Considering covariates in the covariance structure of spatial processes

Alexandra M. Schmidt*
Universidade Federal do Rio de Janeiro, Brazil
Peter Guttorp
University of Washington, USA and
Norwegian Computing Center, Oslo, Norway
Anthony O'Hagan
University of Sheffield, UK

December 2010

Abstract

In spatial statistics one usually assumes that observations are partial realizations of a stochastic process $\{Y(\mathbf{x}), \mathbf{x} \in \mathbb{R}^C\}$, where commonly $C = 2$, and the components of the location vector $\mathbf{x}$ are geographical coordinates. Frequently, it is assumed that $Y(\cdot)$ follows a Gaussian process (GP) with stationary covariance structure. In this setting the usual aim is to make spatial interpolation to unobserved locations of interest, based on observed values at monitored locations. This interpolation is heavily based on the specification of the mean and covariance structure of the GP. In environmental problems the assumption of stationary covariance structures is commonly violated due to local influences in the covariance structure of the process.

We propose models which relax the assumption of stationary GP by accounting for covariate information in the covariance structure of the process. Usually at each location $\mathbf{x}$, covariates related to $Y(\cdot)$ are also observed. We initially propose the use of covariates to allow the latent space model of Sampson & Guttorp to be of dimension $C > 2$. Then we discuss a particular case of the latent space model by using a representation projected down from C dimensions to 2 in order to model the 2D correlation structure better. Inference is performed under the Bayesian paradigm, and Markov chain Monte Carlo methods are used to obtain samples from the resultant posterior distributions under each model. As illustration of the proposed models, we analyze solar radiation in British Columbia, and mean temperature in Colorado.

Key Words: Anisotropy; Deformation; Manifold; Non-stationarity; Projection.

*Address for correspondence: Alexandra M. Schmidt, Departamento de Métodos Estatsticos, Universidade Federal do Rio de Janeiro, Caixa Postal 68530, Rio de Janeiro, RJ, Brazil. CEP 21945-970. *Tel.:* 0055 21 2562-7505 x 204. *Fax:* 0055 21 2562-7374.
E-mail: alex@im.ufrj.br. *Homepage*: www.dme.ufrj.br/∼alex

1 Introduction

Spatial statistics has been receiving a lot of attention in the last two decades. This is related to our ability to store complex datasets, which may often be spatially indexed. Generally, one assumes that $\{Y(\mathbf{x}), \mathbf{x} \in \mathbb{R}^2\}$ is a stochastic process, often taken to be Gaussian. This assumption is quite convenient, as all we need to specify are the mean and covariance functions of a stochastic process. In geostatistics it is common practice to assume that the stochastic process is stationary, which means that the distribution is unchanged when the origin of the index set is translated, and isotropic, that is, the process is invariant under rotations about the origin. Frequently, the covariance function is modelled as the product of a common variance and a valid correlation function, and the correlation structure is commonly assumed to be stationary and isotropic, i.e. a function of the Euclidean distance between locations. There are in the literature many different correlation functions which lead to valid covariance structures. See Cressie (1993), Gneiting (2002) and Banerjee et al. (2004) for examples of such functions. However, Sampson and Guttorp (1992) point out, in the analysis of most spatio-temporal processes underlying environmental studies, there is little reason to expect spatial covariance structures to be stationary over the spatial scales of interest because there may be local influences in the correlation structure of the spatial random process.

Sampson and Guttorp (1992) (S&G hereafter) were among the first to propose a model that relaxes the assumption of isotropy and stationarity. Their idea involves a latent space called D-space, and is based on a nonlinear transformation of the sampling space (which is called G-space) into D-space, within which the spatial structure is stationary and isotropic. Schmidt and O'Hagan (2003) (S&O hereafter) proposed a Bayesian approach to this idea of modelling the covariance structure as a function of the locations in a latent space. Their main contribution is to propose a model whose parameters are estimated in a single framework. More specifically, the unknown function $\mathbf{d}(\cdot)$ that maps the locations from G-space into D-space follows a Gaussian process, *a priori*. Some examples that implement the latent space approach are e.g. Meiring et al. (1998); Le et al. (2001); Sampson et al. (2001); Damian et al. (2003); Guttorp et al. (2007).

In the last 10 years many alternatives have been proposed to the S&G approach. The most successful ones are those based on convolution. Higdon (1998) was the first to propose a moving average convolution approach, based on the fact that any Gaussian process with a specific correlation function can be represented as a convolution between a kernel and a white noise process. Allowing the kernel to vary smoothly across locations results in a valid nonstationary covariance structure. Fuentes and Smith (2000) proposed an alternative approach to that of Higdon (1998). Instead of making the kernel vary, they assumed that the spatial process is a convolution between a fixed kernel and independent Gaussian processes whose parameters are allowed to vary across locations. See Banerjee et al. (2004) for a discussion about these two approaches.

More recently, Paciorek and Schervish (2006) generalize the kernel convolution approach of Higdon (1998). Using a Gaussian kernel, whose covariances vary with

location, they obtain a general form of the covariance function, and show that this can be generalized to correlation functions that are positive definite in Euclidean space of every dimension. Kim et al. (2005) propose a different approach, to decompose the spatial domain into disjoint regions within which the process is assumed to be stationary. This decomposition is done through the use of the Voronoi tesselation. Sampson (2010) discusses constructions for nonstationary spatial processes focusing on the approaches mentioned above.

Let $Y(\mathbf{x}, t)$ denote the value of the process at location $\mathbf{x} = (x_1, x_2)'$ and time t, for $t = 1, 2, \cdots$. Usually, monitoring networks of environmental processes collect information on many different variables of interest. We assume that at each location $\mathbf{x}$ a vector of covariates, say $\mathbf{Z}(\mathbf{x})$, is also observed. Let G-space represent the space defined by the geographical coordinates $\mathbf{x}$, such that $G \subset \mathbb{R}^2$. It is common practice to define the process $Y(\cdot, t)$ in G-space and include the effect of covariates only in its mean structure, frequently assuming a linear relationship between $y(\mathbf{x}, t)$ and covariates $\mathbf{z}(\mathbf{x})$. See Cressie (1993) and Banerjee et al. (2004) for examples. There are also many alternatives which consider nonlinear structures in the mean structure of the process (e.g. Guttorp et al. (2007)). We aim to discuss here how the effect of covariates might be also considered in the covariance structure of a spatial process.

Throughout the paper we let $Y_{it} = Y(\mathbf{x}_i, t)$ for $i = 1, 2, \cdots, n$ and $t = 1, 2, \cdots, T$. Let $\mathbf{Y}_t = (Y_{1t}, Y_{2t}, \cdots, Y_{nt})^T$ for $t = 1, \cdots, T$. We suppose that $\mathbf{Y}_1, \cdots, \mathbf{Y}_T$ are independently distributed with density $N_n(\boldsymbol{\mu}_t, \boldsymbol{\Sigma})$, where $N_n(\boldsymbol{\mu}_t, \boldsymbol{\Sigma})$ stands for the multivariate normal distribution of dimension n with mean vector $\boldsymbol{\mu}_t$ and covariance matrix $\boldsymbol{\Sigma}$. For simplicity of exposition, Section 2 concentrates on modelling the covariance structure $\boldsymbol{\Sigma}$ by assuming $\boldsymbol{\mu}_t = \boldsymbol{\mu}\ \forall t$, and accommodating uniform priors for $\boldsymbol{\mu}$. In particular, this section extends the latent space idea of S&G and S&O for mappings from $\mathbb{R}^2$ onto $\mathbb{R}^C$ $(C > 2)$, such that the spatial stochastic process is defined on a C-dimensional space. Then section 3 discusses a particular case of the model introduced in section 2, by using a representation projected down from C dimensions to 2 in order to model the 2D correlation structure better. Because the covariance structure of this model is simpler than the general one proposed in section 2, therein the mean and covariance structures are jointly estimated. Next section illustrates the performance of the proposed models in analyzing two datasets, solar radiation from British Columbia and mean temperature from Colorado. Finally, Section 5 discusses the advantages and disadvantages of the proposed models and points to future avenues of research.

2 Introducing covariates in the latent space approach

2.1 Modelling the data and correlation structure

In this section we assume it is required to make inference only about $\boldsymbol{\Sigma}$. In this particular case, the data yield an $n \times n$ sample covariance matrix $\mathbf{S}$, obtained from data at n spatial locations $\mathbf{x}_1, \cdots, \mathbf{x}_n$ over T time points. Here we assume $\mathbf{Y}_t \sim$

$N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. After integrating out $\boldsymbol{\mu}$ using a uniform prior, the likelihood for $\boldsymbol{\Sigma}$ has a Wishart form

$$p(\mathbf{S} \mid \boldsymbol{\Sigma}) \propto \mid \boldsymbol{\Sigma} \mid^{-\frac{T-1}{2}} \exp\left\{-\frac{T}{2} \text{ tr } \mathbf{S}\boldsymbol{\Sigma}^{-1}\right\}.$$

Notice further that when computing $\mathbf{S}$ we assume that the temporal effect has been previously removed. The main target here is to model the spatial nonstationarity present in the data. Each element of $\boldsymbol{\Sigma}$ is modelled through

$$\text{Cov } (Y(\mathbf{x}_i, t), Y(\mathbf{x}_j, t)) = \sqrt{v(\mathbf{x}_i)v(\mathbf{x}_j)}\; g(||\mathbf{d}(\mathbf{x}) - \mathbf{d}(\mathbf{x}')||), \tag{2.1}$$

where for all t, $v(\mathbf{x}) = \text{ Var } (Y(\mathbf{x}, t))$, and $g(\cdot)$ denotes the correlation function of the spatial process as a function of the Euclidean distance between locations in D-space. The variances are assumed exchangeable *a priori*. More specifically, *a priori*, the variances $v(\mathbf{x})$ have inverse gamma distributions with mean τ^2 and f degrees of freedom, that is,

$$\begin{aligned} v(\mathbf{x}) \mid \tau^2 &\sim IG(\tau^2(f-2), f) \quad \forall \mathbf{x} \in G \\ \pi(\tau^2) &\propto \tau^{-2}, \end{aligned} \tag{2.2}$$

so that f is fixed and the mean τ^2 has a vague prior distribution. Here, if $V \sim IG(a, d)$ then $f(v) = \frac{(a/2)^{d/2}}{\Gamma(d/2)} v^{-(d+2)/2} \exp\left\{-\frac{a}{2v}\right\}$, $v > 0$.

