\Sigma_\gamma^{-1}(\mathbf{y}_t-\boldsymbol{\theta}_t).$$

The joint posterior density is obtained, up to a normalizing constant, as the product of the above likelihood function and the prior distributions for the parameters in the model, i.e.

$$\pi(\boldsymbol{\xi}|\mathbf{z}_1,\ldots,\mathbf{z}_T)\propto f(\mathbf{z}_1,\ldots,\mathbf{z}_T|\boldsymbol{\xi})\,\pi(\boldsymbol{\xi})\tag{10}$$

where $\pi(\boldsymbol{\xi})$ denotes the prior distribution that is assumed for the parameters in $\boldsymbol{\xi}$ except for the missing data $Z^*(\mathbf{s},t)$.

4.2. The full conditional distributions

We derive the full conditional distributions that are needed for Gibbs sampling under both the above models; see for example Carter and Kohn (1994) for similar calculations in state space models. The full conditional distribution of τ_ϵ^2 is the gamma distribution with parameter $a+Tn/2$ and

$$b+\tfrac{1}{2}\sum_{t=1}^{T}(\mathbf{z}_t-\mathbf{y}_t)'(\mathbf{z}_t-\mathbf{y}_t).$$

The full conditional distribution of $\mathbf{y}_t$ is the multivariate normal distribution $N(V\boldsymbol{\mu}_t,V)$ where

$$V^{-1}=\tau_\varepsilon^2 I+\Sigma_\gamma^{-1},$$
$$\boldsymbol{\mu}_t=\tau_\varepsilon^2\mathbf{z}_t+\Sigma_\gamma^{-1}\boldsymbol{\theta}_t.$$

The full conditional distribution of τ_γ^2 is the gamma distribution with parameters $a + Tn/2$ and

$$b + \tfrac{1}{2} \sum_{t=1}^{T} (\mathbf{y}_t - \boldsymbol{\theta}_t)' V^{-1} (\mathbf{y}_t - \boldsymbol{\theta}_t),$$

where $V_{ij} = (\lambda d_{ij}) K_\kappa(\lambda d_{ij})$. This conjugate distribution is obtained by using the facts

(a) $\Sigma_\gamma = \sigma_\gamma^2 V$ where V is free of σ_γ^2 and
(b) H is invariant with respect to (i.e. free of) σ_γ^2.

The second claim is proved as follows. Note that the matrices A and F are free of σ_γ^2; $B = \tau_\gamma^2 (V^{-1} - V^{-1} FA)$. The eigenvalues of B will be τ_γ^2 multiples of the eigenvalues of $V^{-1} - V^{-1} FA$ (which is free of σ_γ^2). The multiplier τ_γ^2 cancels out when forming H because of the premultiplication by Σ_γ.

The full conditional distribution of $\boldsymbol{\alpha}_t$ is $N(V_t \boldsymbol{\mu}_t, V_t)$ where

$$\begin{aligned}
V_t^{-1} &= I/C_\alpha + Q_\eta, \quad \boldsymbol{\mu}_t = Q_\eta \boldsymbol{\alpha}_{t+1}, && \text{when } t = 0, \\
V_t^{-1} &= H' \Sigma_\gamma^{-1} H + 2Q_\eta, \quad \boldsymbol{\mu}_t = H' \Sigma_\gamma^{-1} \mathbf{y}_t + Q_\eta (\boldsymbol{\alpha}_{t-1} + \boldsymbol{\alpha}_{t+1}), && \text{when } 0 < t < T, \\
V_t^{-1} &= H' \Sigma_\gamma^{-1} H + Q_\eta, \quad \boldsymbol{\mu}_t = H' \Sigma_\gamma^{-1} \mathbf{y}_t + Q_\eta \boldsymbol{\alpha}_{t-1}, && \text{when } t = T.
\end{aligned}$$

Block updating of all the $\boldsymbol{\alpha}_t$, $t = 1, \ldots, T$, can also be considered. However, this will mean storage and inversion of $(p \times T)$-dimensional matrices. Although the matrices will be structured band diagonal matrices, additional programming effort will be required to implement the block updating methods. Componentwise updating, as implemented here, will work fine when the states are not highly correlated.

Missing data, denoted by $Z^*(\mathbf{s}, t)$, are sampled at each MCMC iteration by using the full conditional distribution $N\{Y(\mathbf{s}, t), \sigma_\varepsilon^2\}$.

4.3. Forecasting

The posterior predictive distributions are used to make step ahead predictions (forecasts). The one-step-ahead forecast distribution is given by

$$\pi(\mathbf{z}_{T+1} | \mathbf{z}_1, \ldots, \mathbf{z}_T) = \int \pi(\mathbf{z}_{T+1} | \boldsymbol{\xi})\, \pi(\boldsymbol{\xi} | \mathbf{z}_1, \ldots, \mathbf{z}_T)\, \mathrm{d}\boldsymbol{\xi}, \tag{11}$$

where the likelihood term $\pi(\mathbf{z}_{T+1} | \boldsymbol{\xi})$ is obtained from the hierarchical model (1). The $E(\mathbf{Z}_{T+1} | \mathbf{z}_1, \ldots, \mathbf{z}_T)$ under density (11) provides the optimal one-step ahead forecast under a squared error loss function. To approximate $E(\mathbf{Z}_{T+1} | \mathbf{z}_1, \ldots, \mathbf{z}_T)$ we draw samples $\mathbf{Z}_{T+1}^{(j)}$ from $\pi(\mathbf{z}_{T+1} | \boldsymbol{\xi}^{(j)})$ and form the sample average. Other interesting summary measures, e.g. the 95% predictive intervals, are obtained by appropriately using the samples $\mathbf{z}_{T+1}^{(j)}$; see for example Gelfand (1996).

