# Bayesian Nonparametric Spatial Modeling With Dirichlet Processes Mixing

Alan E. Gelfand, Athanasios Kottas and Steven N. MacEachern*

July 15, 2004

**Abstract**

Customary modeling for continuous point-referenced data assumes a Gaussian process which is often taken to be stationary. When such models are fitted within a Bayesian framework, the unknown parameters of the process are assumed to be random so a random Gaussian process results. Here, we propose a novel spatial Dirichlet process mixture model to produce a random spatial process which is neither Gaussian nor stationary.

We first develop a spatial Dirichlet process model for spatial data and discuss its properties. Due to familiar limitations associated with direct use of Dirichlet process models, we introduce mixing through this process against a pure error process. We then examine properties of models created through such Dirichlet process mixing.

In the Bayesian framework, posterior inference is implemented using Gibbs sampling as we detail. Spatial prediction raises interesting questions but can be handled. Finally, we illustrate the approach using simulated data as well as a dataset involving precipitation measurements over the Languedoc-Roussillon region in southern France.

KEY WORDS: Dependent Dirichlet process; Dirichlet process mixture models; Gaussian processes; Markov chain Monte Carlo; nonstationarity; point referenced spatial data; random distribution.

*A.E. Gelfand is Professor, Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708-0251, USA, A. Kottas is Assistant Professor, Department of Applied Mathematics and Statistics, University of California, Santa Cruz, CA 95064, USA and S.N. MacEachern is Professor, Department of Statistics, Ohio State University, Columbus, OH 43210-1247, USA. The research of the third author was supported in part by the National Science Foundation under Award Number DMS-0072526. The authors thank Doris Damian for providing the French precipitation data and two referees for comments that greatly improved the manuscript.

# 1 Introduction

Point-referenced spatial data is collected in a wide range of contexts. Modeling for such data introduces a spatial process specification either for the data directly or for a set of spatial random effects associated with the mean structure for the data, perhaps on a transformed scale. In virtually all of this work the spatial process specification is parametric. In fact, it is usually a Gaussian process (GP) that is often assumed to be stationary. Within a Bayesian framework the resulting model specification can be viewed as hierarchical. See, e.g. Diggle, Tawn and Moyeed (1998) and Ecker and Gelfand (2003).

The parameters of the spatial process are unknown, and so they are assigned prior distributions resulting in a random GP. Since the parameters in the mean structure and in the covariance structure determine the distribution of the GP, we have a finite dimensional model. The fitting of such models using Markov chain Monte Carlo (MCMC) methods is by now fairly straightforward. See, e.g., Agarwal and Gelfand (2002) and further references therein.

Flexible and computationally tractable ways to remove the stationarity assumption have appeared recently. These include spatially varying kernel convolution ideas as in Higdon, Swall and Kern (1999) and Higdon (2002) as well as a local stationarity approach as in Fuentes and Smith (2001). This work is attracting increased attention but is fully parametric and is still within the setting of GPs.

There is a rich literature on nonparametric modeling for the mean structure in a spatial process setting, much of it drawing from the nonparametric regression literature. See, e.g., Stein (1999) and further references therein. Our interest is entirely in nonparametric modeling for the stochastic mechanism producing the spatial dependence structure. In this regard, there are the *nonparametric* variogram fitting approaches, e.g., Shapiro and Botha (1991) and Barry and Ver Hoef (1996). But these do not fully specify the process; they are nonparametric only in the second moment structure. Arguably, the most significant nonparametric spatial contribution is the "deformation" approach of Sampson and Guttorp (1992). The observed locations in the actual (geographic) space are viewed as a nonlinear transformation of locations in a conceptual (deformed) space where the process is assumed stationary and, in fact, isotropic. This approach has been pursued in a Bayesian context by both Damian, Sampson and Guttorp (2001) and Schmidt and O'Hagan (2003). The former introduces random thin plate splines to implement the transformation. The latter proposes a bivariate GP to explain the movements from conceptual locations to observed locations. Though nonstationarity of the process is introduced through a nonparametric specification of the covariance function, both approaches still employ a GP for the likelihood.

A critical component of the deformation setting is the need for replication. Sampson and Guttorp (1992) use the replications to obtain the sample covariance estimate, a *nonparametric* estimate of the process covariance matrix at the observed sites. In general, for a fully nonparametric modeling approach, replication of some form will be required. With only a single observation from a multivariate distribution, a nonparametric model is not viable; we would fall back on the nonstationary parametric modeling approaches mentioned above. In contrast to the deformation approach, the techniques we develop can be used when replication is present, but where there is only a single observation at any given site.

Here, we introduce a completely different approach based upon the Dirichlet process (DP) (Ferguson, 1973, 1974). In the literature, for continuous measurements, DPs have been used to provide random univariate distributions and, in fact, random multivariate distributions. We only require a continuous baseline or centering distribution and a precision parameter value. What we seek here is a random joint distribution for a stochastic process of random variables, i.e., for the uncountable set of random variables where each is associated with a location in a given region, say $D$. As with GPs, we provide this specification through arbitrary finite dimensional distributions, that is, for $(Y(s_1), ..., Y(s_n))$ where $n$ and the set of $s_i$ are arbitrary. The resulting process is nonstationary and the resulting joint distributions are not normal.

In particular, when $n = 1$ we have $\{F(Y(s)) : s \in D\}$ where $F(Y(s))$ denotes the random distribution for $Y(s)$. With regard to this collection of random distributions we would like to achieve the behavior that MacEachern (2000) discussed in terms of dependent Dirichlet processes (DDPs). More precisely, we want the $F(Y(s))$ to be dependent and, as $s \to s_0$, we want the realized $F(Y(s))$ to converge to the realized $F(Y(s_0))$. We will indicate various ways to accomplish this in section 2. Regardless, this enables us to pool information from nearby spatial locations to better estimate, say $F(Y(s_0))$, analogous to the local learning that is inherent in familiar spatial prediction (or kriging) for say, $Y(s_0)$.

The limitations associated with the use of DPs to model random distributions are well known. Most problematic is that the support for a random $F$ is almost surely the family of discrete distributions. In the high-dimensional multivariate setting we envision, such models lack robustness. As is well established (Antoniak, 1974), DP mixing circumvents these limitations.

Again, inference under such modeling is done within a Bayesian framework. Model fitting is done through MCMC. The computation is demanding but becomes a fairly straightforward extension of existing MCMC routines for DPs (see, e.g., Escobar and West, 1998 and MacEachern, 1998). Nonparametric spatial prediction under such modeling is also possible but raises some interesting technical points which we discuss. We illustrate the entire approach with simulated

data arising from a known nonstationary, non-Gaussian bimodal mixture specification. As a result, we can demonstrate the capability of our modeling formulation. We then turn to the analysis of a real dataset which records precipitation measurements from 39 monitoring stations over the Languedoc-Roussillon region in southern France.

Apart from offering a novel modeling formulation for the nonparametric problem, possible advantages offered by our approach are the following. We achieve a process which is both nonstationary and non-Gaussian. We can draw upon the well-developed theory for DP mixing to facilitate interpretation of our analysis. We can implement the required simulation-based model fitting more easily than other replication-based approaches mentioned above can since we can avail ourselves of established strategies for DP mixture models. We can infer about the unknown (random) distribution that is operating at any given location in the region. We can readily extend our approach to multivariate processes and to spatio-temporal processes.

The plan of the paper is as follows. In section 2 we describe the model including approaches for posterior inference and prior specification. Section 3 addresses spatial prediction. Section 4 provides data illustrations. We conclude with a summary and a discussion of possible extensions in section 5. An Appendix includes the details of the MCMC method for posterior inference.

# 2 Spatial Dirichlet process modeling

Section 2.1 develops the model and its properties. Section 2.2 describes an MCMC method for posterior inference (with the details given in the Appendix). An approach for providing prior specification is discussed in section 2.3.

## 2.1 The Modeling Approach

We begin by developing a model for point referenced spatial data assumed to arise as a sample from a realization of a random field (random process) $\{Y(s) : s \in D\}$, $D \subseteq R^d$. Denote by $\boldsymbol{s}^{(n)} = (s_1, ..., s_n)$ the specific distinct locations in $D$ where the observations are collected. When a Gaussian random field is assumed, a multivariate normal specification for the observed data results. In order to allow for deviations from this, arguably restrictive, assumption, we propose a semiparametric model for the random field with an associated induced model for the distribution of $(Y(s_1), ..., Y(s_n))$. As noted in the introduction, we assume that we have available replicate observations at each location and therefore that the full data set consists of the collection of vectors $\mathbf{Y}_t = (Y_t(s_1), ..., Y_t(s_n))'$, $t = 1,...,T$. In fact, we can accommodate imbalance or missingness in the $Y_t(s_i)$ through customary latent variable methods. We provide further clarification

at the end of section 3.

Central to our approach is the DP, a random probability measure on a space of distribution functions defined on some space $\Theta$ (equipped with a $\sigma$-field $\mathcal{B}$). We use $DP(\nu G_0)$ to denote the DP, where $\nu > 0$ is a scalar (precision parameter) and $G_0$ a specified base distribution defined on $(\Theta, \mathcal{B})$. We recall that a random distribution function on $(\Theta, \mathcal{B})$ arising from $DP(\nu G_0)$ is almost surely discrete and admits the representation $\sum_{l=1}^{\infty} \omega_l \delta_{\theta_l}$, where $\delta_z$ denotes a point mass at $z$, $\omega_1 = z_1$, $\omega_l = z_l \prod_{r=1}^{l-1}(1 - z_r)$, $l$ = 2,3,..., with $\{z_r, r = 1,2,...\}$ i.i.d. from Beta$(1, \nu)$ and independently $\{\theta_l, l = 1,2,...\}$ i.i.d. from $G_0$ (Sethuraman, 1994). In this notation $\theta_l$ is assumed to be scalar or perhaps vector valued, the latter case leading to a multivariate DP.

To model $\boldsymbol{Y}_D \equiv \{Y(s) : s \in D\}$, we conceptually extend $\theta_l$ to a realization of a random field by replacing it with $\theta_{l,D} = \{\theta_l(s) : s \in D\}$. For instance, $G_0$ might be a stationary GP with each $\theta_{l,D}$ being a realization from $G_0$, i.e., a surface over $D$. The resulting random process or distribution, $G$, for $\boldsymbol{Y}_D$ is denoted by $\sum_{l=1}^{\infty} \omega_l \delta_{\theta_{l,D}}$ and the construction will be referred to as a spatial Dirichlet process model. The interpretation is that for $\boldsymbol{s}^{(n)}$ as above, $G$ induces a random probability measure $G^{(\boldsymbol{s}^{(n)})}$ on the space of distribution functions for $(Y(s_1), ..., Y(s_n))$. (To simplify notation, we will use $G^{(n)}$ instead of $G^{(\boldsymbol{s}^{(n)})}$ in what follows.) Since $\omega_l$ does not depend upon the location $s \in D$, we have that $G^{(n)} \sim DP(\nu G_0^{(n)})$, where $G_0^{(n)} \equiv G_0^{(\boldsymbol{s}^{(n)})}$ is the $n$-variate distribution for $(Y(s_1), ..., Y(s_n))$ induced by $G_0$ (e.g., an $n$-variate normal if $G_0$ is taken to be a GP).