S&O concentrates on mappings from $\mathbb{R}^2$ onto $\mathbb{R}^2$. They assume, *a priori*, that the function $\mathbf{d}(\mathbf{x}) = (d_1(\mathbf{x}), \cdots, d_C(\mathbf{x}))'$ follows a Gaussian process prior, such that $\mathbf{d}(\cdot) \sim GP(\mathbf{m}(\cdot), \boldsymbol{\sigma}_d^2 R_d(\cdot, \cdot))$. A potential problem noted in the original S&G work is that the mapping may fold so that two different points in G-space map into the same point in D-space, with the undesirable result that these points become perfectly correlated. The smoothness property of the Gaussian process prior tends to avoid folding but cannot ensure that it will not happen. Although the multivariate normal distribution is unimodal and the uncertainty about the mapping can be controlled through the specification of the prior covariance structure of $\mathbf{d}(\cdot)$ there might be other aspects which influence the spatial process. Considering the approaches in the literature that use the latent space idea, the only one that guarantees non-folding of the mapping is that by Iovleff and Perrin (2004). In their approach the function which maps locations into D-space is guaranteed to be bijective because of the use of the Delaunay triangulation, and so it cannot fold. On the other hand, Monestiez and Switzer (1991) analyse acid rainfall data for which they have tried a mapping into a space of dimension 3. This was done because their first attempt in fitting a model in a space of dimension 2 resulted in folding of their mapping function.

Modeling the spatial correlation $g(\cdot)$ in $\mathbb{R}^C$ When considering the D-space of dimension $C > 2$, one has to ensure the validity of the chosen isotropic covariogram model in $\mathbb{R}^C$. Like in S&O, the function $g(\cdot)$ might be modelled as

$$g(\delta) = \sum_{k=1}^{K} \alpha_k \exp\left\{-\lambda_k \delta^2\right\}, \tag{2.3}$$

where, $\delta = || \mathbf{d}(\mathbf{x}) - \mathbf{d}(\mathbf{x}') || = \sqrt{(d_1(\mathbf{x}) - d_1(\mathbf{x}'))^2 + \cdots + (d_C(\mathbf{x}) - d_C(\mathbf{x}'))^2}$. The difference from S&O is that the correlation is based on the C different directions of the coordinate system of the D-space. Notice that $g(\cdot)$ in (2.3) is a mixture of squared exponential correlation functions, and for each component we assume a correlation structure in D-space, of dimension C. Like in S&O, a nugget effect can be included in the model, by assuming $\lambda_1 \to \infty$, such that we have $g(\delta) = \alpha_1 \Delta(0) + \sum_{k=2}^{K} \alpha_k \exp\{-\lambda_k \delta^2\}$, where $\Delta(0)$ is 1 when $\delta = 0$, and 0 otherwise.

We now define the prior distribution of the parameters in the correlation function $g(\cdot)$ in equation (2.3). Suppose that conditional on K, α_j and λ_l are independent for every $j \neq l$ and $j, l = 1, \ldots, K$ with prior density given by

$$\pi(\alpha_1, \cdots, \alpha_K, \lambda_2, \cdots, \lambda_K \mid K) \propto \prod_{k=2}^{K} \pi_k(\lambda_k) \text{ with } \sum_{k=2}^{K} \alpha_k = 1 \text{ and } \lambda_2 > \cdots > \lambda_K, \quad (2.4)$$

where $\prod_{k=2}^{K} \pi_k(\lambda_k)$ is the kernel of the prior joint density of $\boldsymbol{\lambda}^T = (\lambda_2, \cdots, \lambda_K)$ and $\pi_k(\lambda_k)$ is the kernel of the log-normal density whose associated normal has mean μ_λ and variance σ^2_λ, $k = 2, \cdots, K$. The α_k's have a uniform prior distribution over the $(K-1)$-simplex.

2.2 The $\mathbf{d}(\cdot)$ process in $\mathbb{R}^C$

In the general case of D being of dimension C, the function that maps the locations from G-space onto D-space is a column vector $\mathbf{d}(\cdot)$ of dimension C. Therefore the prior distribution in S&O has to be adapted to this case, that is

$$\mathbf{d}(\mathbf{x}) \sim GP(\mathbf{m}(\mathbf{x}), \boldsymbol{\sigma}^2_d R_d(\mathbf{x}, \mathbf{x})), \quad (2.5)$$

where, now, $\mathbf{m}(\cdot)$ is the mean vector of dimension C. The covariance structure of $\mathbf{d}(\cdot)$ is such that $\boldsymbol{\sigma}^2_d$ is a matrix $C \times C$ and $R_d(\cdot, \cdot)$ measures the prior correlation among the monitored locations such that $R_d(\mathbf{x}, \mathbf{x}) = 1$. It follows that the matrix of the coordinates of the locations is $C \times n$ with $\mathbf{D} = (\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_n)$, and $\mathbf{d}_i = (d_{1i}, d_{2i}, \cdots, d_{Ci})'$, has a matrix normal distribution given by

$$\mathbf{D} \mid \mathbf{m}, \mathbf{R}_d, \boldsymbol{\sigma}^2_d \sim N_{(C \times n)}(\mathbf{m}, \boldsymbol{\sigma}^2_d, \mathbf{R}_d),$$

where $\mathbf{m}$ is the $C \times n$ matrix representing the mean of the configuration of points in D-space; $\boldsymbol{\sigma}^2_d$ is a $C \times C$ matrix representing one's prior belief about the covariance structure among the axes of the coordinate system in D; $\mathbf{R}_d$ is a correlation matrix, $n \times n$, describing the prior information of the spatial correlation structure amongst the sites in D-space. We next discuss the modeling of $\mathbf{m}(\cdot)$, $\mathbf{R}_d$, and $\boldsymbol{\sigma}^2_d$ in $\mathbb{R}^C$.

Assigning 0 mean to $C-2$ axes Usually, the first two components of the prior mean of the vector $\mathbf{d}(\cdot)$ comprise, respectively, the geographical coordinates of the observed locations in G-space. Now, assume that one has no prior information of which variables could be considered as the $C-2$ remaining axis of the coordinate system in D-space. In this case, a naive solution to the folding problem, would

be to assign a zero mean prior to the $C-2$ remaining components of the vector $\mathbf{d}(\cdot)$. In other words, the prior mean of each observed location would be given by $\mathbf{m}(\mathbf{x}_i) = (x_{1_i}, x_{2_i}, 0, \cdots, 0)'$.

This simple approach can be considered when there is lack of bijectivity in an initial mapping into a space of dimension $C = 2$. If there is no other source of information which could be used, this is the natural way to address this problem. Notice that the inclusion of more axes gives the freedom to adjust better the observed correlations, as a correlation function in $\mathbb{R}^{T-1}$ fits the observed correlations perfectly (Sampson and Guttorp, 1992). The use of the zero mean for the $C-2$ axes is just a simple artifact which gives more flexibility to the model, allowing bigger dimensions in D-space.

Making use of covariates for the $C-2$ axes If it is known that a set of $C-2$ covariates, say $\mathbf{z}(\mathbf{x})$, might influence the correlation structure amongst the sites, they can be taken into account in the model by including them as the coordinates of the function $\mathbf{d}(\cdot)$ which maps the locations from G-space into D-space. In this case we suggest to make $\mathbf{m}(\mathbf{x}) = (x_1, x_2, z_1(\mathbf{x}), \cdots, z_{C-2}(\mathbf{x}))'$.

Defining $\mathbf{R}_d$ in $\mathbb{R}^C$ As in S&O, $\mathbf{R}_d$ is the correlation matrix which gives prior information about the shape of the configuration of points in D-space. And this is measured through the correlation amongst the sites in G-space. The aim is to have a smooth mapping, and this is attained by assigning a squared exponential correlation function for $\mathbf{R}_d(\mathbf{x}, \mathbf{x}^*)$, such that $\mathbf{R}_d(\mathbf{x}, \mathbf{x}^*) = \exp\{-(\mathbf{m}(\mathbf{x}) - \mathbf{m}(\mathbf{x}^*))'\mathbf{B}_d(\mathbf{m}(\mathbf{x}) - \mathbf{m}(\mathbf{x}^*))\}$, where $\mathbf{B}_d$ is a fixed $C \times C$ diagonal matrix with $\mathbf{B}_d = \text{diag}\,(b_d, b_d, b_3, \cdots, b_C)$. Notice that the first two components of $\mathbf{m}(\mathbf{x})$ have the same roughness parameter b_d. As we assume $\mathbf{R}_d$ as a fixed matrix there is no problem in assuming different roughness parameters for the different directions. The $C-2$ diagonal elements of $\mathbf{B}_d$ represent a measure of our belief about the degree of smoothness of $\mathbf{d}(\cdot)$ as a function of each one of the $C-2$ covariates independently (Haylock and O'Hagan, 1996). Notice that in $\mathbf{R}_d(\cdot,\cdot)$ we are taking into account the information of the C directions in the prior structure of the covariance function of $\mathbf{d}(\cdot)$.

If one assumes $\mathbf{m}(\mathbf{x}) = (x_1, x_2, 0, \cdots, 0)'$, as the $C-2$ components have no information in G-space, the computation of $\mathbf{R}_d(\mathbf{x}, \mathbf{x}^*)$ simplifies to $\mathbf{R}_d(\mathbf{x}, \mathbf{x}^*) = \exp\{-b_d\,[(x_1 - x_1^*)^2 + (x_2 - x_2^*)^2]\}$.

Defining $\boldsymbol{\sigma}_d^2$ in $\mathbb{R}^C$ The modeling of the prior covariance matrix of $\mathbf{d}(\cdot)$ also depends on $\boldsymbol{\sigma}_d^2$. As in the case of $C = 2$, described in S&O, $\boldsymbol{\sigma}_d^2$ describes the prior covariance amongst the C axes of the coordinate system in D. The likelihood brings information at most about the eigenvalues of $\boldsymbol{\sigma}_d^2$ (Schmidt and O'Hagan, 2003). It follows then that $\boldsymbol{\sigma}_d^2$ is modelled as a diagonal matrix $C \times C$ such that $\boldsymbol{\sigma}_d^2 = \text{diag}\,(\sigma_{d_{11}}^2, \sigma_{d_{22}}^2, \cdots, \sigma_{d_{CC}}^2)$ and independent inverse gamma prior distributions are assigned to each element of the main diagonal of $\boldsymbol{\sigma}_d^2$, such that $\sigma_{d_{jj}}^2 \sim IG(a_j, b_j)$, for known a_j and b_j, $\forall j = 1, 2, \cdots, C$. Small values of $\sigma_{d_{jj}}^2$, $j = 1, 2, \cdots, C$, imply that less distortion is expected *a priori*. Notice that as $\boldsymbol{\sigma}_d^2$ is a random parameter, the

scaling of the D-space is unknown. Also, the correlation function $g(\cdot)$ has unknown roughness parameters λ_k. Therefore, the scaling of the different axes should not affect the assumption of isotropy of the D-space, as these parameters should take into account the different scales of the axes of the coordinate system. The influence of $\boldsymbol{\sigma}_d^2$ in the prior covariance structure amongst the locations of the sites in D-space is also related to the different values of the roughness parameters in the diagonal of $\mathbf{B}_d$ in $\mathbf{R}_d$. It is advised to use moderate values for the elements of $\mathbf{B}_d$ such that locations which are close together in G-space will tend to be mapped close together in D-space, as their prior correlation might be strong, whereas sites which are far apart in G-space will tend to move more independently from the remaining ones, since their moves in D-space are more influenced by the values assigned to the elements of the main diagonal of $\boldsymbol{\sigma}_d^2$. Each value in the main diagonal of $\boldsymbol{\sigma}_d^2$ gives prior information of how far each coordinate will move in the j^{th} direction. The specification of the prior of $\sigma_{d_{jj}}^2$ is related to the covariate being used in the j^{th} direction and therefore, the scale in which it is being measured.