Suppose that we are not only interested in one-step-ahead predictions but also in L-step-ahead predictions where $L > 1$ is a positive integer. We obtain the predictive distribution (11), but here the dynamic parameters, e.g. the $\boldsymbol{\alpha}_t$, are first sampled from their distributions specified by the model; see equation (8). Using these forward values of the parameters we sample $\mathbf{Z}_{T+L}^{(j)}$ from the likelihood. These last samples are then averaged to obtain the estimated forecasts.

Throughout the paper we assume that the mean and variance of the L-step-ahead forecast distribution exist. This assumption is very reasonable in our set-up since we are primarily interested in making short-term forecasts. We can use other summary measures, e.g. the median if the means are not finite. Moreover, in such situations MCMC samples drawn from the forecast

distribution may drift to infinite values, thereby giving an early indication of problems. This may happen if the model is a very poor fit to the data. Some further checks on model validity should be performed before finally abandoning the current models in lieu of new ones.

The predictive distribution (11) is used to obtain simultaneous forecasts for all the monitored sites at any future time point $t > T$. Suppose that it is desired to predict the response at some unmonitored sites at any given time point t where t can be less than or equal to T. The methodology for obtaining the predictive distribution at one particular unmonitored site is given below; the extension for more than one site is straightforward and obvious.

To predict at an unmonitored site, $\mathbf{s}$ say, we use a predictive distribution like equation (11) with the following modifications to account for the spatial correlations between the responses at site $\mathbf{s}$ and at the monitored sites $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n$.

We first obtain the spatial covariance matrix Σ_γ^* of order $n+1$ by using the assumed covariogram (3), i.e.

$$\Sigma_\gamma^* = \begin{pmatrix} \Sigma_\gamma & \Sigma_{12}(\mathbf{s}) \\ \Sigma_{12}'(\mathbf{s}) & \sigma(\mathbf{s},\mathbf{s}) \end{pmatrix},$$

where $\Sigma_{12}(\mathbf{s})$ is the n-dimensional vector with elements $\sigma(\mathbf{s}_i, \mathbf{s})$, $i = 1, \ldots, n$. On the basis of the $n+1$ spatial locations $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n$ and $\mathbf{s}$ we derive the $(n+1) \times p$ matrix H^* by using equation (7) where we replace Σ_γ by Σ_γ^*. Let us partition the matrix H^* as

$$H^* = \begin{pmatrix} H_1^* \\ H_2^* \end{pmatrix}$$

where H_1^* is $n \times p$ and H_2^* is $1 \times p$. We now have that

$$\begin{pmatrix} \mathbf{Y}_t \\ Y(\mathbf{s},t) \end{pmatrix} \sim N(H^* \boldsymbol{\alpha}_t, \Sigma_\gamma^*)$$

by using the model assumption (2). From this multivariate normal distribution we obtain that

$$Y(\mathbf{s},t)|\boldsymbol{\xi} \sim N\{H_2^* \boldsymbol{\alpha}_t + \Sigma_{12}'(\mathbf{s})\Sigma_\gamma^{-1}(\mathbf{Y}_t - H_1^* \boldsymbol{\alpha}_t), \sigma(\mathbf{s},\mathbf{s}) - \Sigma_{12}'(\mathbf{s})\Sigma_\gamma^{-1}\Sigma_{12}(\mathbf{s})\} \tag{12}$$

by using standard methods. Now using model assumption (1) we have that

$$Z(\mathbf{s},t)|\boldsymbol{\xi} \sim N\{Y(\mathbf{s},t), \sigma_\varepsilon^2\},$$

where $Y(\mathbf{s},t)$ follows distribution (12) conditionally on $\boldsymbol{\xi}$. Now the predictive distribution at site $\mathbf{s}$ is given by

$$\pi\{z(\mathbf{s},t)|\mathbf{z}_1, \ldots, \mathbf{z}_T\} = \int \pi\{z(\mathbf{s},t)|\boldsymbol{\xi}\}\, \pi(\boldsymbol{\xi}|\mathbf{z}_1, \ldots, \mathbf{z}_T)\, \mathrm{d}\boldsymbol{\xi}. \tag{13}$$

If we were to forecast the smooth process $\mathbf{Y}_t$ at an unmonitored site $\mathbf{s}$, we use the conditional distribution of $Y(\mathbf{s},t)$ detailed in distribution (12). The average of samples drawn from this conditional distribution is the estimated forecast of the smooth process $\mathbf{Y}_t$ at site $\mathbf{s}$.

If new data were available we can rerun the entire MCMC implementation and predict observations which are future in time. However, there are many approximation methods using importance sampling which can be used as well; see for example Irwin *et al.* (2002) for details.

4.4. Assessing the forecasts

Many graphical diagnostic methods are used to perform diagnostic checking and model validation; see for example Mardia *et al.* (1998). Several validation statistics are also available; see

for example Carroll and Cressie (1996). They make use of the three statistics

$$\mathrm{CR}_1(s_j) = \frac{(1/L)\sum_{t=T+1}^{T+L}\{Z(\mathbf{s}_j,t)-\hat{Z}(\mathbf{s}_j,t)\}}{(1/L)\left\{\sum_{t=T+1}^{T+L}\hat{\sigma}^2_Z(\mathbf{s}_j,t)\right\}^{1/2}},$$

$$\mathrm{CR}_2(s_j) = \left[\frac{(1/L)\sum_{t=T+1}^{T+L}\{Z(\mathbf{s}_j,t)-\hat{Z}(\mathbf{s}_j,t)\}^2}{(1/L)\sum_{t=T+1}^{T+L}\hat{\sigma}^2_Z(\mathbf{s}_j,t)}\right]^{1/2},$$