We next turn to a connection between the spatial DP above and the notion of a dependent Dirichlet process (DDP) as developed by MacEachern (2000). The DDP provides a formal framework within which to describe a stochastic process of random distributions. These distributions are dependent but such that, at each index value, the distribution is a univariate DP. See, e.g., De Iorio et al. (2004) for an illustration in the ANOVA setting. In our setting, $G$ induces a random distribution $G(Y(s))$ for each $s$, hence the set $\mathcal{G}_{\mathcal{D}} \equiv \{G(Y(s)) : s \in D\}$. MacEachern (2000, Theorem 3.1) provides sufficient conditions such that $\mathcal{G}_{\mathcal{D}}$ will be a DDP. In fact, we can envision the possibility of richer spatial DPs by allowing the $z_r$ to vary with $s$, hence the $\omega_l$ to vary with $s$ and thus, connection to more general DDPs. (See Section 5.)

For a stationary $G_0$, i.e., $\mathrm{Cov}(\theta_l(s_i), \theta_l(s_j))$ is a function of $s_i - s_j$ only, the choice of the covariance function determines how smooth process realizations are. Kent (1989), for instance, shows that, if the covariance function admits a second order Taylor-series expansion with remainder that goes to 0 at a rate of $2 + \delta$ for some $\delta > 0$ then $\theta(s_i) - \theta(s_j) \to 0$, almost surely, as $||s_i - s_j|| \to 0$. But then, in the representation of $G$ as $\sum \omega_l \delta_{\theta_{l,D}}$, the continuity of $\theta_{l,D}$ implies that the random marginal distribution of $Y(s_i)$ given $G$, $G(Y(s_i))$, and the random marginal

distribution of $Y(s_j)$ given $G$, $G(Y(s_j))$, are such that the difference between them tends to 0 almost surely, as $||s_i - s_j|| \to 0$. This result is a version of Theorem 3.6 in MacEachern (2000). The implication is that we can learn about $G(Y(s))$ more from data at neighboring locations than from data at locations further away, as in usual spatial prediction.

We digress to note that there are alternative stochastic mechanisms to produce $\theta_{l,D}$ which are continuous. For example, suppose $\theta(s)$ is a parametric function $h(s;\boldsymbol{\gamma})$ which is continuous in $s$ for each $\boldsymbol{\gamma} \in \Gamma$. If $\boldsymbol{\gamma}_l$ is drawn using a distribution over $\Gamma$, then $\theta_{l,D} = \{h(s;\boldsymbol{\gamma}_l) : s \in D\}$ is a random continuous surface over $D$. As a simple example, we might take $h(s;\boldsymbol{\gamma}) = \gamma_0 + \gamma_1 \text{lat}(s) + \gamma_2 \text{lon}(s)$, where $\text{lat}(s)$ and $\text{lon}(s)$ denote the north-south and east-west coordinates of $s$, perhaps after projection, and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2)$. Such a specification in the context of DPs is in the spirit of Müller, Quintana and Rosner (1999). Note that, as a result, the distribution of $\theta_{l,D}$ arises through a finite dimensional distribution specification rather than an uncountable one as in the previous paragraph; the realized $\theta_{l,D}$ will not be as flexible as the previous ones. Also note that, in this case $\theta_l(s) - \theta_l(s') \to 0$ as $||s - s'|| \to 0$ but $\text{Cov}(\theta_l(s), \theta_l(s'))$ is not a function of $s - s'$ as it would be if $\theta_{l,D}$ arose from a stationary process.

Returning to $G$ arising from $G_0$ and $\nu$, note that given $G$, $\text{E}(Y(s) \mid G) = \sum \omega_l \theta_l(s)$ and $\text{Var}(Y(s) \mid G) = \sum \omega_l \theta_l^2(s) - \{\sum \omega_l \theta_l(s)\}^2$. Moreover for a pair of sites $s_i$ and $s_j$,

$$\text{Cov}(Y(s_i), Y(s_j) \mid G) = \sum \omega_l \theta_l(s_i)\theta_l(s_j) - \left\{\sum \omega_l \theta_l(s_i)\right\}\left\{\sum \omega_l \theta_l(s_j)\right\}. \tag{1}$$

Hence, the random process $G$ has heterogeneous variance and is nonstationary. Marginalizing over $G$ simplifies the above expressions. For example, assume a mean zero stationary GP for $G_0$ with variance $\sigma^2$ and correlation function $\rho_\phi(s_i - s_j)$, where the (possibly vector valued) parameter $\phi$ specifies $\rho_\phi(\cdot)$. Then $\text{E}(Y(s)) = 0$, $\text{Var}(Y(s)) = \sigma^2$ and $\text{Cov}(Y(s_i), Y(s_j)) = \sigma^2 \rho_\phi(s_i - s_j)$. Attractively, although $G$ is centered around a stationary process with constant variance, it has nonconstant variance and is nonstationary.

However, the almost sure discreteness of $G$ will be undesirable in practice. This difficulty can be overcome by mixing a pure error process (often referred to as a nugget process, i.e., i.i.d. variables at the locations each having zero mean and variance $\tau^2$) with respect to $G$ to create a random process $F$ which has continuous support. More explicitly, suppose $\theta_D$ given $G$ is a realization from $G$ and $\mathbf{Y}_D - \theta_D$ is a realization from the pure error process. Then, operating formally we find that, marginally, $\mathbf{Y}_D$ arises from the process $F$ which can be defined as the convolution

$$F\left(\mathbf{Y}_D \mid G, \tau^2\right) = \int \mathcal{K}\left(\mathbf{Y}_D - \theta_D \mid \tau^2\right) G\left(d\theta_D\right).$$

*Differentiating* to densities,

$$f\left(\mathbf{Y}_D \mid G, \tau^2\right) = \int k\left(\mathbf{Y}_D - \theta_D \mid \tau^2\right) G\left(d\theta_D\right). \tag{2}$$

Here $\mathcal{K}$ and $k$ denote the joint distribution function and density function, respectively, of the pure error process over $D$. $k$ might denote a $N(0,1)$ or $t_r(0,1)$ density. Hence for any $s$, $f\left(Y(s) \mid G, \tau^2\right) = \int k\left(Y(s) - \theta(s) \mid \tau^2\right) G\left(d\theta(s)\right)$. In other words, $Y(s) = \theta(s) + \epsilon(s)$ where $\theta(s)$ arises from the above spatial DP prior model and $\epsilon(s)$ is $\mathrm{N}(0, \tau^2)$. The customary partitioning into a spatial component and a pure error or nugget component results. Indeed, our modeling *requires* the introduction of a nugget effect. In practice, this is not a limitation. Residuals from the mean would not be expected to be perfectly explained through a spatial story. With our assumed replication, additional variability is introduced so this becomes even more the case. Also, we remark that we are convolving distributions to create the process model rather than convolving process variables to create a process and then extracting the induced distributional model as in Higdon, Swall, and Kern (1999) or Fuentes and Smith (2001).

Evidently, $\tau^2$ could be replaced by $\tau^2 h(s)$ for a known function $h(s)$, e.g., a function of a covariate at $s$. Now, given $\tau^2$, the $\epsilon(s)$ are independent but not identicallly distributed. In fact, again operating formally, (2) could be extended to mix not only on $\theta_D$ but also say on $\tau_D^2 = \left\{\tau^2(s) : s \in D\right\}$. A richer class of models for $\mathbf{Y}_D$ would result. To maintain the partition of $Y(s)$ into a spatial component and an i.i.d. error component would require $\tau^2(s)$ i.i.d. Then, given $\tau^2(s)$, $Y(s) = \theta(s) + \tau(s)z(s)$ where $z(s)$ are i.i.d. with mean 0 and variance 1. Of course, $\tau(s)z(s)$ is just a different error process and, if the distribution of $\tau^2(s)$ were fixed, we could change $k$ accordingly. A nonparametric mixture would require specification for the random measure $G(d\theta(s), d\tau(s))$. In the sequel, we have set $\tau(s) = \tau$ and taken $z(s)$ i.i.d. N(0,1).

For the finite set of locations $(s_1, ..., s_n)$, (2) implies that the joint density of $\mathbf{Y} = (Y(s_1), ..., Y(s_n))'$ given $G^{(n)}$, where $G^{(n)} \sim DP(\nu G_0^{(n)})$, and $\tau^2$ is

$$f\left(\mathbf{Y} \mid G^{(n)}, \tau^2\right) = \int N_n\left(\mathbf{Y} \mid \theta, \tau^2 I_n\right) G^{(n)}\left(d\theta\right), \tag{3}$$

where, to simplify notation, $\theta \equiv \theta^{(\boldsymbol{s}^{(n)})} = (\theta(s_1), ..., \theta(s_n))'$ and $N_p(\cdot \mid \lambda, \Sigma)$ denotes the $p$-variate normal density/distribution (depending on the context) with mean vector $\lambda$ and covariance matrix $\Sigma$. Again, the almost sure representation of $G^{(n)}$ as $\sum \omega_l \delta_{\theta_l}$, where $\theta_l$ is the vector $(\theta_l(s_1), ..., \theta_l(s_n))$, yields that $f(\mathbf{Y} \mid G^{(n)}, \tau^2)$ is almost surely of the form $\sum_{l=1}^{\infty} \omega_l N_n(\mathbf{Y} \mid \theta_l, \tau^2 I_n)$, i.e., a countable location mixture of normals. In fact, assuming the existence of expectations given $G^{(n)}$ and $\tau^2$ (sufficient conditions for which can be obtained and relate to $G_0^{(n)}$), using Fubini's theorem we obtain that $\mathrm{E}(\mathbf{Y} \mid G^{(n)}, \tau^2) = \sum \omega_l \theta_l$ and the covariance ma-

trix $\Sigma_{\mathbf{Y}} \mid G^{(n)}, \tau^2 = \tau^2 I_n + \Sigma_\theta^{(\boldsymbol{s}^{(n)})}$, where $(\Sigma_\theta^{(\boldsymbol{s}^{(n)})})_{i,j} = \mathrm{Cov}(\theta(s_i), \theta(s_j) \mid G^{(n)})$ the covariance arising from (1).