2.3 Inference procedure for the latent space model

The complete set of parameters is $\boldsymbol{\Theta} = \{(v_1, v_2, \cdots, v_n), (\mathbf{d}_1, \ldots, \mathbf{d}_n), (\alpha_1, \cdots, \alpha_K, \boldsymbol{\lambda}), (\tau^2, \boldsymbol{\sigma}_d^2)\}$. According to Bayes' theorem, the posterior for $\boldsymbol{\Theta}$ is proportional to prior times likelihood, that is

$$\begin{aligned}
\pi(\boldsymbol{\Theta} \mid \mathbf{S}) \propto & \mid \boldsymbol{\Sigma} \mid^{-\frac{T-1}{2}} \exp\left\{-\frac{T}{2}\ \mathrm{tr}\ \mathbf{S}\boldsymbol{\Sigma}^{-1}\right\} \left\{\prod_{i=1}^{n} v_i^{-(f+2)/2} \exp\left(-\frac{(f-2)\tau^2}{2v_i}\right)\right\} \\
& \times \mid \boldsymbol{\sigma}_d^2 \mid^{-n/2} \mid \mathbf{R}_d \mid^{-1} \exp\left\{-\frac{1}{2}\ \mathrm{tr}\ (\mathbf{D}-\mathbf{m})'\boldsymbol{\sigma}_d^{-2}(\mathbf{D}-\mathbf{m})\mathbf{R}_d^{-1}\right\} \qquad (2.6) \\
& \times \prod_{c=1}^{C} (\sigma_{d_{cc}}^2)^{-(\beta_c+2)/2} \exp\left\{-\frac{\alpha_c}{2\sigma_{d_{cc}}^2}\right\} \tau^{\frac{(nf-2)}{2}} \left\{\prod_{k=2}^{K} \frac{1}{\lambda_k} \exp\left\{\frac{-(\log(\lambda_k)-\mu_\lambda)^2}{2\sigma_\lambda^2}\right\}\right\}.
\end{aligned}$$

Analytical summarization of (2.6) seems infeasible and we resort to Markov chain Monte Carlo (MCMC) simulation (see e.g. Gamerman and Lopes (2006)).

The algorithm to obtain samples from the posterior in (2.6) is a hybrid Gibbs sampler. Following the full conditionals of each of the parameters we sample v_i using the adaptive rejection sampling (Gilks and Wild, 1992) as its full conditional is log-concave when expressed in terms of $v_i^{-1/2}$; then we sample elements of $\mathbf{D}$ by Metropolis-Hastings steps. The elements of $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_K)$ are sampled through a Metropolis-Hastings step; as $\sum_{k=1}^{K} \alpha_k = 1$, we make independent proposals for the logit of each α_k and obtain $\alpha_K = 1 - \sum_{k=1}^{K-1} \alpha_k$. We also sample $\lambda_2, \cdots, \lambda_K$ through Metropolis-Hastings steps; τ^2 is sampled from the respective gamma distribution; and the variances of the prior distribution of $\mathbf{D}$, $\sigma_{d_{jj}}^2$, $j = 1, \cdots, C$, are sampled from their respective inverse gamma distributions.

The two most complex steps are to update $\mathbf{D}$ and $b_2, \ldots, b_K$. The full conditional posterior distribution of $\mathbf{d}(\cdot)$ is the combination between the prior multivariate Normal distribution and the likelihood, which is invariant under rotation, location, and scale changes of the configuration of points. The Metropolis-Hastings proposal we

use is based on the principal components of the sample covariance matrix $\mathbf{S}$, and instead of sampling locations we sample directions. If we consider sampling the locations separately, one at a time, then those which are highly correlated tend not to move much. In sampling the directions we overcome this problem because the principal components of $\mathbf{S}$ indicate how the locations are correlated. This ensures that sites which are highly correlated, those with small values of the principal components, tend to move together along the D-space. Sampling from the posterior full conditional of the roughness parameters λ_k in $g(\cdot)$ is also challenging. Assume, for simplicity, that the correlation function has only two components ($K = 2$), one of them a nugget effect. Therefore we have only one roughness parameter, λ_2, in the correlation function. The parameter λ_2 brings information about the size of the configuration. The full conditional distribution of λ_2 is tightly concentrated, allowing little movement in the Markov chain. In order to improve mixing our algorithm is based on moving $\mathbf{D} = (\mathbf{d}_1, \ldots, \mathbf{d}_n)$ and λ_2 together, but making proposals only to λ_2 and obtaining the proposal for $\mathbf{D}$ such that the distances amongst the proposed points in D space keep the current correlations the same. See Schmidt and O'Hagan (2003) for more details.

3 A projection model onto the $\mathbb{R}^2$ manifold

The generalization of the previous section, of making the D-space of dimension $C > 2$, aims at overcoming possible folding when the D-space is assumed to be 2D. These foldings might be a result of local influences in the covariance structure of the process. These local influences might be measured through some covariates, and the previous section provided a way of considering covariates in the mapping function $\mathbf{d}(\cdot)$. However, as the number of monitoring locations increase so does the number of parameters to be estimated. This affects the efficiency of the MCMC algorithm because of the unidentifiability problems described in section 2.3. In this section we discuss a simpler approach of the latent space model by defining the stochastic process in a 2D manifold.

Define the C-space as the G-space coordinates and any covariates e.g. elevation, such that $C \geq 2$. The real world is a 2D manifold (or surface) in C-space because at any point in G-space (which is 2D) all the covariates have specific values. We now define a model for the stochastic process in this C-space by assuming a particular specification for the function $\mathbf{d}(\cdot)$ of the previous section. More specifically, we assume the function $\mathbf{d}(\mathbf{x})$ maps the locations from G onto the larger space C. Cooley et al. (2007) use covariates (but not geographic coordinates) to model extreme precipitation.

Different from the previous section, here we assume $\mathbf{Y}_t \sim N(\boldsymbol{\mu}_t, \boldsymbol{\Sigma})$ and we make inference about $\boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma}$ in a single framework. Assuming we observe $\mathbf{y} = (\mathbf{y}_1, \cdots, \mathbf{y}_T)^T$, the likelihood function is

$$p(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \prod_{t=1}^{T} \mid \boldsymbol{\Sigma} \mid^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y}_t - \boldsymbol{\mu}_t)'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_t - \boldsymbol{\mu}_t)\right\}.$$

Consider $\mathbf{x}$ and $\mathbf{x}^*$ arbitrary locations in G-space, and let $\mathbf{d}_P(\mathbf{x}) = (x_1, x_2, z_1(\mathbf{x}), \cdots, z_{C-2}(\mathbf{x}))'$ map the locations from G into the larger C-space. As outlined in (Paciorek and Schervish, 2006), a covariance structure that generalizes to anisotropic covariance functions and account for directionality is one that uses the Mahalanobis distance between locations. Let

$$Mh(\mathbf{d}_P(\mathbf{x}), \mathbf{d}_P(\mathbf{x}^*)) = \sqrt{(\mathbf{d}_P(\mathbf{x}) - \mathbf{d}_P(\mathbf{x}^*))' \, \mathbf{\Phi}^{-1} \, (\mathbf{d}_P(\mathbf{x}) - \mathbf{d}_P(\mathbf{x}^*))}, \tag{3.1}$$

be the Mahalanobis distance between $\mathbf{d}_P(\mathbf{x})$ and $\mathbf{d}_P(\mathbf{x}^*)$, which is a function of the arbitrary positive definite, $C \times C$, matrix $\mathbf{\Phi}$. A valid covariance function might assume, e.g.

$$\text{Cov } (Y(\mathbf{x}, t), Y(\mathbf{x}^*, t)) = \sigma^2 \, g_P(Mh(\mathbf{d}_P(\mathbf{x}), \mathbf{d}_P(\mathbf{x}^*))), \tag{3.2}$$

where σ^2 represents a common variance across the field. Note that this is a particular case of the general covariance structure defined in equation (2.1), as $v(\mathbf{x}) = \sigma^2$, $\forall \, \mathbf{x}$. And the function $g(\cdot)$ of equation (2.1) is replaced by function $g_P(\cdot)$ which is a function of the Mahalanobis distance $Mh(\cdot;\cdot)$. One suggestion is to make $g_P(-Mh(\mathbf{d}_P(\mathbf{x}), \mathbf{d}_P(\mathbf{x}^*)) = \exp\{-Mh(\mathbf{d}_P(\mathbf{x}), \mathbf{d}_P(\mathbf{x}^*))\}$, which provides a valid covariance structure.

As we have different decay parameters in each of the C directions, the mapping function $\mathbf{d}_p(\cdot)$ allows the geographical locations to shrink/stretch linearly in each dimension but not the completely general deformation that the GP in equation (2.5) allows. We denote the mapping function $\mathbf{d}_P(\cdot)$ as a projection model because the inverse function $\mathbf{d}_P^{-1}(\cdot)$ can be viewed as a projection from C onto G-space. And the covariance function above leads to an anisotropic covariance structure when viewed in G-space.