$$\mathrm{CR}_3(s_j) = \left[(1/L)\sum_{t=T+1}^{T+L}\{Z(\mathbf{s}_j,t)-\hat{Z}(\mathbf{s}_j,t)\}^2\right]^{1/2},$$

where $\hat{Z}(\mathbf{s}_j,t)$ is the prediction of $Z(\mathbf{s}_j,t)$ and $\hat{\sigma}^2_Z(\mathbf{s}_j,t)$ is the mean-square prediction error. Then it is recommended that summary statistics be used to compare the models; for example one may find the means of the above three statistics. When forecasts are accurate, the means of $\mathrm{CR}_1(s_j)$ and $\mathrm{CR}_2(s_j)$ should be close to 0 and 1 respectively; the mean of $\mathrm{CR}_3(s_j)$ provides a 'goodness of prediction' and it is expected to be small when predicted values are close to the true values.

The forecasts $\hat{Z}(\mathbf{s}_j,t)$, for $t=T+1,\ldots,T+L$, depend on one another and this fact is ignored when summary statistics are formed from the time-averaged statistics $\mathrm{CR}_1(s_j)$, $\mathrm{CR}_2(s_j)$ and $\mathrm{CR}_3(s_j)$. To overcome this we adopt the weighted distance between the forecasts and the actual observations. Let

$$\mathbf{V} = \begin{pmatrix} \mathbf{Z}_{T+1} \\ \vdots \\ \mathbf{Z}_{T+L} \end{pmatrix}$$

denote the set of observations for which we seek validation. Note that we have observed data $\mathbf{Z}_1,\ldots,\mathbf{Z}_{T+L}$ but we have used only $\mathbf{Z}_1,\ldots,\mathbf{Z}_T$ to fit the model and to obtain the validation forecast for $\mathbf{V}$. Let $\mathbf{v}_{\mathrm{obs}}$ denote the observed data.

Using the implemented MCMC algorithm we draw $\mathbf{V}^{(j)}$, $j=1,\ldots,E$ (where E is a large positive integer), samples from the forecast distribution $\pi(\mathbf{v}|\mathbf{z}_1,\ldots,\mathbf{z}_T)$. The first paragraph in Section 4.3 details how to draw these samples. Now

$$\bar{\mathbf{V}} = \frac{1}{E}\sum_{j=1}^{E}\mathbf{V}^{(j)},$$

$$\hat{\Sigma} = \frac{1}{E-1}\sum_{j=1}^{E}(\mathbf{V}^{(j)}-\bar{\mathbf{V}})(\mathbf{V}^{(j)}-\bar{\mathbf{V}})'$$

unbiasedly estimate the mean vector and the covariance matrix of the forecast distribution $\pi(\mathbf{v}|\mathbf{z}_1,\ldots,\mathbf{z}_T)$ respectively. The ergodicity properties of the MCMC simulation algorithms guarantee that these estimates converge to the true mean and covariance matrix of the forecast distribution when E is large.

Under suitable regularity conditions which guarantee asymptotic normality and for small values of L, the predictive distribution $\pi(\mathbf{v}|\mathbf{z}_1,\ldots,\mathbf{z}_T)$ can be approximated by the nL-dimensional normal distribution with mean $\bar{\mathbf{V}}$ and covariance matrix $\hat{\Sigma}$. Using well-known properties

of the multivariate normal distribution, we have

$$D^2 = (\mathbf{V} - \bar{\mathbf{V}})' \hat{\Sigma}^{-1} (\mathbf{V} - \bar{\mathbf{V}}) \sim \chi^2_{nL}, \qquad \text{approximately.} \tag{14}$$

The approximation arises because $\mathbf{V}$ is only approximately multivariate normal for small values of L for short-term forecasting. A numerical justification for this approximation is provided in Section 5.3.

Our proposed validation statistic is the observed value of D^2, given by

$$D^2_{\text{obs}} = (\mathbf{v}_{\text{obs}} - \bar{\mathbf{V}})' \hat{\Sigma}^{-1} (\mathbf{v}_{\text{obs}} - \bar{\mathbf{V}}). \tag{15}$$

Clearly, D^2_{obs} will increase if there are large discrepancies between the forecast based on the model, $\bar{\mathbf{V}}$, and the observed data, $\mathbf{v}_{\text{obs}}$. Thus D^2_{obs} can be referred to the theoretical values of the χ^2-distribution with nL degrees of freedom. Note also that D^2_{obs} is the Mahalanobis distance when the distributions of $\mathbf{V}_{\text{obs}}$ and $\bar{\mathbf{V}}$ have the common covariance matrix $\hat{\Sigma}$.

Table 1. Values of the predictive model choice criterion for various values of p and λ

p	*Results for the following values of* λ*:*		
	0.3	*0.4*	*0.5*
4	125.3	125.8	128.8
5	120.3	116.9	117.1
6	117.4	112.3	114.4
7	100.8	96.7	97.8
8	109.7	103.5	104.2