Note that we would usually add a constant mean term $\mu$, or more generally, a regression term $X'\boldsymbol{\beta}$, to the kernel of the mixture model in (3) leading to

$$f\left(\mathbf{Y} \mid G^{(n)}, \boldsymbol{\beta}, \tau^2\right) = \int N_n\left(\mathbf{Y} \mid X'\boldsymbol{\beta} + \theta, \tau^2 I_n\right) G^{(n)}\left(d\theta\right), \tag{4}$$

i.e., $\mathrm{E}(\mathbf{Y} \mid G^{(n)}, \boldsymbol{\beta}, \tau^2) = X'\boldsymbol{\beta} + \sum \omega_l \theta_l$. Here $X$ is a $p \times n$ matrix whose $(i,j)$th element is the value of the $i$th covariate at the $j$th location, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients.

Next we bring in the data $\mathbf{Y}_t = (Y_t(s_1), ..., Y_t(s_n))'$ and $X_t$, where $X_t$ is the matrix of covariate values for the $t$th replication, $t = 1,...,T$. (See Remark 2 below.) We assume $\mathbf{Y}_t$ are independent, given $X_t$, $f(\mathbf{Y}_t \mid G^{(n)}, \boldsymbol{\beta}, \tau^2)$ as in (4). A DP prior is placed on $G^{(n)}$, $G^{(n)} \sim DP(\nu G_0^{(n)})$ (induced by the spatial DP prior for $G$ in (2)), with $G_0^{(n)}$ being a multivariate normal with mean zero and covariance matrix $\sigma^2 H_n(\phi)$ (induced by a mean zero GP $G_0$ with variance $\sigma^2$ and correlation function $\rho_\phi(\cdot,\cdot)$ such that $(H_n(\phi))_{i,j} = \rho_\phi(s_i, s_j)$). In view of the earlier discussion, a stationary $\rho_\phi(s_i - s_j)$, possibly isotropic $\rho_\phi(\|s_i - s_j\|)$, specification for $H_n(\phi)$ will be assumed. For convenience, we illustrate throughout with an exponential correlation function $\rho_\phi(\| \cdot \|) = \exp(-\phi \| \cdot \|)$, $\phi > 0$. The full Bayesian model is completed by placing (independent) priors on $\boldsymbol{\beta}$, $\tau^2$, $\nu$, $\sigma^2$ and $\phi$. If we associate with each $\mathbf{Y}_t$ a $\theta_t = (\theta_t(s_1), ..., \theta_t(s_n))'$, with the $\theta_t$, $t = 1,...,T$, being i.i.d. realizations from $G^{(n)}$, the following semiparametric hierarchical model emerges (hereafter we use the bracket notation for marginal and conditional densities)

$$\begin{array}{ccc}
\mathbf{Y}_t \mid \theta_t, \boldsymbol{\beta}, \tau^2 & \overset{ind.}{\sim} & N_n(\mathbf{Y}_t \mid X_t'\boldsymbol{\beta} + \theta_t, \tau^2 I_n), t = 1, ..., T \\
\theta_t \mid G^{(n)} & \overset{i.i.d.}{\sim} & G^{(n)}, t = 1, ..., T \\
G^{(n)} \mid \nu, \sigma^2, \phi & \sim & DP(\nu G_0^{(n)}); G_0^{(n)}(\cdot \mid \sigma^2, \phi) = N_n(\cdot \mid 0_n, \sigma^2 H_n(\phi)) \\
\boldsymbol{\beta}, \tau^2 & \sim & N_p(\boldsymbol{\beta} \mid \boldsymbol{\beta}_0, \Sigma_{\boldsymbol{\beta}}) \times \mathrm{IGamma}(\tau^2 \mid a_\tau, b_\tau) \\
\nu, \sigma^2, \phi & \sim & \mathrm{Gamma}(\nu \mid a_\nu, b_\nu) \times \mathrm{IGamma}(\sigma^2 \mid a_\sigma, b_\sigma) \times [\phi],
\end{array} \tag{5}$$

the prior on $\phi$ depending on the specific form of the covariance function in $H_n(\phi)$. Illustratively, we adopt the exponential covariance function with $\phi$ as the decay parameter, yielding $(H_n(\phi))_{i,j} = \exp(-\phi\|s_i - s_j\|)$, $\phi > 0$. We take a uniform prior for $\phi$ on $[0, b_\phi]$. The hyperparameters of the priors for $\boldsymbol{\beta}$, $\tau^2$, $\nu$, $\sigma^2$ and $\phi$ are fixed. We discuss their specification in section 2.3. Inference for the model parameters and, most importantly, spatial prediction at new locations requires the joint posterior of all parameters of the DP mixture model (5).

We conclude with two remarks regarding the model in (5).

Remark 1: Consider the stationary GP model that results from (5) when $\nu \to \infty$. In this case,

the $\theta_t$, conditionally on $\sigma^2$ and $\phi$, become i.i.d. $G_0^{(n)}$; we have a different realization of the base spatial GP $G_0$ for each replication. In section 4.1 we compare this GP mixture model with the spatial DP mixture model (5).

Remark 2: If $X_t(s) = X(s)$ we have i.i.d. replications of the spatial process. If $X(s)$ is $p \times 1$, we need to conceptualize $p$ surfaces over $s \in D$ in order to define $Y(s_i)$ at each of the $s_i$. With interest in prediction at $s_0$, we would need $X(s_0)$. For explaining say, pollution levels, examples of such surfaces include elevation and perhaps inverse distance from an emissions source. Time dependence of covariates presents no problem for fitting the model or for making predictions. If $X_t(s)$ varies with $t$ and say, $t$ indexes time, $X_t(s)$ might be the temperature at location $s$ at time $t$. Temporal dependence in the $X_t$ surfaces induces temporal dependence in the $Y_t$. Additionally, we can introduce a temporal structure to $\boldsymbol{\beta}$, extending the notation to $\boldsymbol{\beta}_t$ and, for instance, modeling through an autoregressive specification. This induces temporal dependence in the $Y_t$ through changing relationships with the covariates.

## 2.2 Simulation-based Model Fitting

Simulation based model fitting for DP mixture models proceeds by marginalizing over the random mixing distribution resulting in a finite dimensional parameter vector. Model (5) is a *conjugate* DP mixture model, the term referring to the fact that $G_0^{(n)}$ is a conjugate prior for the likelihood in the first stage of the hierarchical model. Gibbs sampling for such DP mixture models has been developed by Escobar (1994), West, Müller and Escobar (1994), Escobar and West (1995) and Bush and MacEachern (1996).

After the marginalization over $G^{(n)}$, the joint posterior becomes $[\boldsymbol{\theta}, \boldsymbol{\beta}, \tau^2, \nu, \sigma^2, \phi \mid \text{data}]$, where $\boldsymbol{\theta} = (\theta_1, ..., \theta_T)$ and data is $\{(\mathbf{Y}_t, X_t) : t = 1, ..., T\}$. The discreteness of the random distribution $G^{(n)}$ plays central role in the implementation of the algorithm since with positive probability some of the $\theta_t$ are identical (Antoniak, 1974), hence with positive probability a clustering of the $\theta_t$ is induced. Denote by $T^*$ the number of distinct elements, i.e., the number of clusters, of the vector $(\theta_1, ..., \theta_T)$ and by $\boldsymbol{\theta}^* = (\theta_1^*, ..., \theta_{T^*}^*)$ the vector of distinct $\theta_t$. Since $G_0^{(n)}$ is continuous, the vector of configuration indicators $\mathbf{w} = (w_1, ..., w_T)$ is defined by $w_t = j$ if and only if $\theta_t = \theta_j^*$ for $t$ = 1,...,$T$, and this vector determines the clusters. Let $T_j$ be the number of members of cluster $j$, i.e., $T_j = \#\{t : w_t = j\}$, $j$ = 1,...,$T^*$ (with $\sum_{j=1}^{T^*} T_j = T$). We note here that each $\theta_t$ is a function of $w_t$ and $\boldsymbol{\theta}^*$, each $w_t$ is a function of $\theta_t$ and $\boldsymbol{\theta}^*$, and each $\theta_j^*$ as well as $T^*$ are functions of $\boldsymbol{\theta}$. Hence $(\boldsymbol{\theta}^*, \mathbf{w})$ is an equivalent representation of $\boldsymbol{\theta}$. In particular, it is straightforward to convert the posterior draws from $[\boldsymbol{\theta}, \boldsymbol{\beta}, \tau^2, \nu, \sigma^2, \phi \mid \text{data}]$ to posterior draws from $[\boldsymbol{\theta}^*, \mathbf{w}, T^*, \boldsymbol{\beta}, \tau^2, \nu, \sigma^2, \phi \mid \text{data}]$. The details of a Gibbs sampler to draw

from $[\boldsymbol{\theta}, \beta, \tau^2, \nu, \sigma^2, \phi \mid \text{data}]$ are given in the Appendix.

### 2.3 Prior Specification

Turning to prior specifications, first, consider $\phi$ which has a uniform prior on $(0, b_\phi]$. Although it is not directly connected to the random nonstationary $G$, $\phi$ determines the range $(3/\phi)$ of the stationary baseline GP $G_0$. (For an isotropic covariance function that decreases to 0 as distance goes to $\infty$, the range is the distance at which correlation becomes 0.05.) We use this range interpretation to specify $b_\phi$. In particular, the range is usually presumed to be at least one tenth (and often closer to half) of the maximum interpoint distance over the region D. But since $3/b_\phi < 3/\phi$, we conservatively specify $3/b_\phi = d \max ||s_i - s_j||$ for a small value of $d$ ($d = 0.01$ was used for the examples in section 4). Based on experimentation with several data sets (including the ones in section 4), we have found that this approach yields a rather noninformative prior for $\phi$ as its posterior mass is always concentrated on values quite smaller than $b_\phi$.

Regarding $\boldsymbol{\beta}$, $\tau^2$ and $\sigma^2$, we set $a_\tau = a_\sigma = 2$ yielding priors for $\tau^2$ and $\sigma^2$ with means $b_\tau$ and $b_\sigma$, respectively, and infinite variance. We simplify $\boldsymbol{\beta}$ to a constant mean $\mu$ with prior $N(\mu \mid m, v^2)$. (We have no covariates in our data examples and the ensuing discussion can be easily extended to accommodate $\boldsymbol{\beta}$ if we did.) Letting $\nu \to \infty$, the marginal mean and variance for $Y_t(s)$ are $m$ and $b_\tau + b_\sigma + v^2$, respectively. Now suppose we compute rough estimates for the center and range of the set of $Y_t(s)$, say $c$ and $r$. Based on the above, we can take $m = c$ and $b_\tau + b_\sigma + v^2 = 3(r/4)^2$ (or $3(r/6)^2$) using 3 as an additional variance inflation factor. Splitting $3(r/4)^2$ equally between the three summands, we can set $b_\tau = b_\sigma = v^2 = (r/4)^2$. In this way, the data is used only to provide rough centering for the priors for $\mu$, $\tau^2$, and $\sigma^2$.