We now discuss the prior distribution of the parameters in the model. We assume σ^2 follows an inverse gamma prior distribution with parameters a_σ and b_σ, that is, $\sigma^2 \sim IG(a, b)$. For $\mathbf{\Phi}$, one possibility is to assume it as a diagonal matrix, such that the element in the i^{th} diagonal is associated with the decay of the correlation in the i^{th} direction. As we assume the first two components of $\mathbf{d}_P(\mathbf{x})$ as the geographical coordinates, one might assign the same decay parameter for the first two directions, such that $\mathbf{\Phi}^{-1} = diag(1/\phi_1, 1/\phi_1, 1/\phi_3, \cdots, 1/\phi_C)$. We assume independent, inverse gamma prior distributions for $\phi_1, \phi_3, \cdots, \phi_C$ with parameters e_i and h_i. When fixing e_1 and h_1 in the prior distribution of ϕ_1, one suggestion is to assign its prior mean such that the practical range (when the correlation is equal to 0.05) is reached at half of the maximum distance between geographical locations and the variance is fixed at some reasonably large value. For the other decay parameters, $\phi_3, \cdots, \phi_C$, the prior distributions are also chosen following the idea of practical range but considering the scale of each direction separately. A more general possibility is to assume an inverse Wishart prior distribution for $\mathbf{\Phi}$ with a relatively small number of degrees of freedom ν, such that $\nu > C - 1$, and a diagonal scale matrix $\mathbf{V}$. The elements of the main diagonal of $\mathbf{V}$ can be fixed by following the idea of practical range discussed above.

Note that in the case of a diagonal matrix $\mathbf{\Phi}$, if we assume an identity mapping function, such that $\mathbf{d}_P(\mathbf{x}) = \mathbf{x}$, and use the same decay parameter ϕ_1 for both geographical directions, we obtain an isotropic correlation function.

Modelling the mean structure As this model has a significant smaller number of parameters when compared to the general one discussed in section 2, we propose to estimate the mean and the covariance structures of the process in a single framework. In particular, we assume $\boldsymbol{\mu}_t$ is decomposed as the sum of two components, a purely spatial term and a temporal one, such that

$$\boldsymbol{\mu}_t = \mathbf{U}'\boldsymbol{\beta} + \mathbf{F}_t'\boldsymbol{\theta}_t. \tag{3.3}$$

The matrix $\mathbf{U}$, $n \times q$, contains the q covariates that vary only across space and might be a subset of the covariates $\mathbf{Z}(\mathbf{x})$, such that $q \leq C - 2$. The coefficient vector $\boldsymbol{\beta}$, $q \times 1$, measures the effect of each covariate on the mean structure $\boldsymbol{\mu}_t$. One suggestion is to assign a q-variate normal prior distribution to $\boldsymbol{\beta}$, with mean vector $\mathbf{m}_\beta$, and diagonal covariance matrix $\mathbf{C}_\beta$.

The temporal structure is captured through dynamic linear models (West and Harrison, 1997), such that $\mathbf{F}_t$ is a $n \times p$ matrix containing parameters which describe the mean temporal structure of $\mathbf{Y}_t$, e.g. a baseline, seasonal components, time-varying covariates, etc. The p-dimensional coefficient vector $\boldsymbol{\theta}_t$ evolves smoothly with time, such that $\boldsymbol{\theta}_t = \mathbf{G}\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t$ with $\boldsymbol{\omega}_t \sim N(0, \mathbf{W})$, and $\boldsymbol{\theta}_0 \sim N(\mathbf{m}_0, \mathbf{C}_0)$, with $\mathbf{m}_0$ and $\mathbf{C}_0$ known. $\mathbf{G}$ is a $p \times p$ known matrix, and $\mathbf{W}$ is a $p \times p$ matrix, describing the covariance among the elements of $\boldsymbol{\theta}_t$. In particular, we assume $\mathbf{W}$ as a diagonal matrix with elements $\mathbf{W} = diag(W_1, \cdots, W_p)$, and we assign independent inverse gamma prior distributions to each W_i, such that $W_i \sim IG(a_W, b_W)$, $i = 1, \cdots, p$.

A nugget effect ν^2 can be accommodated in this model by making

$$\mathbf{\Sigma}_{ij} = \sigma^2 \exp\{-Mh(\mathbf{d}_P(\mathbf{x}_i), \mathbf{d}_P(\mathbf{x}_j))\} + \nu^2 \, \Delta(\mathbf{d}_{P_i}, \mathbf{d}_{P_j}), \tag{3.4}$$

where $\Delta(\mathbf{d}_{P_i}, \mathbf{d}_{P_j}) = 1$ if $\mathbf{x}_i = \mathbf{x}_j$, and 0 otherwise. If a nugget effect is included in the model, we assume $\nu^2 \sim IG(c, d)$, *a priori*, for known c and d.

3.1 Inference procedure for the projection model

The parameter vector to be estimated in the projection model is $\boldsymbol{\vartheta} = (\boldsymbol{\beta}, \sigma^2, \mathbf{\Phi}, \nu^2, \mathbf{W}, \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_T)$. Following the prior specification discussed above, and the Bayes' theorem, the posterior distribution of $\boldsymbol{\vartheta}$ is proportional to

$$\begin{aligned}
\pi(\boldsymbol{\vartheta} \mid \mathbf{y}) \propto & \prod_{t=1}^{T} \mid \mathbf{\Sigma} \mid^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y}_t - \mathbf{U}'\boldsymbol{\beta} - \mathbf{F}_t'\boldsymbol{\theta}_t)'\mathbf{\Sigma}^{-1}(\mathbf{y}_t - \mathbf{U}'\boldsymbol{\beta} - \mathbf{F}_t'\boldsymbol{\theta}_t)\right\} \\
& \times \left[\prod_{t=1}^{T} \mid \mathbf{W} \mid^{-1/2} \exp\left\{-0.5\,(\boldsymbol{\theta}_t - \mathbf{G}\boldsymbol{\theta}_{t-1})'\,\mathbf{W}^{-1}\,(\boldsymbol{\theta}_t - \mathbf{G}\boldsymbol{\theta}_{t-1})\right\}\right] \\
& \times \exp\left\{-0.5(\boldsymbol{\theta}_0 - \mathbf{m}_0)'\mathbf{C}_0^{-1}(\boldsymbol{\theta}_0 - \mathbf{m}_0)\right\} \exp\left\{-0.5(\boldsymbol{\beta} - \mathbf{m}_\beta)'\mathbf{C}_\beta^{-1}(\boldsymbol{\beta} - \mathbf{m}_\beta)\right\} \\
& \times \prod_{i=1}^{p}\left[W_i^{-a_w+1} \exp\left\{-\frac{b_w}{W_i}\right\}\right] (\sigma^2)^{-a_\sigma-1} \exp\left\{-\frac{b_\sigma}{\sigma^2}\right\} (\nu^2)^{-c-1} \exp\left\{-\frac{d}{\nu^2}\right\} \prod_{\substack{i=1 \\ i\neq 2}}^{C} \phi_i^{-(e_i-1)} \exp(-h_i/\phi_i).
\end{aligned}$$

Analogously to the previous model, we resort to MCMC to obtain samples from the posterior distribution of $\boldsymbol{\vartheta}$. In particular, we use the Gibbs sampling with some steps of the Metropolis-Hastings algorithm. The coefficient vector $\boldsymbol{\beta}$ has a multivariate normal posterior full conditional distribution. The elements of $\boldsymbol{\theta}$ are sampled through the forward filtering backward (FFBS) sampling algorithm, introduced by Frühwirth-Schnater (1994). The posterior full conditional distribution of each W_i follows an inverse gamma distribution. The elements of $\boldsymbol{\Phi}$, σ^2, and ν^2 are sampled through Metropolis-Hastings steps.

3.2 Spatial interpolation

From a Bayesian point of view, spatial interpolation is based on the posterior predictive distribution, which we now obtain for the projection model.

We aim at predicting the process at a set of locations which have not been observed, say $\mathbf{Y}_t^{out} = (Y(\mathbf{x}_{u1}, t), \cdots, Y(\mathbf{x}_{uL}, t))'$ at unobserved locations $\mathbf{x}_{u1}, \cdots, \mathbf{x}_{uL}$. For each time t, samples $Y_t(\mathbf{x})$ are being generated from the multivariate normal distribution, $N_n(\boldsymbol{\mu}_t, \boldsymbol{\Sigma})$, the posterior predictive distribution, $p(\mathbf{y}_t^{out}|\mathbf{y})$, is given by

$$p(\mathbf{y}_t^{out}|\mathbf{y}) = \int_{\boldsymbol{\vartheta}} p(\mathbf{y}_t^{out}|\mathbf{y}, \boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta}|\mathbf{y})d\boldsymbol{\vartheta}. \tag{3.5}$$

From the theory on the multivariate normal distribution (Anderson, 1984), it follows that the joint distribution of $\mathbf{Y}_t$ and $\mathbf{Y}_{t_u}$, conditioned on $\boldsymbol{\vartheta}$, is given by

$$\begin{pmatrix} \mathbf{Y}_t^{out} \\ \mathbf{Y}_t \end{pmatrix} \mid \boldsymbol{\vartheta} \sim N_{n+L}\left(\begin{pmatrix} \boldsymbol{\mu}_t^{out} \\ \boldsymbol{\mu}_t \end{pmatrix}; \begin{pmatrix} \boldsymbol{\Sigma}^{out} & \boldsymbol{\Psi}' \\ \boldsymbol{\Psi} & \boldsymbol{\Sigma} \end{pmatrix} \right), \tag{3.6}$$

where $\boldsymbol{\mu}_t^{out}$ is a L-dimensional vector representing the mean of the unobserved locations at time t; $\boldsymbol{\mu}_t$ is a vector comprising the mean of the observed sites; $\boldsymbol{\Sigma}^{out}$ is a covariance matrix of dimension L and each of its element is the covariance of the process between unobserved locations. Each line of the matrix $\boldsymbol{\Psi}$, $n \times L$, represents the covariance between the i^{th} monitored location and the j^{th} unobserved one, $i = 1, \cdots, n$ and $j = 1, \cdots, L$. From the theory of the multivariate normal distribution we have that

$$\mathbf{Y}_t^{out}|\mathbf{y}_t, \boldsymbol{\vartheta} \sim N_L\left(\boldsymbol{\mu}_t^{out} + \boldsymbol{\Psi}^T\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}_t - \boldsymbol{\mu}_t\right); \boldsymbol{\Sigma}^{out} - \boldsymbol{\Psi}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Psi}\right). \tag{3.7}$$

The integration in (3.5) does not have an analytical solution, however approximations can be easily obtained through Monte Carlo methods (Gamerman and Lopes, 2006). For each sample q, $q = 1, \cdots, Q$, obtained from the MCMC algorithm, we can obtain an approximation for (3.5), by sampling from the distribution in (3.7) and computing

$$p(\mathbf{y}_t^{out}|\mathbf{y}_t) \approx \frac{1}{Q}\sum_{q=1}^{Q} p(\mathbf{y}_t^{out}|\boldsymbol{\vartheta}^q). \tag{3.8}$$

The approximation above is also suitable to compare models. Usually, one holds a set of observations out from the inference procedure and uses the posterior predictive distribution to compare all fitted models in terms of their prediction abilities. Greater values of $p(\mathbf{y}_t^{out}|\mathbf{y}_t)$ in (3.8) point to the best model.