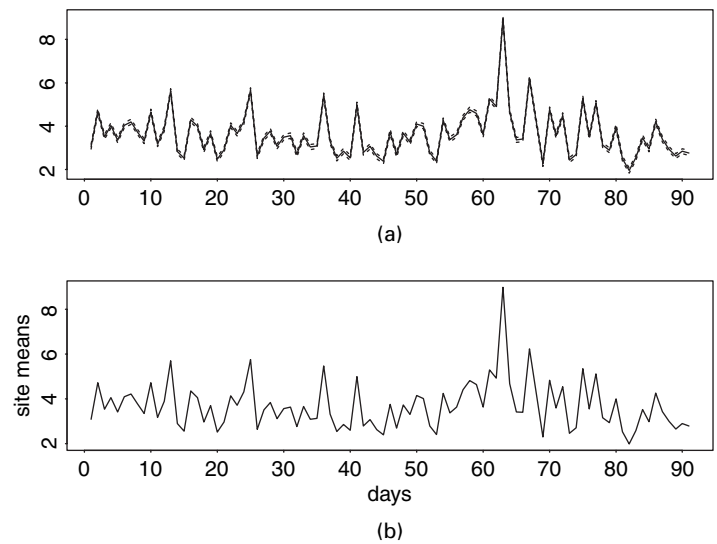


Fig. 6. (a) Marginal posterior means and 95% credible intervals of α_{t1} and (b) mean observed time series: the time unit is 3 days

5. The New York City data example

5.1. Model choice

We return to the example that was discussed in Section 2. We first choose the parameters p and λ by using the following well-known predictive model choice criterion (see for example Laud and Ibrahim (1995)):

$$\text{PMCC} = \sum ([Z(\mathbf{s},t)_{\text{obs}} - E\{Z(\mathbf{s},t)_{\text{rep}}\}]^2 + \text{var}\{Z(\mathbf{s},t)_{\text{rep}}\}),$$

where the summation is taken over all the nT observations except for the missing observations and $Z(\mathbf{s},t)_{\text{rep}}$ is a future observation corresponding to $Z(\mathbf{s},t)$ under the model assumed. The estimated values of PMCC are reported in Table 1. The model with $p=7$ and $\lambda=0.4$ is seen to be the best model and henceforth we work with this model. Table 1 also shows that the model choice criterion is not greatly sensitive to the choice of λ among the values that are considered. We have also computed the model choice criterion for $\lambda=0.2$ and $\lambda=0.1$. For those values the criterion values were higher than the values corresponding to each value of p reported in Table 1.

The chosen value of $\lambda=0.4$ corresponds to an approximate range of 10 miles in spatial dependence since the covariogram decays to 0.05 for $\lambda=0.4$ and $d=10$. The choice of $p=7$ is seen to

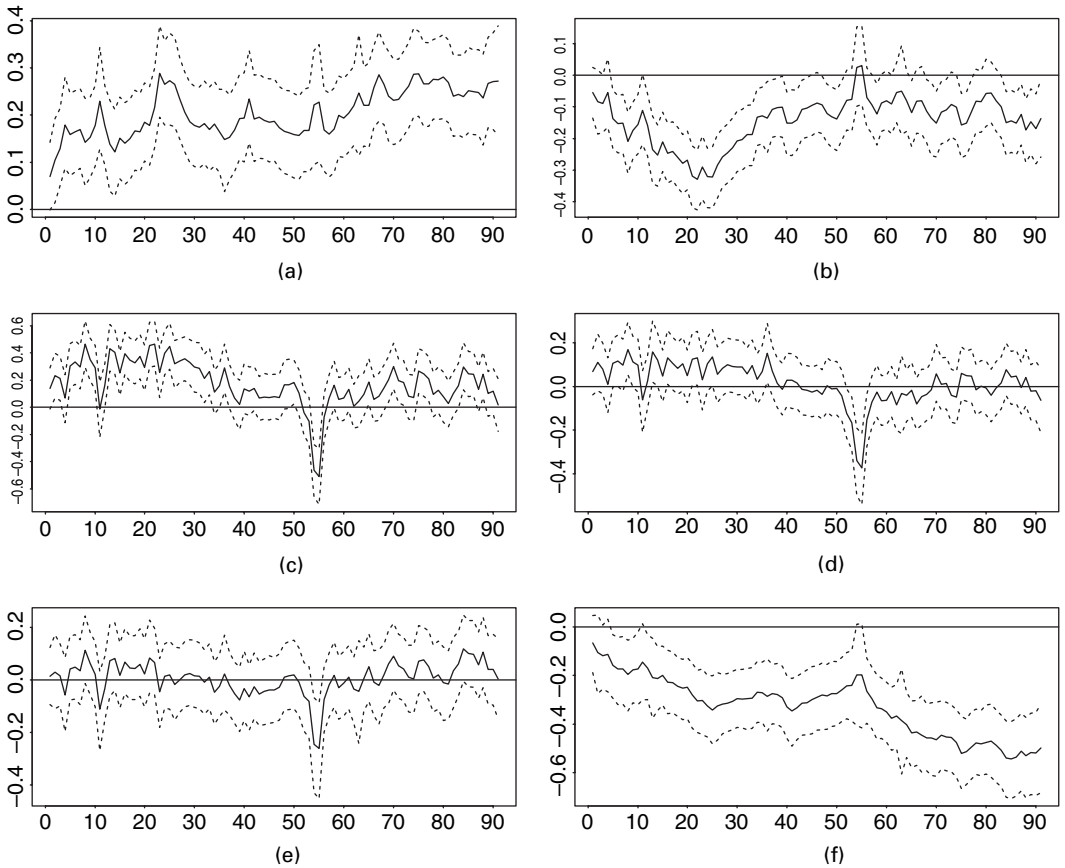


Fig. 7. Marginal posterior means and 95% credible intervals of α_{ti} for (a) $i=2$, (b) $i=3$, (c) $i=4$, (d) $i=5$, (e) $i=6$ and (f) $i=7$: the horizontal line at 0.0 has been superimposed to see the significance of the states; the time unit is 3 days

be about half the maximum number of principal fields possible. We shall further examine the choice by monitoring the components of $\boldsymbol{\alpha}$ for this optimal model.