Finally, the parameter $\nu$ in the DP mixture model (5) controls the distribution of the number of distinct elements of the vector $(\theta_1, ..., \theta_T)$ and hence the number of distinct components of the mixture. (See Antoniak, 1974, and Escobar and West, 1995, for details.) Therefore, prior information about the number of components can be incorporated through the prior for $\nu$. In the absence of strong prior information in this direction, it appears natural to choose values for $a_\nu$ and $b_\nu$ that yield Gamma priors for $\nu$ that place mass both on small and large values. (In Section 4 we use a rather dispersed Gamma(3,.005) prior. Experimentation with other choices revealed little prior sensitivity.) Practical experience with model (5) suggests that there is posterior learning for $\nu$ when sample sizes are moderate to large (e.g., $T > 50$).

## 3 Bayesian nonparametric spatial prediction

Arguably, the main objective of spatial modeling is prediction (e.g., kriging) of a random realization of the process at locations where the random field $\{Y(s) : s \in D\}$ is not observed. Let $\tilde{\boldsymbol{s}}^{(m)} = (\tilde{s}_1, ..., \tilde{s}_m)$ be a collection of such new locations. Denoting by $\tilde{\mathbf{Y}}_0 = (Y_0(\tilde{s}_1), ..., Y_0(\tilde{s}_m))'$ the prediction at the new locations in $\tilde{\boldsymbol{s}}^{(m)}$, interest lies in the posterior predictive distribution $[\tilde{\mathbf{Y}}_0 \mid \tilde{X}_0, \text{data}]$, where, if covariates are employed, $\tilde{X}_0$ is the $p \times m$ matrix of covariate values associated with $\tilde{\mathbf{Y}}_0$. As the notation suggests, it is useful to recognize that in such prediction we are not attempting to interpolate $Y_t(\tilde{s}_i)$ for one of the T replications. Rather, we are predicting for a new replication of the spatial process (hence the subscripting by "0") with associated new covariate surfaces, which yield the matrix $\tilde{X}_0$ at the new sites and a $p \times n$ matrix $X_0$ at the observed sites.

To motivate the approach that yields $[\tilde{\mathbf{Y}}_0 \mid \tilde{X}_0, \text{data}]$, consider first model (5), for the observed data at the locations in $\boldsymbol{s}^{(n)}$,

$$\prod_{t=1}^{T}[\mathbf{Y}_t \mid \theta_t, \boldsymbol{\beta}, \tau^2] \prod_{t=1}^{T}[\theta_t \mid G^{(n)}][G^{(n)} \mid \nu, \sigma^2, \phi][\boldsymbol{\beta}][\tau^2][\nu][\sigma^2][\phi], \tag{6}$$

with the distributional specifications given in (5). After integrating out $G^{(n)}$, the MCMC algorithm, detailed in section 2.2 and the Appendix, yields posterior samples $(\boldsymbol{\theta}_b, \boldsymbol{\beta}_b, \tau_b^2, \nu_b, \sigma_b^2, \phi_b)$, $b = 1, ..., B$, from $[\boldsymbol{\theta}, \boldsymbol{\beta}, \tau^2, \nu, \sigma^2, \phi \mid \text{data}]$, or, equivalently samples $(\boldsymbol{\theta}_b^*, \mathbf{w}_b, T_b^*, \boldsymbol{\beta}_b, \tau_b^2, \nu_b, \sigma_b^2, \phi_b)$, $b = 1, ..., B$, from $[\boldsymbol{\theta}^*, \mathbf{w}, T^*, \boldsymbol{\beta}, \tau^2, \nu, \sigma^2, \phi \mid \text{data}]$.

Using the structure of DPs, for a new $\theta_0 = (\theta_0(s_1), ..., \theta_0(s_n))'$, associated with a new $\mathbf{Y}_0 = (Y_0(s_1), ..., Y_0(s_n))'$ at the locations in $\boldsymbol{s}^{(n)}$ where we observe data, we have

$$[\theta_0 \mid \boldsymbol{\theta}^*, \mathbf{w}, T^*, \nu, \sigma^2, \phi] = \frac{\nu}{\nu + T} G_0^{(n)}(\theta_0 \mid \sigma^2, \phi) + \frac{1}{\nu + T} \sum_{j=1}^{T^*} T_j \delta_{\theta_j^*}(\theta_0). \tag{7}$$

Moreover, the posterior predictive distribution for $\mathbf{Y}_0$ is given by

$$\begin{aligned} [\mathbf{Y}_0 \mid X_0, \text{data}] &= \int [\mathbf{Y}_0 \mid \boldsymbol{\theta}^*, \mathbf{w}, T^*, \boldsymbol{\beta}, \tau^2, \nu, \sigma^2, \phi][\boldsymbol{\theta}^*, \mathbf{w}, T^*, \boldsymbol{\beta}, \tau^2, \nu, \sigma^2, \phi \mid \text{data}] \\ &= \int\int [\mathbf{Y}_0 \mid \theta_0, \boldsymbol{\beta}, \tau^2][\theta_0 \mid \boldsymbol{\theta}^*, \mathbf{w}, T^*, \nu, \sigma^2, \phi][\boldsymbol{\theta}^*, \mathbf{w}, T^*, \boldsymbol{\beta}, \tau^2, \nu, \sigma^2, \phi \mid \text{data}]. \end{aligned} \tag{8}$$

The second integral expression in (8) indicates how a sample $(\mathbf{Y}_{0b}, b = 1, ..., B)$ from $[\mathbf{Y}_0 \mid X_0, \text{data}]$ can be obtained using the MCMC output. For each $b = 1, ..., B$, we first draw $\theta_{0b}$ from $[\theta_0 \mid \boldsymbol{\theta}_b^*, \mathbf{w}_b, T_b^*, \nu_b, \sigma_b^2, \phi_b]$, using (7), and then draw $\mathbf{Y}_{0b}$ from $N_n(X_0'\boldsymbol{\beta}_b + \theta_{0b}, \tau_b^2 I_n)$.

To further highlight the mixture structure, note that, using (7), we can also write

$$[\mathbf{Y}_0 \mid X_0, \text{data}] = \int (\tfrac{\nu}{\nu+T} N_n(\mathbf{Y}_0 \mid X_0'\boldsymbol{\beta}, \tau^2 I_n + \sigma^2 H_n(\phi)) + \tfrac{1}{\nu+T} \sum_{j=1}^{T^*} T_j N_n(\mathbf{Y}_0 \mid X_0'\boldsymbol{\beta} + \theta_j^*, \tau^2 I_n)) \\ [\boldsymbol{\theta}^*, \mathbf{w}, T^*, \boldsymbol{\beta}, \tau^2, \nu, \sigma^2, \phi \mid \text{data}]. \tag{9}$$

The integrand in (9) is a mixture with $T^*+1$ components, the last $T^*$ components (that dominate when $\nu$ is small relative to $T$) yielding a discrete location mixture of $n$-variate normals with the locations defined through the distinct $\theta_j^*$. The posterior predictive density for $\mathbf{Y}_0$ is obtained by averaging this mixture with respect to the posterior of $\boldsymbol{\theta}^*$, $T^*$ and all other parameters.

Turning next to prediction at the new locations $\tilde{\boldsymbol{s}}^{(m)}$, if we were able to observe replicates $\tilde{\mathbf{Y}}_t = (Y_t(\tilde{s}_1), ..., Y_t(\tilde{s}_m))'$, $t = 1,...,T$, the analogous expression to (6) for the full Bayesian model would be

$$\prod_{t=1}^{T} [\mathbf{Y}_t \mid \theta_t, \boldsymbol{\beta}, \tau^2] \prod_{t=1}^{T} [\tilde{\mathbf{Y}}_t \mid \tilde{\theta}_t, \boldsymbol{\beta}, \tau^2] \prod_{t=1}^{T} [(\theta_t, \tilde{\theta}_t) \mid G^{(n+m)}] [G^{(n+m)} \mid \nu, \sigma^2, \phi] [\boldsymbol{\beta}] [\tau^2] [\nu] [\sigma^2] [\phi], \tag{10}$$

where $\tilde{\theta}_t = (\theta_t(\tilde{s}_1), ..., \theta_t(\tilde{s}_m))'$ is the parameter vector associated with $\tilde{\mathbf{Y}}_t$. This is the model that corresponds to the complete data vector $\left\{((\mathbf{Y}_t, \tilde{\mathbf{Y}}_t), (X_t, \tilde{X}_t)) : t = 1, ..., T\right\}$, over the $n+m$ locations $(\boldsymbol{s}^{(n)}, \tilde{\boldsymbol{s}}^{(m)})$, requiring partial (of dimension $n+m$) random realizations $G^{((\boldsymbol{s}^{(n)}, \tilde{\boldsymbol{s}}^{(m)}))} \equiv G^{(n+m)}$ from the spatial DP.

Note that the second term in (10) can be eliminated by integrating out the latent (unobserved) $\tilde{\mathbf{Y}}_t$, $t = 1,...,T$. Moreover, marginalizing over $G^{(n+m)}$ we obtain

$$\prod_{t=1}^{T} [\mathbf{Y}_t \mid \theta_t, \boldsymbol{\beta}, \tau^2] [(\theta_1, \tilde{\theta}_1), ..., (\theta_T, \tilde{\theta}_T) \mid \nu, \sigma^2, \phi] [\boldsymbol{\beta}] [\tau^2] [\nu] [\sigma^2] [\phi]. \tag{11}$$