4 Applications

We fit the proposed models to two datasets, solar radiation in British Columbia and mean temperature in Colorado. In particular, as the number of locations for the temperature data is relatively large, $n = 151$, we only fit the projection model to this data, as the MCMC algorithm for the general latent space model tends to take extremely long to converge when n is too big.

4.1 Solar radiation in British Columbia

Here we revisit the solar radiation data set collected in southwestern British Columbia, Canada, which was analysed by S&G and S&O. Accurate mesoscale prediction of solar radiation is important for the development of solar energy systems. The network consists of $n = 12$ monitoring stations and we have available a sample covariance matrix obtained after removing the temporal effect from the original data, as analysed by S&G. The sample covariance is based on observations at $n = 12$ locations, from March 22nd to September 20th, for 4 years, from 1980 until 1983, so that $T = 732$.

Station 1 lies at a very different elevation compared to the remaining ones. The analysis performed by S&O led to a lack of bijectivity of the mapping. Hay (1984) gives a description of the location of the monitored sites. He goes on to say that the network includes a transect along the major orographically induced climatic gradient in Vancouver (from south to north), a transect from the coast to a distant but reasonably accessible inland location and a number of stations located in the central urban area and sparsely populated rural areas. We would expect that the correlation among the locations in the North-South direction would change faster than in the East-West direction. Because of these geographical characteristics of the network, we make use of the coordinate system in G-space standardized in all directions. The coordinates of the monitoring stations are shown in Table 1.

We fit three different models to this dataset. The first is the same fitted by S&O which assumes the D-space to be 2D. More specifically, the $\mathbf{d}(\cdot)$ function maps locations from $\mathbb{R}^2$ onto $\mathbb{R}^2$ and the prior mean of $\mathbf{d}(\cdot)$ assumes $\mathbf{m}(\mathbf{x}_i)$ to be the identity function, that is $\mathbf{m}(\mathbf{x}_i) = \mathbf{x}_i'$. This model is denoted by $\mathbb{R}^2 \to \mathbb{R}^2$. The second model assumes the D-space to be 3D, and uses elevation in the prior mean of $\mathbf{d}(\cdot)$ such that $\mathbf{m}(\mathbf{x}_i) = (\mathbf{x}_i, elevation_i)'$. This model is denoted by $\mathbb{R}^2 \to \mathbb{R}^3$. As models $\mathbb{R}^2 \to \mathbb{R}^2$ and $\mathbb{R}^2 \to \mathbb{R}^3$ differ only in the specification of the mapping function $\mathbf{d}(\cdot)$ we assume the same prior distribution for all parameters which are common in both models. More specifically, we assume the correlation function $g(\cdot)$ in equation (2.3) to have $K = 3$ components, a nugget plus two components of squared exponential correlation functions. For λ_k, the decay parameters in the correlation function $g(\cdot)$, we assume a log-normal prior distribution with mean 0.1 and variance such that the probability of λ_k being greater than 2 is 0.01. The prior degrees of freedom of the variances were set equal to 10, because we do not have strong prior belief about the variances of the process at different locations being similar. For this example we set b_d, the roughness parameter of the prior correlation

Table 1: Latitude, longitude and elevation of the 12 monitoring sites of the solar radiation data set.

Site	Latitude ($^{o'}$)	Longitude ($^{o'}$)	Elevation (m)
1	49.23	123.05	1128
2	49.19	123.04	114
3	49.16	123.07	122
4	49.11	123.10	5
5	49.00	123.08	3
6	49.13	122.42	5
7	49.08	122.17	125
8	49.02	122.17	60
9	49.01	122.22	61
10	49.06	122.38	11
11	49.13	123.06	63
12	49.16	123.15	93

function of the locations in D space, to 1.7. As previously discussed, $\sigma^2_{d_{jj}}$ ($j = 1, 2$ in the $\mathbb{R}^2 \to \mathbb{R}^2$ model, and $j = 1, 2, 3$ in the $\mathbb{R}^2 \to \mathbb{R}^3$ model) controls how far the sites are expected to move *a priori*. We assign inverse gamma priors to $\sigma^2_{d_{jj}}$, where due to the coordinate system $E(\sigma^2_{d_{jj}}) = 0.5$ with 30 degrees of freedom, *a priori*.

A projection model (PM) We also fit a model with covariance structure based on the projection proposal described in section 3. As for this data we only have the sample covariance matrix available, based on equation (2.1), we propose to model the covariance as

$$\Sigma^{PM}_{ij} = \sqrt{v(\mathbf{x}_i)\, v(\mathbf{x}_j)} \left[a_1 \Delta(\mathbf{d}_{P_i}, \mathbf{d}_{P_j}) + a_2 \ \exp\left\{ - \sqrt{\left(\mathbf{d}_{P_i} - \mathbf{d}_{P_j}\right)' \mathbf{\Phi}^{-1} \left(\mathbf{d}_{P_i} - \mathbf{d}_{P_j}\right)} \right\} \right],$$

where $\mathbf{d}_{P_i} = (\mathbf{x}_i, elevation_i)'$, and $\mathbf{\Phi} = diag(\phi_1, \phi_1, \phi_2)$. For ϕ_1 and ϕ_2 we assign inverse gamma prior distributions, with parameters based on the idea of practical range. Notice that this is a much simpler version of the latent space model as we do not need to estimate the locations of the sites in D-space. The parameters to be estimated are $\mathbf{\Theta}^{PM} = ((v_1, \cdots, v_n), a_1, a_2, \phi_1, \phi_2)$. The MCMC algorithm is similar to the one described in section 2.3, with the advantage that here, at each iteration of the MCMC, we skip the steps of sampling $\mathbf{d}$ and $\boldsymbol{\sigma}^2_{d_{ii}}$, $i = 1, 2, 3$.

For model $\mathbb{R}^2 \to \mathbb{R}^2$ we have run the chain for $90,000$ iterations and, to avoid autocorrelation within the chains, we stored every 35^{th} element after considering a burn in of $20,000$ iterations. For model $\mathbb{R}^2 \to \mathbb{R}^3$, the MCMC was run for $200,000$ iterations, with a burn in of $20,000$ and after the burn in, we kept every 90^{th} iteration. For the projection model (PM), we let the MCMC run for $50,000$ iterations,

considered a burn in of $10,000$ iterations, and stored every 40^{th} iteration. For all models, convergence was checked using the tools available in the software CODA (Plummer et al., 2006). Model $\mathbb{R}^2 \to \mathbb{R}^3$ required a longer chain to reach convergence when compared to models $\mathbb{R}^2 \to \mathbb{R}^2$ and PM. Recall that PM does not have unknown locations in D-space, therefore does not suffer from the unidentifiability problems discussed in section 2.3.

For models $\mathbb{R}^2 \to \mathbb{R}^2$ and $\mathbb{R}^2 \to \mathbb{R}^3$ we obtain the estimated correlations versus the Euclidean distance among locations in D-space, and compare these estimates with the observed correlations versus the distances in G-space to check if the models are correcting the anisotropy present in the data. Model $\mathbb{R}^2 \to \mathbb{R}^3$ tends to provide estimated correlations closer to the theoretical correlation induced by the resepctive model, than model $\mathbb{R}^2 \to \mathbb{R}^2$ (Figure 1).

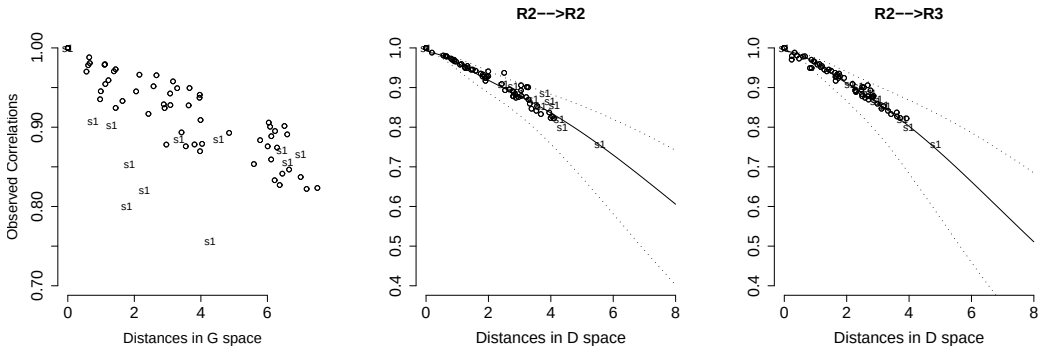

Figure 1: Observed correlation versus distance in G-space (1st column), posterior mean (solid line) and associated 95% posterior credible intervals (dashed lines) of the correlation function for the solar radiation data set under models $\mathbb{R}^2 \to \mathbb{R}^2$, and $\mathbb{R}^2 \to \mathbb{R}^3$ (2nd and 3rd columns) . The abscissae of the plots in the 2nd and 3rd columns represents the average distance obtained in D-space.

For the latent space idea it is interesting to investigate how the locations moved in D-space in order to provide an isotropic covariance structure. Following Schmidt and O'Hagan (2003) we used the Procrustes superimposition to compare the original locations with their mapping onto D-space for model $\mathbb{R}^2 \to \mathbb{R}^3$ (Figure 2). One of the difficulties in considering a D-space of dimension bigger than 2 is how to

present the shape of the configuration in the latent space. Here we use a $2D$ plot in which the third coordinate is represented by the radius of circles centered at the respective sites in D-space. Those sites which are labelled by $-s_i$ mean that they have a negative third coordinate. Sites s_4, s_5, s_6, s_{11} and s_{12} have a negative value of the third coordinate. Site 1 is quite far from all the others, and site 2 is the closest one to it. Sites 4 and 12 are quite close in D-space and both are closer to site 5. The latter being the least correlated with site 1 as it moves further down in the direction of the third coordinate (Figure 2).