5.2. Analysis

The estimates of σ_ε^2 and σ_γ^2 under the model chosen are 0.0356 and 0.0172 respectively. The standard deviations are estimated to be 0.0032 and 0.0046 respectively. The MCMC chains for these two parameters were monitored to detect possible problems in convergence. However, no such problems were found in the current implementation.

We plot the MCMC estimates of α_{t1} for all values of t along with the 95% credible intervals in Fig. 6. Since the first column of the matrix H is a unit vector, α_{t1} will estimate the mean of the time series that is observed at the different sites. To see this we plot the mean time series that is obtained by averaging the response from all the sites in Fig. 6(b). As expected the plots in Figs 6(a) and 6(b) look virtually the same. This justifies our previous claim that model (8) captures the main temporal structures in the data.

The plots of the remaining six components of $\boldsymbol{\alpha}_t$ along with their 95% credible intervals appear in Fig. 7. In Fig. 7 we have also plotted a horizontal line at zero to see the significance of the α_{ti}, $i = 2, \ldots, 7$, for the entire range of t. The second and third components α_{t2} and α_{t3}

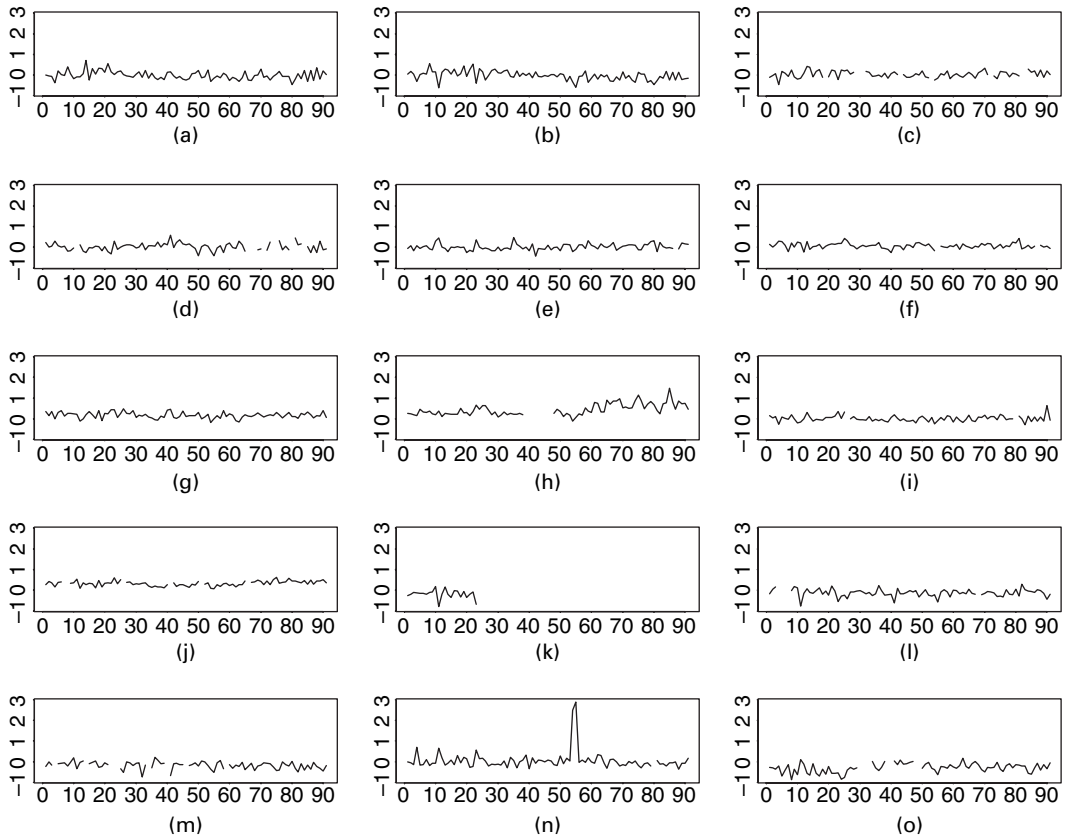


Fig. 8. Time series plots of the residuals from (a) site 1, (b) site 2, (c) site 3, (d) site 4, (e) site 5, (f) site 6, (g) site 7, (h) site 8, (i) site 9, (j) site 10, (k) site 11, (l) site 12, (m) site 13, (n) site 14 and (o) site 15: the time unit is 3 days

are seen to be significant for all values of t. The remaining four components are significant at different times but are not significant for all values of t. The two components α_{t5} and α_{t6} are significant for only a few values of t. Fig. 7 also shows that none of the seven components of $\boldsymbol{\alpha}_t$ can be removed to obtain a more parsimonious model as all the components are significant at least for some values of t.

The time series plots of the raw residuals, the differences between the observed and the fitted, are given in Fig. 8. As expected, the residual plots do not show any spatial or temporal patterns. The plot for site 14, however, shows high residual values for June 10th and 13th. As mentioned