The second term in (11) can be obtained using the DP structure. It involves point masses and a continuous part corresponding to $G_0^{(n+m)} = N_{n+m}(\cdot \mid 0_{n+m}, \sigma^2 H_{n+m}(\phi))$, where, analogous to $H_n(\phi)$, $H_{n+m}(\phi)$ is the $(n+m) \times (n+m)$ matrix with elements that depend on locations in $\boldsymbol{s}^{(n)}$ and $\tilde{\boldsymbol{s}}^{(m)}$. In particular, the $(\theta_t, \tilde{\theta}_t)$ are no longer independent in (11). However, having integrated $G^{(n+m)}$ out, the clusters $(\boldsymbol{\theta}^*, \tilde{\boldsymbol{\theta}}^*)$ in $(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$, where $\tilde{\boldsymbol{\theta}}^* = (\tilde{\theta}_1^*, ..., \tilde{\theta}_{T^*}^*)$ denotes the $T^*$ distinct values of $\tilde{\theta}_t$ in $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, ..., \tilde{\theta}_T)$, are i.i.d. $G_0^{(n+m)}$. In addition, $(\boldsymbol{\theta}^*, \tilde{\boldsymbol{\theta}}^*)$, $\mathbf{w}$ and $T^*$ specify $(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$. Hence (11) can be written as

$$\prod_{t=1}^{T} [\mathbf{Y}_t \mid \theta_t, \boldsymbol{\beta}, \tau^2] [(\boldsymbol{\theta}^*, \tilde{\boldsymbol{\theta}}^*) \mid T^*, \sigma^2, \phi] [\mathbf{w}, T^* \mid \nu, \sigma^2, \phi] [\boldsymbol{\beta}] [\tau^2] [\nu] [\sigma^2] [\phi],$$

where $[(\boldsymbol{\theta}^*, \tilde{\boldsymbol{\theta}}^*) \mid T^*, \sigma^2, \phi] = \prod_{j=1}^{T^*} N_{n+m}(\theta_j^*, \tilde{\theta}_j^* \mid 0_{n+m}, \sigma^2 H_{n+m}(\phi))$. Therefore, finally, model (10) can be expressed as

$$\prod_{t=1}^{T}[\mathbf{Y}_t \mid \theta_t, \boldsymbol{\beta}, \tau^2] \prod_{j=1}^{T^*}[\tilde{\theta}_j^* \mid \theta_j^*, \sigma^2, \phi] \prod_{j=1}^{T^*} N_n(\theta_j^* \mid 0_n, \sigma^2 H_n(\phi))[\mathbf{w}, T^* \mid \nu, \sigma^2, \phi][\boldsymbol{\beta}][\tau^2][\nu][\sigma^2][\phi], \tag{12}$$

where $[\tilde{\theta}_j^* \mid \theta_j^*, \sigma^2, \phi]$ is the conditional ($m$-variate normal) distribution for $\tilde{\theta}_j^*$ given $\theta_j^*$ emerging from $N_{n+m}(\theta_j^*, \tilde{\theta}_j^* \mid 0_{n+m}, \sigma^2 H_{n+m}(\phi))$.

Based on (12), the joint posterior $[(\boldsymbol{\theta}^*, \tilde{\boldsymbol{\theta}}^*), \mathbf{w}, T^*, \boldsymbol{\beta}, \tau^2, \nu, \sigma^2, \phi \mid \text{data}]$ can be decomposed as follows

$$\prod_{j=1}^{T^*}[\tilde{\theta}_j^* \mid \theta_j^*, \sigma^2, \phi][\boldsymbol{\theta}^*, \mathbf{w}, T^*, \boldsymbol{\beta}, \tau^2, \nu, \sigma^2, \phi \mid \text{data}], \tag{13}$$

where $[\boldsymbol{\theta}^*, \mathbf{w}, T^*, \boldsymbol{\beta}, \tau^2, \nu, \sigma^2, \phi \mid \text{data}]$ is the posterior corresponding to model (6) that can be sampled, using only the observed data, employing the MCMC algorithm of section 2.2. (Note that, excluding the second term, model (12) is the one that arises from (6) once we marginalize over $G^{(n)}$.) Hence, according to (13), we can sample the joint posterior $[(\boldsymbol{\theta}^*, \tilde{\boldsymbol{\theta}}^*), \mathbf{w}, T^*, \boldsymbol{\beta}, \tau^2, \nu, \sigma^2, \phi \mid \text{data}]$ after fitting model (6) by drawing, for each $j = 1, ..., T_b^*$, $\tilde{\theta}_{jb}^*$ from $[\tilde{\theta}_j^* \mid \theta_{jb}^*, \sigma_b^2, \phi_b]$, $b = 1, ..., B$. For any number $m$ of new locations, the additional sampling required is from $m$-variate normal distributions.

Now the method to obtain the posterior predictive distribution for $(\mathbf{Y}_0, \tilde{\mathbf{Y}}_0)$ extends the previous one for $[\mathbf{Y}_0 \mid X_0, \text{data}]$. Analogous to (8), the expression for $[(\mathbf{Y}_0, \tilde{\mathbf{Y}}_0) \mid (X_0, \tilde{X}_0), \text{data}]$ is given by

$$\begin{aligned}[(\mathbf{Y}_0, \tilde{\mathbf{Y}}_0) \mid (X_0, \tilde{X}_0), \text{data}] \;=\; & \int\int [(\mathbf{Y}_0, \tilde{\mathbf{Y}}_0) \mid (\theta_0, \tilde{\theta}_0), \boldsymbol{\beta}, \tau^2][(\theta_0, \tilde{\theta}_0) \mid (\boldsymbol{\theta}^*, \tilde{\boldsymbol{\theta}}^*), \mathbf{w}, T^*, \nu, \sigma^2, \phi] \\ & [(\boldsymbol{\theta}^*, \tilde{\boldsymbol{\theta}}^*), \mathbf{w}, T^*, \boldsymbol{\beta}, \tau^2, \nu, \sigma^2, \phi \mid \text{data}],\end{aligned}$$

where $\tilde{\theta}_0 = (\theta_0(\tilde{s}_1), ..., \theta_0(\tilde{s}_m))'$ is associated with $\tilde{\mathbf{Y}}_0$,

$$[(\theta_0, \tilde{\theta}_0) \mid (\boldsymbol{\theta}^*, \tilde{\boldsymbol{\theta}}^*), \mathbf{w}, T^*, \nu, \sigma^2, \phi] = \tfrac{\nu}{\nu+T} G_0^{(n+m)}((\theta_0, \tilde{\theta}_0) \mid \sigma^2, \phi) + \tfrac{1}{\nu+T} \sum_{j=1}^{T^*} T_j \delta_{(\theta_j^*, \tilde{\theta_j}^*)}(\theta_0, \tilde{\theta}_0) \tag{14}$$

and $[(\mathbf{Y}_0, \tilde{\mathbf{Y}}_0) \mid (\theta_0, \tilde{\theta}_0), \boldsymbol{\beta}, \tau^2] = N_{n+m}(\cdot \mid (X_0, \tilde{X}_0)'\boldsymbol{\beta} + (\theta_0, \tilde{\theta}_0), \tau^2 I_{n+m})$.

Having obtained samples $((\boldsymbol{\theta}_b^*, \tilde{\boldsymbol{\theta}}_b^*), \mathbf{w}_b, T_b^*, \boldsymbol{\beta}_b, \tau_b^2, \nu_b, \sigma_b^2, \phi_b)$, $b = 1, ..., B$, from the posterior $[(\boldsymbol{\theta}^*, \tilde{\boldsymbol{\theta}}^*), \mathbf{w}, T^*, \boldsymbol{\beta}, \tau^2, \nu, \sigma^2, \phi \mid \text{data}]$, as described above, to sample $[(\mathbf{Y}_0, \tilde{\mathbf{Y}}_0) \mid (X_0, \tilde{X}_0), \text{data}]$, we draw, for each $b = 1, ..., B$, $(\theta_{0b}, \tilde{\theta}_{0b})$ from $[(\theta_0, \tilde{\theta}_0) \mid (\boldsymbol{\theta}_b^*, \tilde{\boldsymbol{\theta}}_b^*), \mathbf{w}_b, T_b^*, \nu_b, \sigma_b^2, \phi_b]$, based on (14),

and then draw $(\mathbf{Y}_{0b}, \tilde{\mathbf{Y}}_{0b})$ from $[(\mathbf{Y}_0, \tilde{\mathbf{Y}}_0) \mid (\theta_{0b}, \tilde{\theta}_{0b}), \boldsymbol{\beta}_b, \tau_b^2]$. The resulting samples $(\mathbf{Y}_{0b}, \tilde{\mathbf{Y}}_{0b})$, $b = 1, ..., B$, can be used to study the entire posterior predictive distribution at any combination of locations in $\boldsymbol{s}^{(n)}$ and $\tilde{\boldsymbol{s}}^{(m)}$. We provide several illustrations in section 4.

Regarding implementation, the MCMC output can either be stored and used to predict at the new locations or a draw from $[(\mathbf{Y}_0, \tilde{\mathbf{Y}}_0) \mid (X_0, \tilde{X}_0), \text{data}]$ can be obtained at each iteration of the MCMC (after a suitable burn-in period). The latter option yields considerable savings in storage but necessitates specification of the number of and choice of new locations where prediction is sought.

Consideration of the prediction problem clarifies how to fit the model when some data are missing. In this case, we partition the vector $\mathbf{Y}_t$ into two pieces, the observed portion $\mathbf{Y}_{t,o}$, and the unobserved portion $\mathbf{Y}_{t,u}$, the latter playing a role analogous to $\tilde{\mathbf{Y}}_t$ in (10). Taking $q$ as the dimension of $\mathbf{Y}_{t,u}$, the conditional distribution $[\mathbf{Y}_{t,u}|\theta_t, \boldsymbol{\beta}, \tau^2]$ is $N_q(X'_{t,u}\boldsymbol{\beta} + \theta_{t,u}, \tau^2 I_q)$, where $\theta_t = (\theta_{t,o}, \theta_{t,u})$ and $X_{t,u}$ is the $p \times q$ matrix of covariate values associated with $\mathbf{Y}_{t,u}$. The algorithm in the Appendix is augmented with a step where the $\mathbf{Y}_{t,u}$ are generated for all replicates $t$ with missing data. All other generations remain unchanged. This development also provides a means of interpolation of $\tilde{\mathbf{Y}}_t$ or $\tilde{\theta}_t$ for each of the $T$ replicates, allowing us to address the within-replicate kriging problem.

# 4 Data illustrations

We illustrate the fitting and performance of our semiparametric spatial modeling approach using both simulated data (section 4.1) and real data (section 4.2). The real data is that used by Meiring et al. (1997) and Damian, Sampson and Guttorp (2001). Unfortunately, it has no covariates. For our simulation, introduction of some covariate structure would be arbitrary and apart from our primary objectives which are to retrieve multimodality and nonstationarity in analyzing the spatial effects in our model. For all the examples, we followed the suggestions in section 2.3 for prior specification.