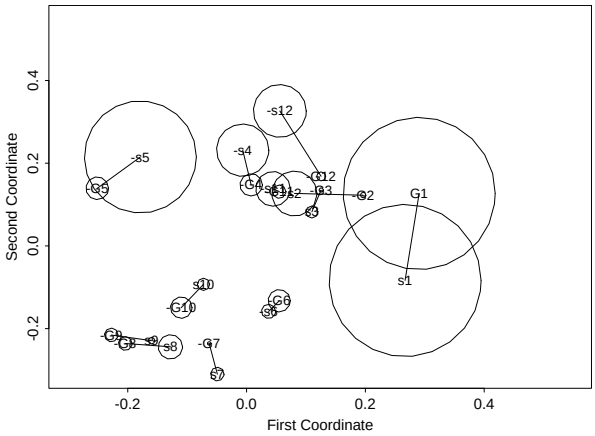

Figure 2: Procrustes superimposition of the posterior mean of the locations in D-space onto G-space under model $\mathbb{R}^2 \rightarrow \mathbb{R}^3$. The original coordinates are labelled as G_i, whereas the coordinates in D-space are labelled s_i. Sites labelled with $-s_i$ represent a negative posterior mean of the third coordinate in D-space.

4.1.1 Prediction of the augmented covariance matrix

To compare the different models we held out site 6 from the inference procedure, fitted all three models and predicted the augmented covariance under each of the fitted models. The algorithm to obtain the predicted augmented covariance matrix follows Schmidt and O'Hagan (2003). Once we have samples from the posterior distribution of $\boldsymbol{\theta}$, we can estimate the covariance between a unmonitored location and the monitored ones following the model specification in equation (2.1). Note that the posterior distribution of the variance at a unmonitored location is equal to its prior distribution. Also, the posterior distribution of $\mathbf{d}(\cdot)$ at a unmonitored location is obtained through the properties of the partition of the multivariate normal distribution. See Schmidt and O'Hagan (2003) for details.

We obtain quite similar interquartile ranges of the posterior predictive distribution for the covariance between site 6 and all the others. The most complex model, $\mathbb{R}^2 \rightarrow \mathbb{R}^3$, tends to provide smaller ranges of the interquartile ranges. Except for

sites 7, 8 and 9 the project model tends to provide very similar results to the most complex fitted models (Figure 3).

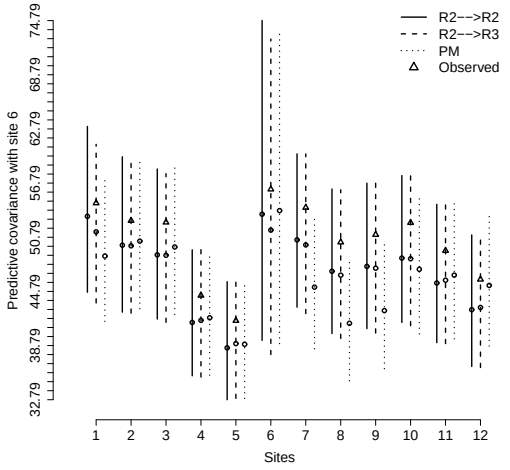

Figure 3: Interquartile ranges of the predictive augmented covariance matrix between site 6 and all the others, when site 6 is held out from the inference procedure. The triangle represents the observed covariance and the open circles represent the median of the posterior predictive distribution under the respective model. The solid line represents the ranges under model $\mathbb{R}^2 \to \mathbb{R}^2$, the dashed line under model $\mathbb{R}^2 \to \mathbb{R}^3$, and the dotted line under model PM.

4.2 Mean temperature in Colorado

As pointed out by Paciorek and Schervish (2006) the Geophysical Statistics Project at the National Centre for Atmospheric Research has posted a useful subset of the United States climate record over the past century from a large network of weather stations. In particular the state of Colorado presents interesting variations in terms of topography. There are in Colorado 367 locations which have colocated information about temperature, precipitation, latitude, longitude and elevation (`http://www.image.ucar.edu/Data/US.monthly.met/CO.shtml`). We concentrated on a subset of locations after removing from the data locations which had more than 3 missing values during January 1991 and December 1997. We model the monthly mean temperature, obtained as the average between the monthly maximum and minimum temperatures, observed at $n = 151$ locations, between 1991 and 1997. The final sample has less than 0.8% of missing observations. From a Bayesian point of view this is not a problem as these missing observations are viewed as parameters and are estimated together with the remaining parameters of the model. For spatial interpolation purposes, we held the time series of $L = 20$ locations out from the sample to compare the fitted models.

The elevations present in the sample vary from 811m up to 3537m. Apparently, many locations are close together when mapped in the geographical space, but once elevation is taken into account they seem to be more distant from each other (1st panel of figure 4). Exploratory data analysis suggests that elevation is negatively related with mean temperature. The time series in figure 4 show a clear seasonal pattern.

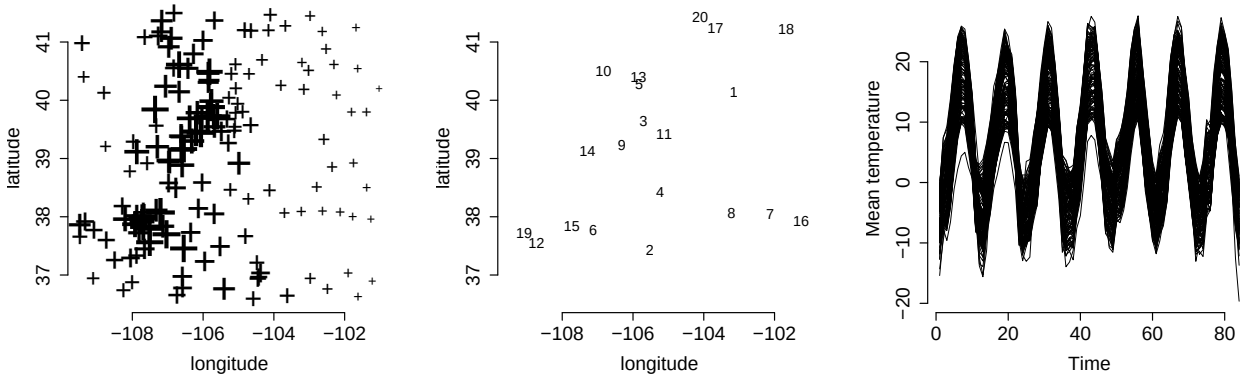

Figure 4: Geographical locations and observed time series of the mean temperature data in Colorado. In the first panel the size of the crosses are proportional to observed elevation, the middle panel shows the location of the 20 stations held out for spatial interpolation, and the third panel shows all 151 observed time series.

Following the model proposed in section 3, the mean vector $\boldsymbol{\mu}_t$ in equation (3.3) is assumed to have components

$$\mu_t(\mathbf{x}_i) = U(\mathbf{x}_i)\beta + \mathbf{F}(\mathbf{x}_i)\boldsymbol{\theta}_t \quad \text{with} \quad \boldsymbol{\theta}_t = \mathbf{G}\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t,$$

where $U(\mathbf{x}_i)$ is standardized elevation at location $\mathbf{x}_i$, $\boldsymbol{\theta}_t = (\theta_{t1}, \theta_{t2}, \theta_{t3})'$, $\mathbf{F}(\mathbf{x}_i) = \begin{pmatrix} 1 & 1 & 0 \end{pmatrix}, i = 1, \cdots, n, W = diag(W_1, W_2, W_2)$, and $\mathbf{G} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos 2\pi/12 & \sin 2\pi/12 \\ 0 & -\sin 2\pi/12 & \cos 2\pi/12 \end{pmatrix}$,

to account for a baseline and a seasonal pattern in each time series. We fitted two models with exact the same mean structure but different covariance structures:

- Isotropic model (IM): for this model we assume $\mathbf{d}_P(\mathbf{x}_i) = \mathbf{x}_i$ in equation (3.4) and $\mathbf{\Phi} = diag(\phi_1, \phi_1)$;

- Projection model (PM): here we assume $\mathbf{d}_P(\mathbf{x}_i) = (\mathbf{x}_i, U(\mathbf{x}_i))'$ in equation (3.4) and $\mathbf{\Phi} = diag(\phi_1, \phi_1, \phi_2)$.

We assumed the following prior specification for both models. For β we assigned a zero mean normal distribution with variance equal to 100. For $\boldsymbol{\theta}_0$ we assumed a zero mean multivariate normal distribution, with a diagonal covariance matrix with elements fixed at 100. For σ^2 and ν^2 we assigned inverse gamma prior distributions with infinite variance and mean based on the mean of ordinary least square fits to each of the observed time series. For the decay parameters in the correlation function we assigned inverse gamma prior distributions with infinite variance and mean based on the idea of practical range, described in section 3. For each model we let the chains run for 40,000 iterations, considered 5,000 as burn in, and stored every 30^{th} iteration to avoid autocorrelation within each chain. Convergence for each model was checked by running two chains starting from very different values.

The projection model provides an estimate of the nugget effect which is considerably smaller than that obtained under the isotropic model (3rd column of Figure 5). There are many sites which are very close in G-space but they become further apart when elevation is taken into account (first panel of Figure 4). The projection model seems to capture this information, as the differences in the observed measurements for sites close together in G-space might not be due to measurement error. This suggests that the presence of elevation in the covariance structure of model PM is capturing some structure left in the residual of the isotropic model, even after accounting for its influence in the mean of the process. The posterior distribution of the coefficient of elevation, β, seems to be affected by the presence of elevation in the covariance structure of the underlying spatial process (1st column of Figure 5).

Although not shown here, the parameters in the temporal structure of the model ($\boldsymbol{\theta}$) seem not to be affected by the different assumptions about the covariance structure $\mathbf{\Sigma}$.

The decay parameter related to the geographical space seems to be smaller under PM than under IM. This seems reasonable as in the PM the observed correlation is being modelled as a function of geographical and elevation distances (Figure 6).

The effect of the different estimates of the decay parameters in the respective correlation functions can be further noticed when we look at the estimate of the correlation between a particular location and all the others in the sample (Figure 7). We present the posterior mean of the correlation between the highest, as well as the lowest, points in the sample and all the others, under models IM and PM. Different from IM, PM provides estimated correlations which change with direction when viewed in G-space (Figure 7).