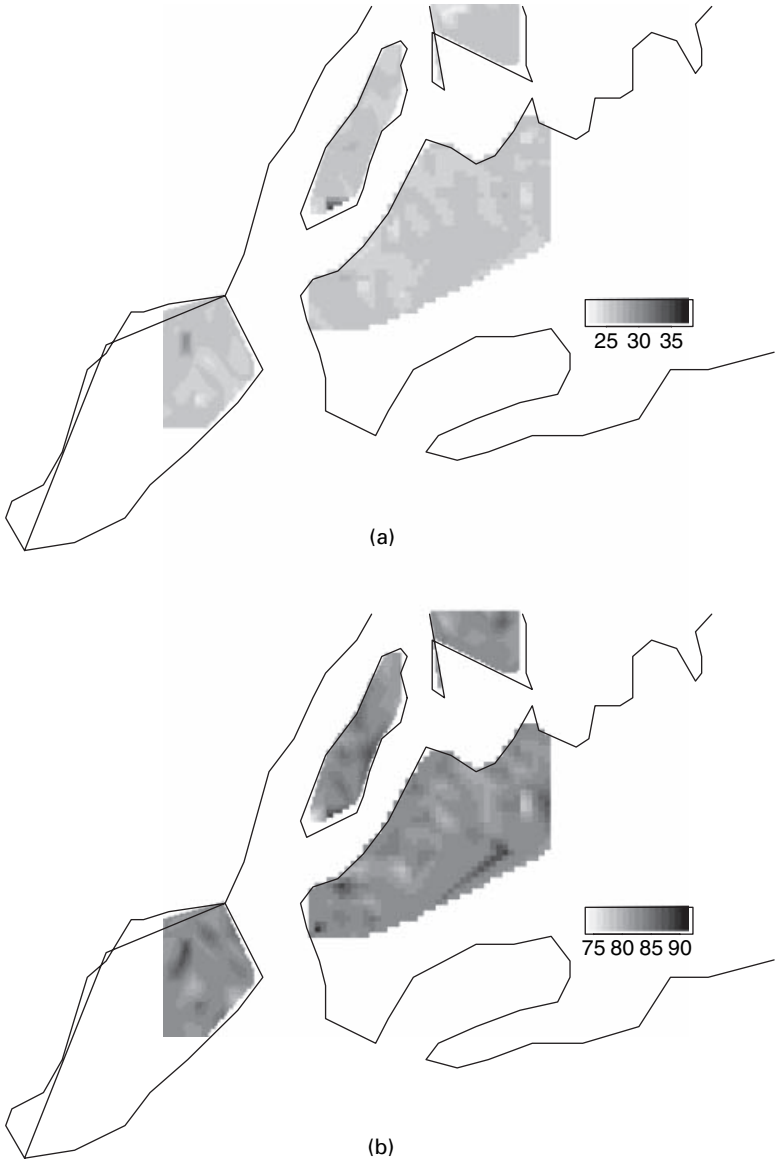
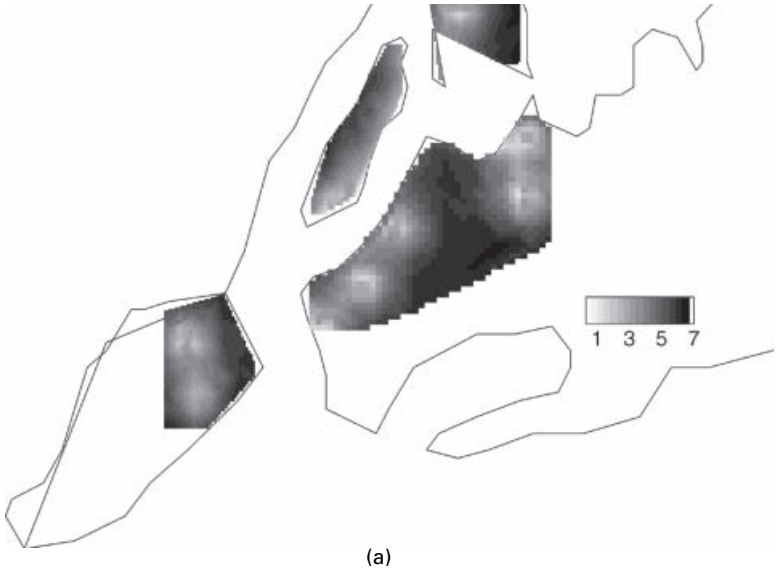


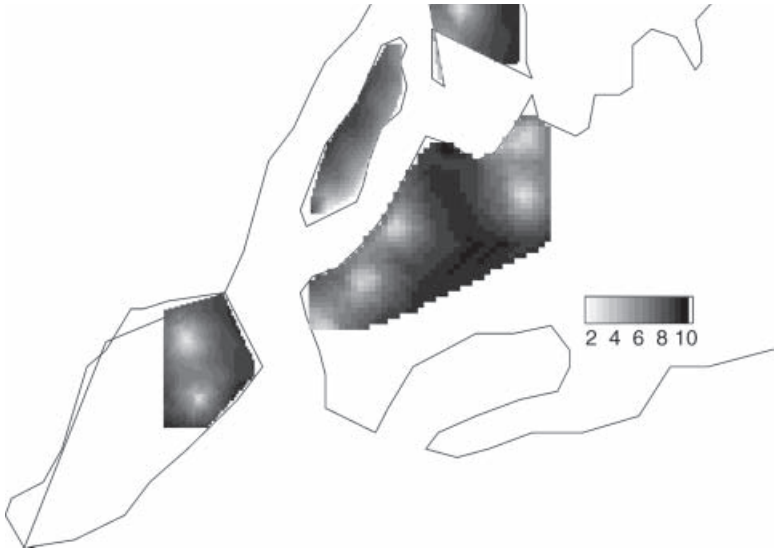
Fig. 9. Model predicted maps for (a) July 4th and (b) July 7th: these predictions should be compared with the observed data plotted in Fig. 4

previously in Section 2 these two observations are outliers and consequently the fitted model shows some lack of fit for these two observations. We have also examined the variogram of the fitted values as was done for the data in Fig. 3 and this looked very similar to Fig. 3. This is expected since the model provides a very good fit to the data, as suggested by the above residual plots.