## 4.1 Simulation Experiment

We can propose interesting nonstationary, non-Gaussian models to simulate from using two-component mixtures of independent GPs. The general version sets process $k$ to be a GP with constant mean $\mu_k$ and covariance function $\sigma(s)\sigma(s') \exp(-\phi_k||s - s'||)$, $k = 1, 2$, where process 1 is sampled with probability $\alpha$ and process 2 with probability $1 - \alpha$. Hence, for the resulting

process $Z_D$, $\mathrm{E}(Z(s)) = \alpha\mu_1 + (1-\alpha)\mu_2$ and

$$\begin{aligned}\mathrm{Cov}(Z(s), Z(s')) &= \alpha(1-\alpha)(\mu_1-\mu_2)^2 + \\ &\quad \sigma(s)\sigma(s')\left\{\alpha\exp(-\phi_1||s-s'||) + (1-\alpha)\exp(-\phi_2||s-s'||)\right\}.\end{aligned}$$

A convenient choice for $\sigma^2(s)$ is $\sigma^2(s) = \sigma^2\left\{(\mathrm{lat}(s) - \mathrm{midlat})^2 + (\mathrm{lon}(s) - \mathrm{midlon})^2\right\}$ where $\mathrm{lat}(s)$ and $\mathrm{lon}(s)$ denote the latitude and longitude (or projections thereof) for location $s$ and $\mathrm{midlat} = (\max \mathrm{lat}(s) + \min \mathrm{lat}(s))/2$ and $\mathrm{midlon} = (\max \mathrm{lon}(s) + \min \mathrm{lon}(s))/2$. Since $\sigma(s) \to 0$ as $s \to$ (midlat,midlon), to ensure that there will be a far-from-degenerate spatial term associated with each location, we set $\tilde{\sigma}(s) = \max(\sigma(s), 1)$. Finally, we generate the data from the process $Z(s) + e(s)$ where $e(s)$ is a pure error process with variance $\tau^2$.

We simulated 75 replications at the 39 locations given in Figure 1. In fact, these locations are identical to the 39 sites in the Languedoc-Roussillon region which provide the French precipitation data for section 4.2. We have explored simulation using several choices in the above specification. The two cases we report here are case I: $\mu_1 = \mu_2 (= 0)$, $\phi_1 = \phi_2 (= 0.0025)$, $\sigma = 0.0025$ and $\tau^2 = 1$ yielding a nonstationary GP, and case II: $\tilde{\sigma}(s) = \sigma (= 0.5)$, $\phi_1 = \phi_2 = (0.0025)$, $\tau^2 = 0.5$, $\mu_1 = -2$, $\mu_2 = 2$ and $\alpha = 0.75$ yielding a stationary process which is non-Gaussian and, in fact, has bimodal univariate and bivariate densities.

We note that neither of these cases is included in our specification (5). The parameter values used to generate the data under cases I and II have no connection to any of the parameters in (5). Hence our simulation study focuses on comparison of posterior mean covariances with sample covariances, comparison of posterior predictive densities with the data or true sampling densities, both in univariate and bivariate fashion, and demonstration of and assessment of nonstationarity.

Regarding case I, in Figure 2 we plot posterior predictive densities for six sites where data were generated as well as six new sites, $\tilde{s}_1 = (7250, 18500)$, $\tilde{s}_2 = (6316, 17452)$, $\tilde{s}_3 = (6000, 18200)$, $\tilde{s}_4 = (6298, 18245)$, $\tilde{s}_5 = (7250, 18870)$, $\tilde{s}_6 = (7500, 19000)$ (denoted by a, b, c, d, e, f, respectively, in Figure 1). Predictive inference is arguably quite accurate considering the fairly small sample size. Of course, the more general mixture model, described below (2), would capture more successfully the smaller variances in the center of the region (e.g., sites $s_{29}$ and $\tilde{s}_1$). Figure 3 illustrates the departure from isotropy of the process as well as the accuracy of posterior inference for covariances under our fitting model. Here, each panel includes, for a specific site $s_i$, $\mathrm{Cov}(Y(s_i), Y(s_j) \mid \mathrm{data})$, $j \neq i$, along with the corresponding sample covariances. Results are presented for six representative sites $s_i$.

Of most interest for the isotropic process of case II is the bimodality of its univariate and bivariate densities. Our model was very successful in capturing this feature as revealed by

univariate and bivariate posterior predictive densities for several combinations of sites $s_i$ and the new sites $\tilde{s}_1$,...,$\tilde{s}_6$. We provide some illustrations in Figure 4.

Turning to comparison with the GP mixture model (as defined in Remark 1 of section 2.1), for case II, the superior predictive performance of the spatial DP mixture model for non-standard distributional shapes is evident from Figure 4. For case I, predictive inference under the nonparametric model was, again, more accurate although, in this case, differences from the parametric model were less pronounced. (Therefore, we have not overlaid these results on Figures 2 and 3 to avoid cluttering them.)

Lastly, regarding formal model comparison, we note that, to date, there is little work on comparison of a nonparametric model with another nonparametric model, with a semiparametric model or with a fully parametric model. Standard penalized likelihood methods are not applicable to the nonfinite dimensional likelihoods that arise under nonparametric modeling. Rather, it seems more sensible to work in predictive space. One possibility is to attempt to calculate Bayes factors, following the approach laid out in Basu and Chib (2003). Of course, one might prefer not to reduce a model to a single number in order to make comparison. Graphical comparison of predictive performance as in, e.g., Figure 4, may be more illuminating, particularly if only a few models are being considered.

### 4.2 Precipitation Data from a Region in Southern France

The French precipitation data we examine were discussed in Meiring et al. (1997) and have also been studied in Damian, Sampson and Guttorp (2001). There are, initially, 108 altitude-adjusted 10-day aggregated precipitation records from November and December of 1975 through 1992 for the 39 sites in Figure 1. Concerns regarding missing values and too many zeros for some of the records led us to reduce to 75 replicates which have been log-transformed (after adding 1 to all measurements) with site-specific means removed. (If we did not remove site-specific means we would have to include $\mu(s_i)$ in the models. For prediction, this would necessitate specifying a $\mu(s)$ process. Site-specific centering avoids this and, since primary interest is in the shape of predictive distributions and in the association in bivariate predictive distributions, seems appropriate.) Damian, Sampson and Guttorp (2001) use the same transformation but work with the full dataset. Their approach reveals relatively low spatial covariance in the central region and higher spatial covariance in the northeast region. The corresponding result in Meiring et al. (1997) is qualitatively similar, although in that paper the data were further standardized by dividing by site-specific standard deviations.

Preliminary exploration of the version of the dataset we consider also suggests that spatial

association is higher in the northeast than in the southwest. For instance, Figure 5 presents plots of sample covariances between each one of six sites (from the different subregions) and the other 38 sites vs. distance. The corresponding posterior mean covariances, included in the plots, indicate a good fit to the data as do the posterior predictive densities at the 39 observation sites. (See Figure 6 for results at 9 of the sites.) We note the ability of the model to capture different distributional shapes, including fairly symmetrical densities (e.g., sites $s_4$, $s_{27}$) and skewed densities (e.g., sites $s_{15}$, $s_{22}$).

Figure 5 immediately provides evidence of departure from isotropy (compare, e.g., sites 1 and 30). In an effort to examine nonstationarity, we selected 21 pairs of sites across the region, all pairs having the same separation vector and separated in a South-North direction. The pairs are shown in panel (a) of Figure 7. Panel (b) shows the associated posterior mean covariances suggesting departure from stationarity, encouraging a nonstationary spatial specification.

In the interest of validation for spatial prediction, we removed two sites from each of the three subregions in Figure 1, specifically, sites $s_4$, $s_{35}$, $s_{29}$, $s_{30}$, $s_{13}$, $s_{37}$, and refitted the model using only the data from the remaining 33 sites. Figures 8 and 9 provide results for new sites $(\tilde{s}_1, ..., \tilde{s}_6) = (s_4, s_{35}, s_{29}, s_{30}, s_{13}, s_{37})$. Posterior predictive densities are compared with the data at these sites from the full dataset, including all 39 sites, that were not used to fit the model in this validation exercise. Figure 10 shows, for each $i = 1, ..., 6$, plots of $\mathrm{Cov}(Y(\tilde{s}_i), Y(s_j) \mid \text{data})$, for all 33 sites $s_j$, vs. distance and corresponding plots with sample covariances based on the full dataset. Noting the instability of sample covariances for relatively small sample sizes, our model has done quite well.

Predictive inference can be extended to enable, up to interpolation, various spatial surfaces of interest. Given the site-specific centering, we would not expect prediction of a mean surface to reveal much. Spatial pattern in spread and skewness are potentially more interesting. Here, we use non-moment based measures, the interquartile range (.75 quantile − .25 quantile) for spread and the Bowley coefficient ([.75 quantile + .25 quantile − 2 median]/[.75 quantile − .25 quantile]) for skewness (see, e.g., Groeneveld and Meeden, 1984), which takes values in $(-1, 1)$ with negative (positive) values corresponding to left (right) skewness and 0 corresponding to symmetry. In Figure 11 we provide a predictive interquartile range surface, a predictive Bowley skewness coefficient surface, and a predictive median surface. We see evidence that spread and skewness vary spatially over the region with spread greatest in the northeastern part of the region and skewness generally to the left but least in the center. The median surface is essentially centered around 0 with slight depression in the center, slight elevation to the southwest and northeast.

## 5 Summary and extensions

The familiar spatial process modeling form (see, e.g., Cressie, 1993, p.112) sets $Y(s) = \mu(s) + \theta(s) + \epsilon(s)$ where $\mu(s)$ denotes the mean structure, $\theta(s)$ denotes a stationary spatial GP and $\epsilon(s)$ denotes a white noise process. Through DP mixing we have provided an approach for removing the restrictive assumptions on $\theta(s)$. That is, within the Bayesian framework, customarily, $\theta(s)$ would come from a random mean 0 GP, i.e., $\sigma^2$ and $\phi$ would be random. For us, $\theta(s)$ comes from a random process which is non-Gaussian and nonstationary with nonconstant variance which is "centered" around such a random GP. We have noted that the resulting model can be interpreted as a DDP and have shown that fitting such a model can be implemented by established methods for DP mixture models. We have illustrated with both simulated and real data examples.

It is evident that suitable finite mixture models can achieve the same objectives that we have for our modeling. However, such models are far more demanding to specify - how many mixture components, mean and covariance structure for each component, identifiability restrictions -and to fit. The DP mixing machinery only requires a base process specification with marginalization over the random process model enabling straightforward fitting.

Consistent with our modeling approach, if we sought a nonparametric mean strucuture in the covariates, we might DP mix on $\boldsymbol{\beta}$, as well, following the ideas of Mukhopadhyay and Gelfand (1997). An extension of interest would be to implement the generalized nugget structure discussed below (2). Nonhomogeneous and non-Gaussian $\epsilon(s)$ would add considerable flexibility to the model. In fact, if say the $Y(s)$ are binary (binomial) or counts, extension to a spatial generalized linear model would replace the nugget with a first stage exponential family specification. Here, $f\left(Y(s)|\beta, \theta(s)\right) \propto \exp(\eta(s)Y(s) + \chi(\eta(s)))$ where $\eta(s) = x'(s)\beta + \theta(s)$.