Regarding the $L = 20$ locations held out from the inference procedure, apparently model PM tends to provide ranges of the 95% posterior predictive distribution

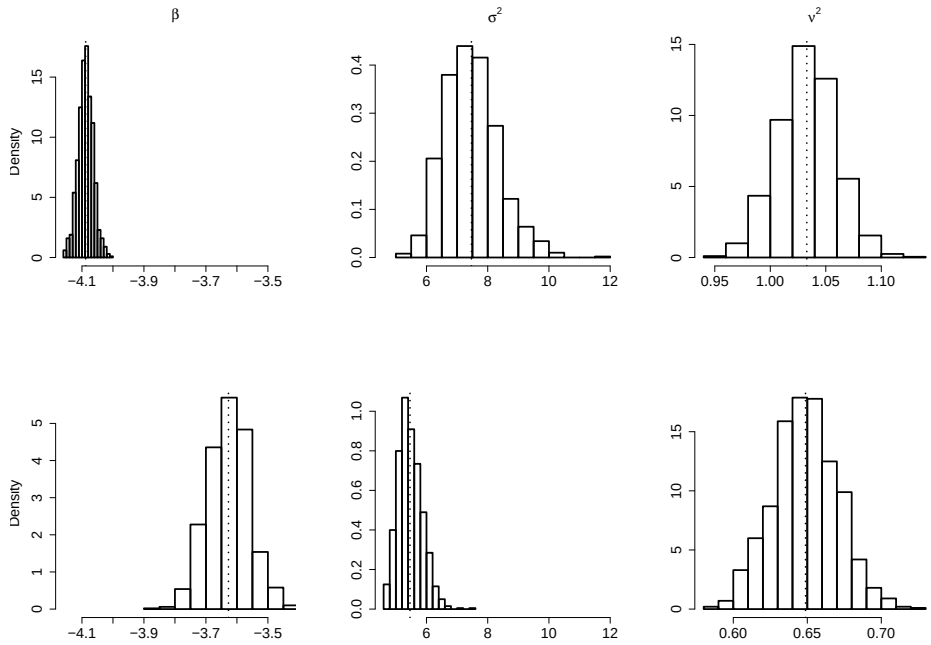

Figure 5: Posterior distribution of the coefficient of elevation, β, the variance of the spatial process, σ^2, and the nugget effect ν^2, under the isotropic and projection models (rows) for the mean temperature data. In each panel, the vertical dotted line is the posterior mean of the respective parameter.

slightly smaller than those obtained under model IM (Figure 8). The predictive ability of each model can be compared using different measurements. The mean square error (MSE) is based on the average square difference between the posterior mean of the predictive distributions and the observed values. Based on all estimates, 20 locations and 84 instants in time for each location, the MSE under PM is equal to 1.876, whereas under IM it is equal to 1.999. A similar result is obtained if we compare the average range of the 95% posterior predictive interval obtained under each model. For IM, the average range is equal to 5.043, whereas for PM it is equal to 4.396. These measurements suggest that model PM performs better than IM in terms of spatial interpolation.

To compare observed and fitted values, Bastos and O'Hagan (2009) suggest a generalization of a chi-square test for correlated observations. For each time t we have the predicted values at the 20 locations held out from the inference procedure. Following Bastos and O'Hagan (2009) we compute the Mahalanobis distance between the predicted and observed values. More specifically, for each time t we compute,

$$D_t(\mathbf{y}_t^{out}) = (\mathbf{y}_t^{out} - \boldsymbol{\mu}_t^*)^T {\boldsymbol{\Sigma}^*}^{-1} (\mathbf{y}_t^{out} - \boldsymbol{\mu}_t^*), \tag{4.1}$$

where $\boldsymbol{\mu}_t^*$ and $\boldsymbol{\Sigma}^*$ are the moments of the normal distribution defined in equation (3.7). As for each time $t = 1, \cdots, 84$ we are estimating the process at 20 locations,

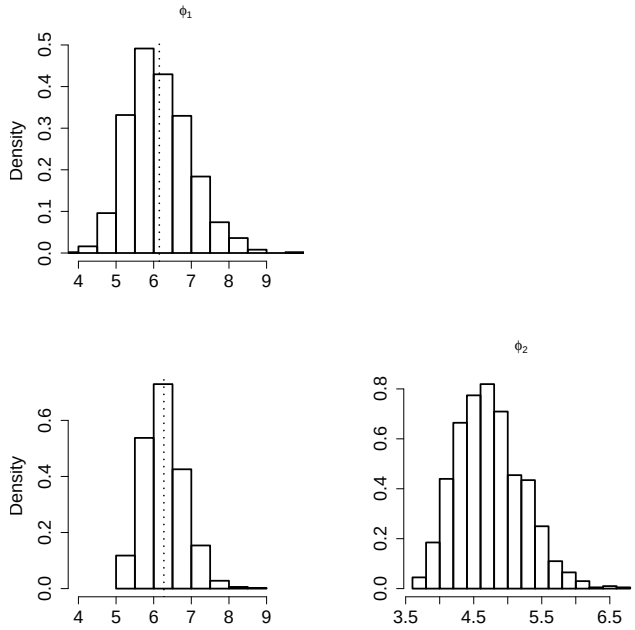

Figure 6: Posterior distribution of ϕ_1 and ϕ_2 for the isotropic and projection models (rows). In each panel, the vertical dotted line is the posterior mean of the respective parameter.

we expect $D_t(\cdot)$ to be equal to 20. Bastos and O'Hagan (2009) mention that extreme values (large or small) of $D_t(\cdot)$ indicate a conflict between the observed and predicted values. For this example, for each time t there seems to be a different model closer to 20 (left panel of Figure 9). However, if we compute $\frac{1}{84}\sum_{t=1}^{84} \mid D_t(\mathbf{y}_t^{out}) - 20 \mid$, the average of the absolute differences between the observed distance and 20, PM provides a value equal to 15.78, whereas IM results on 16.95. The predictive likelihood also provides evidence that PM performs better in terms of spatial interpolation as the average of the predictive likelihood under IM is -33.044, and equal to -30.275 under PM (right panel of Figure 9).

5 Discussion

We proposed models which make use of covariate information in the covariance structure of spatial processes. This is useful when modelling nonstationary spatial processes, as when predicting the process at unobserved locations of interest, the predictions are heavily dependent on the specification of the underlying spatial covariance structure. Following the model proposed by S&O we assume the latent space to be of dimension $C > 2$. This generalization seems appealing, but some points are naturally raised for discussion. Increasing the dimension of the latent space results in the increase of the parameters to be estimated. This affects directly the efficiency of the MCMC, making the convergence of the chains more difficult.

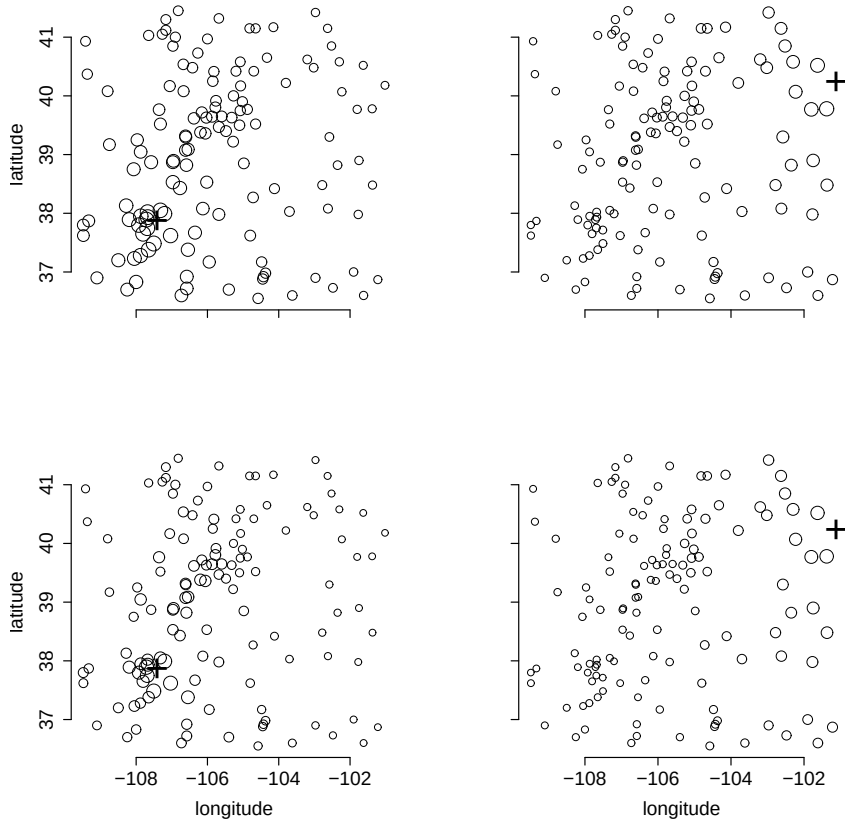

Figure 7: Posterior mean of the correlation of the highest and lowest points, marked by + (coordinates $(-107.512, 37.714, 3537)$, $(-101.02, 40.18, 811)$) (columns) with all the others in the sample for the isotropic and projection models (rows) .

The latent model has a complex hierarchical structure. It is important to recognize that without constraints there is an inherent unidentifiability due to the way $\mathbf{d}(\cdot)$ appears in the $g(\cdot)$ function in equation (2.1). First, any transformation of coordinates $\mathbf{d}(\cdot)$ in D which leaves distances unchanged (translation and/or rotation) is observationally equivalent. Second, because $g(\cdot)$ has unspecified roughness parameters λ_k, any transformation which multiplies all distances in D space by a constant is also unidentifiable. Although our MCMC algorithm is built to recognize these unidentifiability problems, as the number of sites increases this may cause the MCMC to take very long to converge.

In the solar radiation example, there were no apparent problems of convergence because of the low number of gauged sites and the great number of replicates in time. In the general case of making the D-space C dimensional, there are $n + Cn + 2K + C + 1$ parameters to be estimated. Therefore, there is a problem of parsimony related to the choice of the dimension of D-space. The bigger the dimension of

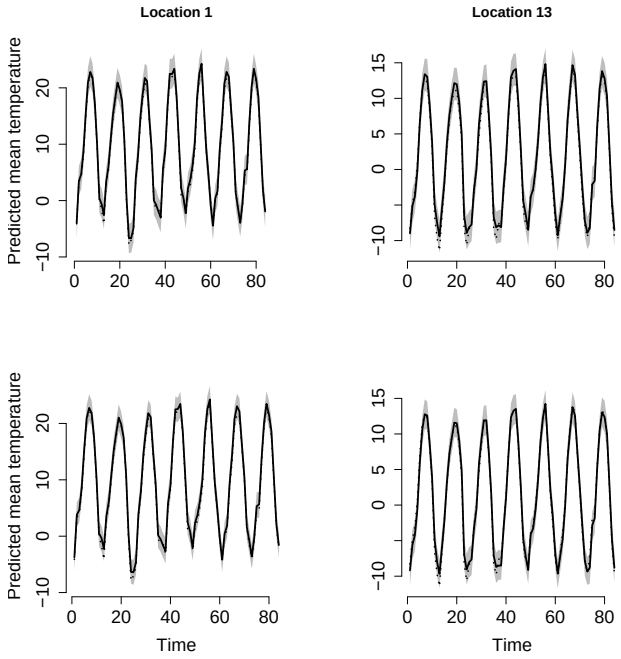

Figure 8: Posterior summary of the predictive distribution for locations 1 and 13 (see middle panel of Figure 4) held out from the inference procedure under models IM and PM (rows). In each panel, the solid line is the mean of the posterior predictive distribution, the dashed line is the actual observed value, and the shaded areas represent the 95% posterior predictive credible intervals.

the D-space the better is the fit of the observed correlations, but the bigger is the number of the parameters in the model. Also, making the D-space of dimension greater than 2 makes it more difficult to visualize the shape of the locations in this space. Here we managed to show the Procrustes superimposition of the locations in D onto the original configuration in G-space, by using $2D$ plots with circles around the gauged sites. The radius of these circles were proportional to the absolute value of the component which represents the third coordinate of the sites.