We now return to the peculiarity of the data as plotted in Fig. 4. We spatially predict the level of the response on 625 locations on land for July 4th and 7th. Note that these are spatial predictions and are not temporal forecasts. Moreover, no cross-validation is done here. We use all the data for model fitting and then we predict at the new locations using the Bayesian predic-



(a)



(b)

Fig. 10. Standard deviation of the predicted maps for (a) July 4th and (b) July 7th

tive distribution (13). Note that we require the matrix H^* to obtain this predictive distribution. Here we first obtain the matrix H^* (640×7) for all the 640 sites (15 monitoring sites and 625 locations for predictions) and then use H_1^* (15×7) for model fitting and use H_2^* (625×7) for prediction.

The two spatial prediction surfaces each with 640 predictions (at the 15 monitored and 625 unmonitored sites) are linearly interpolated and then plotted in Fig. 9. The plot for July 4th shows two hot spots, one each in Manhattan and in Staten Island. These two hot spots also remain on July 7th, but more hot spots emerge on July 7th possibly because of the after-effect of the July 4th firework celebrations. This reinforces the fact that there are different spatial patterns at different locations and at different time points. A comparison between these and the data plots in Fig. 4 shows that there is very good agreement between the model predictions and the observed data.

The standard deviations of the predictions are plotted in Fig. 10. The standard deviations are smaller for the locations which are near the observed sites. As expected a good predictor should be able to predict better for the sites which are close to the observation sites than the sites which are far away.

Why is the prediction map for July 4th much lighter than the same for July 7th? This is explained by the two different types of variation in the data for two days; see Fig. 5. The data for July 4th have a long left-hand tail whereas the data for July 7th have a long right-hand tail. The small data values in the long left-hand tail have influenced the predicted surface for July 4th to be lighter in colour, and the large data values in the long right-hand tail have influenced the surface for July 7th to be darker.

5.3. Cross-validation

We examine the cross-validation statistic D^2 that was proposed in Section 4.4. A referee has expressed concern regarding the asymptotic normal approximation and hence the asymptotic χ^2 -approximation for D^2 that is given in expression (14). We address the concern as follows. We only consider cross-validation for one and two time steps in advance since the main motivation here is short-term forecasting of spatiotemporal processes. For the one-step-ahead predictions D^2 will be approximately χ^2 distributed with 15 degrees of freedom and for the two-step-ahead predictions D^2 will have 30 degrees of freedom approximately.

We estimate $\bar{\mathbf{V}}$ and $\hat{\Sigma}$ by using 10000 MCMC samples from the predictive distributions of the one- and two-step-ahead predictions. Subsequently, we draw 1000 independent random samples, $V^{(j)}$, $j = 1, \ldots, 1000$, from the corresponding predictive distributions and form the statistic D^2 in each case. Note that the samples are not drawn from the approximate multivariate normal distribution.

The histogram of the 1000 D^2-values and the density of the theoretical χ^2-distributions are plotted in Fig. 11. Figs 11(a) and 11(b) show that the data histogram in each case is a very good approximation for the corresponding theoretical χ^2-distribution. Moreover, to see the goodness of fit we run the Kolmogorov–Smirnov goodness-of-fit test using the 1000 simulated values. The p-value of the test is 0.85 for the one-step-ahead prediction and 0.21 for the two-step-ahead predictions. These high p-values indicate that the distributions of the observed D^2-values can be taken to be the corresponding theoretical χ^2-distributions as claimed in distribution (14).

Now we evaluate the forecasting performance of the model by using D^2_{obs} as given in equation (15). Using the current model the D^2_{obs}-values are 17.7 with 15 degrees of freedom for the one-step-ahead forecasts and 37.9 with 30 degrees of freedom for the two-step-ahead forecasts. These values clearly indicate that the model is forecasting the data well.

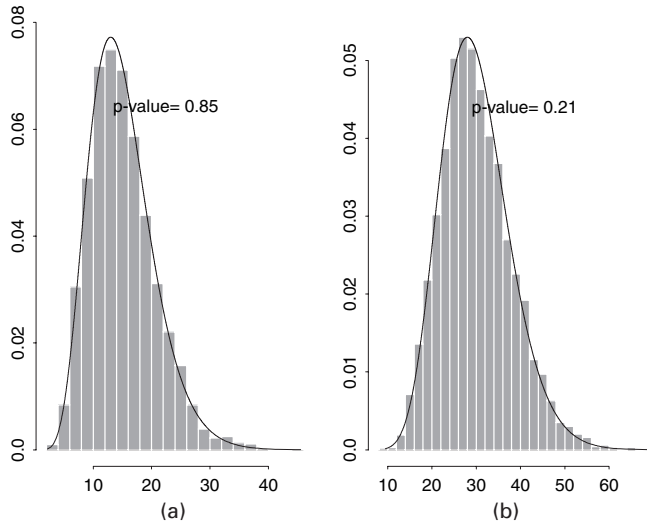


Fig. 11. χ^2-approximation of D^2 for (a) the one-step-ahead forecasts and (b) the two-step-ahead forecasts: the p-values are those of the Kolmogorov–Smirnov goodness-of-fit test

6. Discussion

We have proposed a Bayesian model for analysing spatiotemporal data. The model has been implemented in a full Bayesian set-up using MCMC sampling. We have implemented the models in a simulation example (documented in an unpublished technical report version of the current paper by the same authors) which validated our MCMC code. However, for brevity we do not present the example here.

The principal kriging functions that were used in the model that we proposed are basis functions which are optimal for spatial predictions alone. The comparative models using polynomial-type regressors do not use these optimal functions and hence may provide less accurate forecasts especially for extrapolation.

We have applied our model on the air pollution data, and using new cross-validation methods we have shown that the model is adequate for short-term forecasting. Our use of Bayesian predictive densities for spatial predictions makes our method optimal in the sense of Wikle and Cressie (1999). The models proposed work even when the numbers of sites are moderately large, although as expected the computations become more intensive as the number of sites increases. The well-known advantages of the fully implemented MCMC methods, however, justify their use for small to moderate data sets.