The $\omega_l$ could be allowed to follow a distribution other than that given by Sethuraman's rule, as in Hjort (2000) and Ishwaran and James (2001). This leads to more general spatial models. Such models can be fit with MCMC algorithms similar to those we present here, and may have advantageous asymptotic behavior, as described by Hjort (2000).

Allowing the $\omega_l$ to vary with location leads to more general spatial DPs in the spirit of MacEachern (1999). Now we allow the realization at location $s$ to come from a different surface than that for the realization at location $s'$. And it seems that the chance of this should depend upon the distance between $s$ and $s'$. More demanding modeling is required. Formalization of such generalized spatial DPs adds flexibility to the specification and exposes interesting modeling opportunities but introduces much more demanding computation. We will report on this work in a future manuscript.

Lastly, we can envision spatio-temporal extensions with several possible model formulations.

For instance, with time discretized, at time t, we could write $\theta_t(s)$ to allow dynamic evolution, e.g., $\theta_t(s) = \theta_{t-1}(s) + \eta_t(s)$ where $\eta_t(s)$ are independent replications from a spatial DP.

# References


Agarwal, D.K., and Gelfand, A.E. (2002), "Slice Gibbs Sampling for Simulation Based Fitting of Spatial Data Models," Technical Report, Department of Statistics, University of Connecticut.

Antoniak, C.E. (1974), "Mixtures of Dirichlet Processes With Applications to Nonparametric Problems," *The Annals of Statistics*, 2, 1152-1174.

Barry, R.P., and Ver Hoef, J.M. (1996), "Blackbox Kriging: Spatial Prediction without Specifying Variogram Models," *Journal of Agricultural, Biological, and Environmental Statistics*, 1, 297-322.

Basu, S., and Chib, S. (2003), "Marginal Likelihood and Bayes Factors for Dirichlet Process Mixture Models," *Journal of the American Statistical Association*, 98, 224-235.

Blackwell, D., and MacQueen, J.B. (1973), "Ferguson Distributions via Pólya Urn Schemes," *The Annals of Statistics*, 1, 353-355.

Bush, C.A., and MacEachern, S.N. (1996), "A Semiparametric Bayesian Model for Randomised Block Designs," *Biometrika*, 83, 275-285.

Cressie, N.A.C. (1993), *Statistics for Spatial Data* (Revised Edition), New York: Wiley.

Damian, D., Sampson, P.D., and Guttorp, P. (2001), "Bayesian Estimation of Semi-parametric Non-stationary Spatial Covariance Structures," *Environmetrics*, 12, 161-178.

De Iorio, M., Müller, P., Rosner, G.L., and MacEachern, S.N. (2004), "An ANOVA Model for Dependent Random Measures," *Journal of the American Statistical Association*, 99, 205-215.

Diggle, P.J., Tawn, J.A., and Moyeed, R.A. (1998), "Model-based Geostatistics (with discussion)," *Applied Statistics*, 47, 299-350.

Ecker, M.D., and Gelfand, A.E. (2003), "Spatial Modeling and Prediction under Stationary Non-geometric Range Anisotropy," *Environmental and Ecological Statistics*, 10, 165-178.

Escobar, M.D. (1994), "Estimating Normal Means With a Dirichlet Process Prior," *Journal of the American Statistical Association*, 89, 268-277.

Escobar, M.D., and West, M. (1995), "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association*, 90, 577-588.

Escobar, M.D., and West, M. (1998), "Computing Nonparametric Hierarchical Models," in *Practical Nonparametric and Semiparametric Bayesian Statistics*, eds. D. Dey, P. Müller and D. Sinha, New York: Springer, pp. 1-22.

Ferguson, T.S. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, 1, 209-230.

Ferguson, T.S. (1974), "Prior Distributions on Spaces of Probability Measures," *The Annals of Statistics*, 2, 615-629.

Fuentes, M., and Smith, R.L. (2001), "A New Class of Nonstationary Spatial Models," Technical Report, Department of Statistics, North Carolina State University.

Groeneveld, R.A., and Meeden, G. (1984), "Measuring Skewness and Kurtosis," *The Statistician*, 33, 391-399.

Higdon, D. (2002), "Space and Space-time Modeling Using Process Convolutions," in Quantitative Methods for Current Environmental Issues, eds. C.W. Anderson, V. Barnett, P.C. Chatwin and A.H. El-Shaarawi, London: Springer Verlag.

Higdon, D., Swall, J., and Kern, J. (1999), "Non-Stationary Spatial Modeling," *Bayesian Statistics 6*, eds. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, Oxford University Press.

Hjort, N.L. (2000), "Bayesian Analysis for a Generalised Dirichlet Process Prior," Statistical Research Report, Department of Mathematics, University of Oslo.

Ishwaran, H., and James, L.F. (2001), "Gibbs Sampling Methods for Stick-Breaking Priors," *Journal of the American Statistical Association*, 96, 161-173.

Kent, J.T. (1989), "Continuity Properties for Random Fields," *The Annals of Probability*, 17, 1432-1440.

MacEachern, S.N. (1998), "Computational Methods for Mixture of Dirichlet Process Models," in *Practical Nonparametric and Semiparametric Bayesian Statistics*, eds. D. Dey, P. Müller and D. Sinha, New York: Springer, pp. 23-43.

MacEachern, S.N. (1999), "Dependent Nonparametric Processes," in *ASA Proceedings of the Section on Bayesian Statistical Science*, Alexandria, VA: American Statistical Association, pp. 50-55.

MacEachern, S.N. (2000), "Dependent Dirichlet Processes," Technical Report, Department of Statistics, The Ohio State University.

Meiring, W., Monestiez, P., Sampson, P.D., and Guttorp, P. (1997), "Developments in the Modelling of Nonstationary Spatial Covariance Structure from Space-Time Monitoring Data," in *Geostatistics Wollongong '96*, vol. 1, eds. E.Y. Baafi and N. Schofield, Dordrecht: Kluwer Academic, pp. 162-173.

Mukhopadhyay, S., and Gelfand, A.E. (1997), "Dirichlet Process Mixed Generalized Linear Models," *Journal of the American Statistical Association*, 92, 633-639.

Müller, P., Quintana, F., and Rosner, G. (1999), "Hierarchical Meta-Analysis over Related Non-parametric Bayesian Models," ISDS Discussion Paper 99-22, Duke University.

Sampson, P.D., 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 Structure via Spatial Deformations," *Journal of the Royal Statistical Society, Series B*, 65, 743-758.

Sethuraman, J. (1994), "A constructive definition of Dirichlet priors," *Statistica Sinica,* 4, 639-650.

Shapiro, A., and Botha, J. (1991), "Variogram Fitting with a General Class of Conditionally Nonnegative Definite Functions," *Computational Statistics and Data Analysis*, 11, 87-96.

Stein, M.L. (1999), *Interpolation of Spatial Data: Some Theory for Kriging*, New York: Springer Verlag.

West, M., Müller, P., and Escobar, M.D. (1994), "Hierarchical Priors and Mixture Models, With Application in Regression and Density Estimation," in *Aspects of Uncertainty: A Tribute to D.V. Lindley*, eds. A.F.M. Smith and P. Freeman, New York: Wiley, pp. 363-386.

# Appendix

Here, we provide the details for MCMC posterior simulation for model (5). As discussed in section 2.2, we can employ Gibbs sampling to draw from $[\boldsymbol{\theta}, \boldsymbol{\beta}, \tau^2, \nu, \sigma^2, \phi \mid \text{data}]$. The required full conditionals are:

(a) $[(\theta_t, w_t) \mid \{(\theta_{t'}, w_{t'}), t' \neq t\}, \boldsymbol{\beta}, \tau^2, \nu, \sigma^2, \phi, \text{data}]$, for $t = 1,...,T$

(b) $[\theta_j^* \mid \mathbf{w}, T^*, \boldsymbol{\beta}, \tau^2, \sigma^2, \phi, \text{data}]$, for $j = 1,...,T^*$

(c) $[\boldsymbol{\beta} \mid \boldsymbol{\theta}, \tau^2, \text{data}]$ and $[\tau^2 \mid \boldsymbol{\theta}, \boldsymbol{\beta}, \text{data}]$

(d) $[\nu \mid T^*, \text{data}]$, $[\sigma^2 \mid \boldsymbol{\theta}^*, T^*, \phi]$ and $[\phi \mid \boldsymbol{\theta}^*, T^*, \sigma^2]$

The first set of full conditionals in (a) can be obtained using the Pólya urn representation of the DP (Blackwell and MacQueen, 1973, Ferguson, 1973). We use the subscript "$(-t)$" to denote all relevant quantities when $\theta_t$ is removed from the vector $\boldsymbol{\theta}$. Hence $T^*_{(-t)}$ refers to the number of clusters in $(\theta_{t'}, t' \neq t)$ and $T_{j(-t)}$ to the number of elements in cluster $j$, $j = 1,...,T^*_{(-t)}$, with $\theta_t$ removed. Then for each $t = 1,...,T$ the full conditional in (a) is given by

$$\frac{q_0 h(\theta_t \mid \mathbf{Y}_t, \boldsymbol{\beta}, \tau^2, \sigma^2, \phi) + \sum_{j=1}^{T^*_{(-t)}} T_{j(-t)} q_j \delta_{\theta_j^*}(\theta_t)}{q_0 + \sum_{j=1}^{T^*_{(-t)}} T_{j(-t)} q_j},$$

where $q_0 = \nu \int f_{N_n}(\mathbf{Y}_t \mid X_t'\boldsymbol{\beta} + \theta, \tau^2 I_n) G_0^{(n)}(d\theta \mid \sigma^2, \phi)$, $q_j = f_{N_n}(\mathbf{Y}_t \mid X_t'\boldsymbol{\beta} + \theta_j^*, \tau^2 I_n)$, $j = 1,...,T^*_{(-t)}$ and $h(\theta_t \mid \mathbf{Y}_t, \boldsymbol{\beta}, \tau^2, \sigma^2, \phi) = \nu q_0^{-1} f_{N_n}(\mathbf{Y}_t \mid X_t'\boldsymbol{\beta} + \theta_t, \tau^2 I_n) g_0^{(n)}(\theta_t \mid \sigma^2, \phi)$. Here $g_0^{(n)}(\cdot \mid \sigma^2, \phi)$ is the density corresponding to $G_0^{(n)}(\cdot \mid \sigma^2, \phi)$. Using standard calculations with multivariate normal distributions we obtain

$$q_0 = \frac{\nu (\det(\Lambda))^{1/2} \exp\left\{-0.5\tau^{-2}((\mathbf{Y}_t - X_t'\boldsymbol{\beta})'(I_n - \tau^{-2}\Lambda)(\mathbf{Y}_t - X_t'\boldsymbol{\beta}))\right\}}{(2\pi\tau^2\sigma^2)^{n/2}(\det(H_n(\phi)))^{1/2}},$$

where $\Lambda = (\tau^{-2} I_n + \sigma^{-2} H_n^{-1}(\phi))^{-1}$. Moreover, $h(\theta_t \mid \mathbf{Y}_t, \boldsymbol{\beta}, \tau^2, \sigma^2, \phi)$ is the density of a multivariate normal distribution with mean vector $\tau^{-2}(\Lambda(\mathbf{Y}_t - X_t'\boldsymbol{\beta}))$ and covariance matrix $\Lambda$. Note that once $\theta_t$ is updated $w_t$ is also implicitly updated. In fact, before proceeding to the updating of $\theta_{t+1}$, it is necessary to redefine $T^*$, $\theta_j^*$, $j = 1,...,T^*$, $w_t$, $t = 1,...,T$ and $T_j$, $j = 1,...,T^*$ which also define $T^*_{(-t)}$ and $T_{j(-t)}$, $j = 1,...,T^*_{(-t)}$, after the removal of $\theta_{t+1}$.