As an alternative to the general model, we introduced in section 3 the projection model which has a significant smaller number of parameters but is still able to capture interesting correlation structures present in the data. As the covariates are a function of the geographical locations, the spatial process can be viewed as defined in a 2D manifold, and the covariance structure proposed in equation (3.2) provides an anisotropic covariance function in $\mathbb{R}^2$. Because of its considerably smaller number of parameters, the MCMC did not take long to converge, even for the mean temperature data with $n = 131$ locations. In this example, the spatial interpolation to unmonitored locations under the projection model performed better than the usual, simpler, isotropic model. Although both models include elevation in their respective mean structure, the projection model still performs better because it is able to reflect that not only do temperature tend to be lower in the mountains but

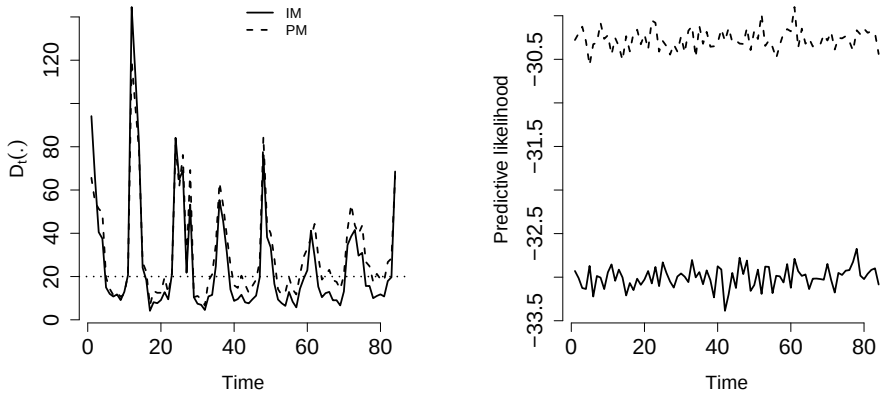

Figure 9: Values of $D_t(\cdot)$ (see equation (4.1)) and posterior predictive likelihood (see equation (3.8)) for each time $t = 1, 2, \cdots, 84$ based on the 20 locations held out from the inference procedure (see Figure 4). In both panels, the solid line corresponds to IM and the dashed line to PM.

they are less highly correlated for given geographical distance.

This projection model can be used as an investigation tool to understand better the covariance structure of the spatial process of interest, and see which covariates might provide better estimates for unmonitored locations of interest. Here we explored an exponential correlation function in the projection model using the Mahalanobis distance between $\mathbf{d}_P(\mathbf{x})$ and $\mathbf{d}_P(\mathbf{x}^*)$, for $\mathbf{x}$ and $\mathbf{x}^* \in G$-space. But different correlation functions might be used in equation (3.2), for example, one possibility is to assume $g_P(\cdot,\cdot)$ as the product of valid correlation functions in each of the C directions, e.g. $g_P(\mathbf{d}_P(\mathbf{x}), \mathbf{d}_P(\mathbf{x}^*), \mathbf{\Phi}) = \prod_{i=1}^{C} g_{P_i}(|\mathbf{d}_P(\mathbf{x})_i - \mathbf{d}_P(\mathbf{x}^*)_i|, \phi_i)$, where each $g_{P_i}(\cdot, \phi_i)$ is a valid correlation function with parameter vector ϕ_i, and $\mathbf{d}_P(\mathbf{x})_i$ is the i^{th} component of the vector $\mathbf{d}_P(\mathbf{x})$. Stein (1999)[p. 54] mentions that the behavior of such covariance functions depend heavily on the choice of the axes. Therefore, it is crucial to understand the process under study in order to use suitable covariates in the covariance structure. Because of its relative simplicity, the projection model allows that the mean and covariance structures of the underlying Gaussian process are estimated in a single framework.

We believe the use of covariates might also be explored in other approaches that handle nonstationary spatial processes. Calder (2008) makes use of wind measurements in the kernel convolution approach of Higdon (1998). However, the information about the wind field is taken at a single location, and this is used to estimate

directional variograms which are used to build the covariance matrix of the Gaussian kernels of the convolution. We are currently investigating ways of including the wind field (considering measurements over a grid of locations) on the covariance structure of environmental spatial processes.

Acknowledgements

The authors thank Paul Sampson and Claudia Tebaldi for providing the solar radiation and the temperature data, respectively. They also thank an associate editor and two anonymous reviewers whose suggestions greatly improved the presentation of the paper. A. M. Schmidt was partially supported by FAPERJ and CNPq. Schmidt and Guttorp are grateful to *Núcleo de Apoio à Pesquisa em Modelagem Estocástica e Complexidade* (NUMEC)), USP, Brazil, for giving the opportunity to discuss initial ideas on this project during the *Workshop on Stochastic Processes Applied to Spatial Statistics: Multi-scenario analysis and stochasticity in environmental prediction*, in December 2007.

References

Anderson, T. (1984) *An Introduction to Multivariate Statistical Analysis.* John Wiley & Sons, Inc.

Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2004) *Hierarchical Modeling and Analysis of Spatial Data.* New York: Chapman and Hall.

Bastos, L. and O'Hagan, A. (2009) Diagnostics for Gaussian process emulators. *Technometrics*, **19**, 39–48.

Calder, C. A. (2008) A dynamic process convolution approach to modeling ambient particulate matter concentrations. *Environmetrics*, **19**, 39–48.

Cooley, D., Nychka, D. and Naveau, P. (2007) Bayesian spatial modelling of extreme precipitation return levels. *Journal of the American Statistical Association*, **102**, 824–840.

Cressie, N. (1993) *Statistics for Spatial Data. Revised Edition.* John Wiley & Sons, Inc.

Damian, D., Sampson, P. and Guttorp, P. (2003) Variance modeling for nonstationary spatial processes with temporal replication. *Journal of Geophysical Research Atmospheres*, **108**, (D24) Art. No. 8778.

Frühwirth-Schnater, S. (1994) Data augmentation and dynamic linear models. *Journal of Time Series Analysis*, **15**, 183–202.

Fuentes, M. and Smith, R. (2000) Modeling nonstationary spatial processes as a convolution of local stationary processes. *Tech. rep.*, North Carolina State University, USA.

Gamerman, D. and Lopes, H. F. (2006) *Markov Chain Monte Carlo - Stochastic Simulation for Bayesian Inference*. Chapman & Hall, 2nd Edition.

Gilks, W. and Wild, P. (1992) Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, **41**, 337–48.

Gneiting, T. (2002) Compactly supported correlation functions. *Journal of Multivariate Analysis*, **83**, 493–508.

Guttorp, P., Fuentes, M. and Sampson, P. (2007) Using transforms to analyze space-time processes. In *Statistics Methods of Spatio-Temporal Systems*, 77–150. V. Isham, B. Finkelstadt and L. Held (editors). Boca Raton: Chapman and Hall/CRC.

Hay, J. (1984) An assessment of the mesoscale variability of solar radiation at the earth's surface. *Solar Energy*, **32**, 425–434.

Haylock, R. and O'Hagan, A. (1996) On inference for outputs of computationally expensive algorithms with uncertainty on the inputs. In *Bayesian Statistics 5*, 629–637. Bernardo, J.M., Berger, J.O., Dawid, A.P. and Smith, A.F.M. (editors).

Higdon, D. (1998) A process-convolution approach to modelling temperatures in the North-Atlantic. *Journal of Environmental Engineering and Science*, **5**, 173–190.

Iovleff, S. and Perrin, O. (2004) Estimating a nonstationary spatial structure using simulated annealing. *Journal of Computational and Graphical Statistics*, **13**, 90–105.

Kim, H.-M., Mallick, B. K. and Holmes, C. C. (2005) Analyzing nonstationary spatial data using piecewise Gaussian processes. *Journal of the American Statistical Association*, **100**, 653–668.

Le, N., Sun, L. and Zidek, J. (2001) Spatial prediction and temporal backcasting for environmental fields having monotone data patterns. *The Canadian Journal of Statistics*, **29**, 529–554.

Meiring, W., Guttorp, P. and Sampson, P. (1998) Space-time estimation of grid-cell hourly ozone levels for assessment of a deterministic model. *Environmental and Ecological Statistics*, **5**, 197–222.

Monestiez, P. and Switzer, P. (1991) Semiparametric estimation of nonstationary spatial covariance models by metric multidimensional scaling. *Tech. rep.*, Department of Statistics - Stanford University, USA.

Paciorek, C. J. and Schervish, M. J. (2006) Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, **17**, 483–506.

Plummer, M., Best, N., Cowles, K. and Vines, K. (2006) CODA: Convergence diagnosis and output analysis for MCMC. *R News*, **6**, 7–11. URL`http://CRAN.R-project.org/doc/Rnews/`.

Sampson, P. (2010) Constructions for nonstationary spatial processes. In *Handbook of Spatial Statistics*, pp. 119–130. Alan E. Gelfand and Peter J. Diggle and Montserrat Fuentes and Peter Guttorp (eds.). Boca Raton: Chapman and Hall/CRC.

Sampson, P., Damian, D., Guttorp, P. and Holland, D. M. (2001) Deformation-based nonstationary spatial covariance modelling and network design. In *Spatio-temporal modelling of environmental processes*, 125–132. Colecion "Treballs D'Informatica I Technologia", Núm. 10, J. Mateu and M. Fuentes (editors). Castellon, Spain: Universitat Jaume I.

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

Schmidt, A. M. and O'Hagan, A. (2003) Bayesian inference for nonstationary spatial covariance structures via spatial deformations. *Journal of the Royal Statistical Society, Series B*, **65**, 743–775.

Stein, M. (1999) *Interpolation of Spatial Data*. Springer-Verlag, New York, Inc.

West, M. and Harrison, P. J. (1997) *Bayesian Forecasting and Dynamic Models*. Springer-Verlag New York, Second Edition.