Acknowledgements

The authors thank John Kent and Richard Smith for helpful discussions. They thank David Holland of the US Environmental Protection Agency for providing the data; they also thank the Joint Editor, an Associate Editor and two referees for many helpful comments and suggestions.

References

Allcroft, D. J. and Glasbey, C. A. (2003) A latent Gaussian Markov random-field model for spatiotemporal rainfall disaggregation. *Appl. Statist.*, **52**, 487–498.

Berger, J. O., de Oliveira, V. and Sansó, B. (2001) Objective Bayesian analysis of spatially correlated data. *J. Am. Statist. Ass.*, **96**, 1361–1374.
Bookstein, F. L. (1989) Principal warps: thin-plate splines and the decomposition of deformations. *IEEE Trans. Pattn Anal. Mach. Intell.*, **11**, 567–585.
Brown, P. E., Diggle, P. J., Lord, M. E. and Young, P. C. (2001) Space–time calibration of radar rainfall data. *Appl. Statist.*, **50**, 221–241.
Carroll, S. S. and Cressie, N. (1996) A comparison of geostatistical methodologies used to estimate snow water equivalent. *Wat. Resour. Bull.*, **32**, 267–278.
Carter, C. and Kohn, R. (1994) On Gibbs sampling for state space models. *Biometrika*, **81**, 541–553.
Cressie, N. (1994) Comment on "An approach to statistical spatial-temporal modeling of meteorological fields" by M. S. Handcock and J. R. Wallis. *J. Am. Statist. Ass.*, **89**, 379–382.
Dye, T., Miller, D. and MacDonald, C. (2002) Summary of $PM_{2.5}$ forecasting program development and operations for Salt Lake City, Utah during winter 2002. *Technical Report*. Sonoma Technology, Petaluma.
Ecker, M. D. and Gelfand, A. E. (1997) Bayesian variogram modeling for an isotropic spatial process. *J. Agric. Biol. Environ. Statist.*, **2**, 607–617.
Gelfand, A. E. (1996) Model determination using sampling based methods. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter), pp. 145–161. London: Chapman and Hall.
Gelfand, A. E., Banerjee, S. and Gamerman, D. (2004) Spatial process modelling for univariate and multivariate dynamic spatial data. *Environmetrics*, to be published.
Gelfand, A. E. and Sahu, S. K. (1999) Identifiability, improper priors, and Gibbs sampling for generalized linear models. *J. Am. Statist. Ass.*, **94**, 247–253.
Gelfand, A. E., Sahu, S. K. and Carlin, B. P. (1995) Efficient parametrization for normal linear mixed models. *Biometrika*, **82**, 479–488.
Goodall, C. and Mardia, K. V. (1994) Challenges in multivariate spatio-temporal modeling. In *Proc. 17th Int. Biometric Conf., Hamilton, Aug. 8th–12th*, pp. 1–17. Hamilton: McMaster University Press.
Irwin, M. E., Cressie, N. and Johannesson, G. (2002) Spatial-temporal non-linear filtering based on hierarchical statistical models (with discussion). *Test*, **11**, 249–302.
Kent, J. T. and Mardia, K. V. (1994) The link between Kriging and thin-plate splines. In *Probability, Statistics and Optimisation* (ed. F. P. Kelly), pp. 324–329. New York: Wiley.
Kent, J. T. and Mardia, K. V. (2002) Modelling strategies for spatial-temporal data. In *Spatial Cluster Modelling* (eds A. Lawson and D. Denison), pp. 214–226. London: Chapman and Hall.
Kyriakidis, P. C. and Journel, A. G. (1999) Geostatistical space-time models: a review. *Math. Geol.*, **31**, 651–684.
Laud, P. W. and Ibrahim, J. G. (1995) Predictive model selection. *J. R. Statist. Soc.* B, **57**, 247–262.
Mardia, K. V. and Goodall, C. (1993) Spatial-temporal analysis of multivariate environmental monitoring data. In *Multivariate Environmental Statistics* (eds G. P. Patil and C. R. Rao), pp. 347–386. Amsterdam: Elsevier.
Mardia, K. V., Goodall, C., Redfern, E. J. and Alonso, F. J. (1998) The Kriged Kalman filter (with discussion). *Test*, **7**, 217–252.
Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979) *Multivariate Analysis*. London: Academic Press.
Matérn, B. (1986) *Spatial Variation*. Berlin: Springer.
Sansó, B. and Guenni, L. (1999) Venezuelan rainfall data analysed by using a Bayesian space–time model. *Appl. Statist.*, **48**, 345–362.
Sansó, B. and Guenni, L. (2000) A nonstationary multisite model for rainfall. *J. Am. Statist. Ass.*, **95**, 1089–1100.
Smith, R. L., Kolenikov, S. and Cox, L. H. (2003) Spatio-temporal modelling of $PM_{2.5}$ data with missing values. *J. Geophys. Res. Atmos.*, **108**.
Stroud, J. R., Müller, P. and Sansó, B. (2001) Dynamic models for spatiotemporal data. *J. R. Statist. Soc.* B, **63**, 673–689.
West, M. and Harrison, J. (1997) *Bayesian Forecasting and Dynamic Models*. New York: Springer.
Wikle, C. K., Berliner, L. M. and Cressie, N. (1998) Hierarchical Bayesian space-time models. *Environ. Ecol. Statist.*, **5**, 117–154.
Wikle, C. K. and Cressie, N. (1999) A dimension-reduced approach to space-time Kalman filtering. *Biometrika*, **86**, 815–829.