Once step (a) is completed we have a specific configuration $\mathbf{w} = (w_1, ..., w_T)$ and the associated $T^*$ cluster values. Step (b) amounts to moving these cluster locations in order to improve mixing of the Gibbs sampler (Bush and MacEachern, 1996). For each $j = 1,...,T^*$,

$$[\theta_j^* \mid \mathbf{w}, T^*, \boldsymbol{\beta}, \tau^2, \sigma^2, \phi, \text{data}] \propto g_0^{(n)}(\theta_j^* \mid \sigma^2, \phi) \prod_{\{t: w_t = j\}} f_{N_n}(\mathbf{Y}_t \mid X_t'\boldsymbol{\beta} + \theta_j^*, \tau^2 I_n).$$

Again, it is straightforward to show that this full conditional is a multivariate normal with covariance matrix $\Lambda_j = (T_j\tau^{-2}I_n + \sigma^{-2}H_n^{-1}(\phi))^{-1}$ and mean vector $\tau^{-2}\Lambda_j \sum_{\{t:w_t=j\}} (\mathbf{Y}_t - X_t'\boldsymbol{\beta})$.

The full conditionals in (c) do not involve the DP part of the model and hence have exactly the same form as in the corresponding parametric hierarchical model. Specifically, $[\boldsymbol{\beta} \mid \boldsymbol{\theta}, \tau^2, \text{data}]$ is a $N_p(\boldsymbol{\beta} \mid \tilde{\boldsymbol{\beta}}_0, \tilde{\Sigma}_{\boldsymbol{\beta}})$ distribution with $\tilde{\Sigma}_{\boldsymbol{\beta}} = (\Sigma_{\boldsymbol{\beta}}^{-1} + \tau^{-2}\sum_{t=1}^{T} X_t X_t')^{-1}$ and $\tilde{\boldsymbol{\beta}}_0 = \tilde{\Sigma}_{\boldsymbol{\beta}}(\Sigma_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta}_0 + \tau^{-2}\sum_{t=1}^{T} X_t(\mathbf{Y}_t - \theta_t))$, and $[\tau^2 \mid \boldsymbol{\theta}, \boldsymbol{\beta}, \text{data}]$ is an IGamma$(\tau^2 \mid \tilde{a}_\tau, \tilde{b}_\tau)$ distribution with $\tilde{a}_\tau = a_\tau + 0.5nT$ and $\tilde{b}_\tau = b_\tau + 0.5\sum_{t=1}^{T}(\mathbf{Y}_t - X_t'\boldsymbol{\beta} - \theta_t)'(\mathbf{Y}_t - X_t'\boldsymbol{\beta} - \theta_t)$.

Turning to step (d) of the Gibbs sampler, the precision parameter $\nu$ can be updated using the augmentation idea from Escobar and West (1995). Briefly, we introduce an auxiliary variable $\eta$ and, having completed steps (a) to (c) of the $b$th Gibbs iteration, we draw $\eta^{(b)}$ from $[\eta \mid \nu^{(b-1)}, \text{data}]$ and then $\nu^{(b)}$ from $[\nu \mid \eta^{(b)}, T^{*(b)}, \text{data}]$, where $[\eta \mid \nu, \text{data}] = \text{Beta}(\nu+1, T)$ and $[\nu \mid \eta, T^*, \text{data}]$ is the two component mixture of gamma distributions $p$ Gamma$(a_\nu + T^*, b_\nu - \log(\eta))$ + $(1-p)$ Gamma$(a_\nu + T^* - 1, b_\nu - \log(\eta))$, where $p = (a_\nu + T^* - 1)/\{T(b_\nu - \log(\eta)) + a_\nu + T^* - 1\}$. Alternatively, it is possible to remove the updating for $\nu$ from the Gibbs sampler by integrating over this parameter and modifying appropriately the full conditionals in step (a) (MacEachern, 1998).

The full conditionals for $\sigma^2$ and $\phi$ are obtained by combining their priors with the likelihood arising from $g_0^{(n)}(\theta_j^* \mid \sigma^2, \phi)$, $j = 1,...,T^*$, since, marginalizing over $G^{(n)}$, the $\theta_j^*$ are i.i.d. $G_0^{(n)}(\cdot \mid \sigma^2, \phi)$. In particular, the full conditional for $\sigma^2$ is proportional to $\sigma^{-2(a_\sigma+1)} \exp(-b_\sigma/\sigma^2) \prod_{j=1}^{T^*} \sigma^{-n} \exp(-\theta_j^{*\prime} H_n^{-1}(\phi)\theta_j^*/(2\sigma^2))$, a form which yields an IGamma$(\sigma^2 \mid \tilde{a}_\sigma, \tilde{b}_\sigma)$, with $\tilde{a}_\sigma = a_\sigma + 0.5nT^*$ and $\tilde{b}_\sigma = b_\sigma + 0.5\sum_{j=1}^{T^*} \theta_j^{*\prime} H_n^{-1}(\phi)\theta_j^*$. Finally, the full conditional for $\phi$ is proportional to

$$[\phi](\det(H_n(\phi)))^{-T^*/2} \exp(-\sum_{j=1}^{T^*} \theta_j^{*\prime} H_n^{-1}(\phi)\theta_j^*/(2\sigma^2)), \tag{15}$$

where $(H_n(\phi))_{i,j} = \exp(-\phi\|s_i - s_j\|)$, $\phi > 0$. The fact that $\phi$ appears in all the off-diagonal elements of $H_n(\phi)$ indicates that, regardless of the prior for $\phi$, (15) does not allow for straightforward sampling. We have found that discretizing the full conditional for $\phi$ and sampling it directly yields a method that is both efficient and numerically stable. Hence we work with a grid of values $\phi_\ell$, $\ell = 1,...,L$, in $(0, b_\phi]$ and a discrete uniform prior $[\phi_\ell] = P(\phi = \phi_\ell) = L^{-1}$. (How to choose $b_\phi$ is discussed in section 2.3.)

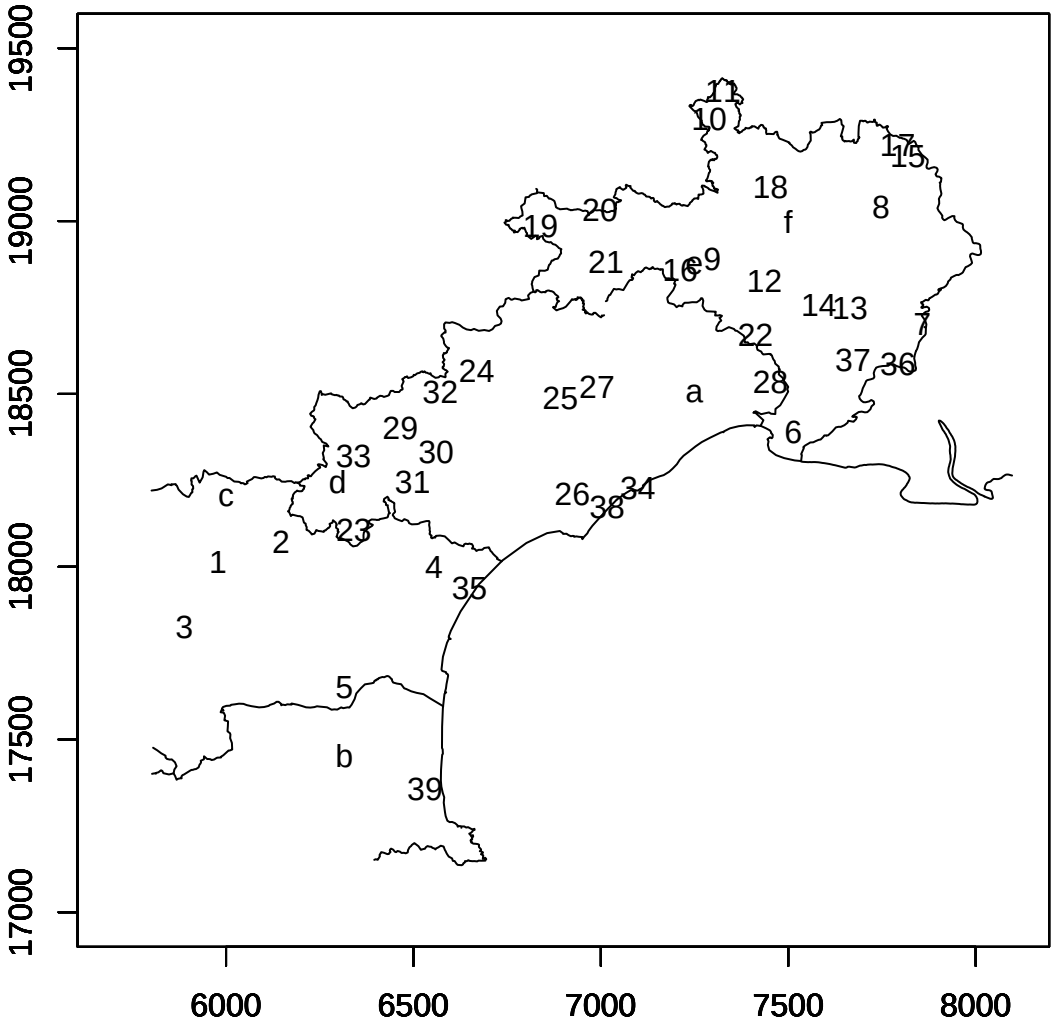



Figure 1: Geographic map of the Languedoc-Roussillon region in southern France including the locations of the 39 sites where the precipitation data have been observed. The six new sites considered for spatial prediction in the simulation experiment are denoted by a, b, c, d, e and f. The boundaries of three French departments are also drawn.

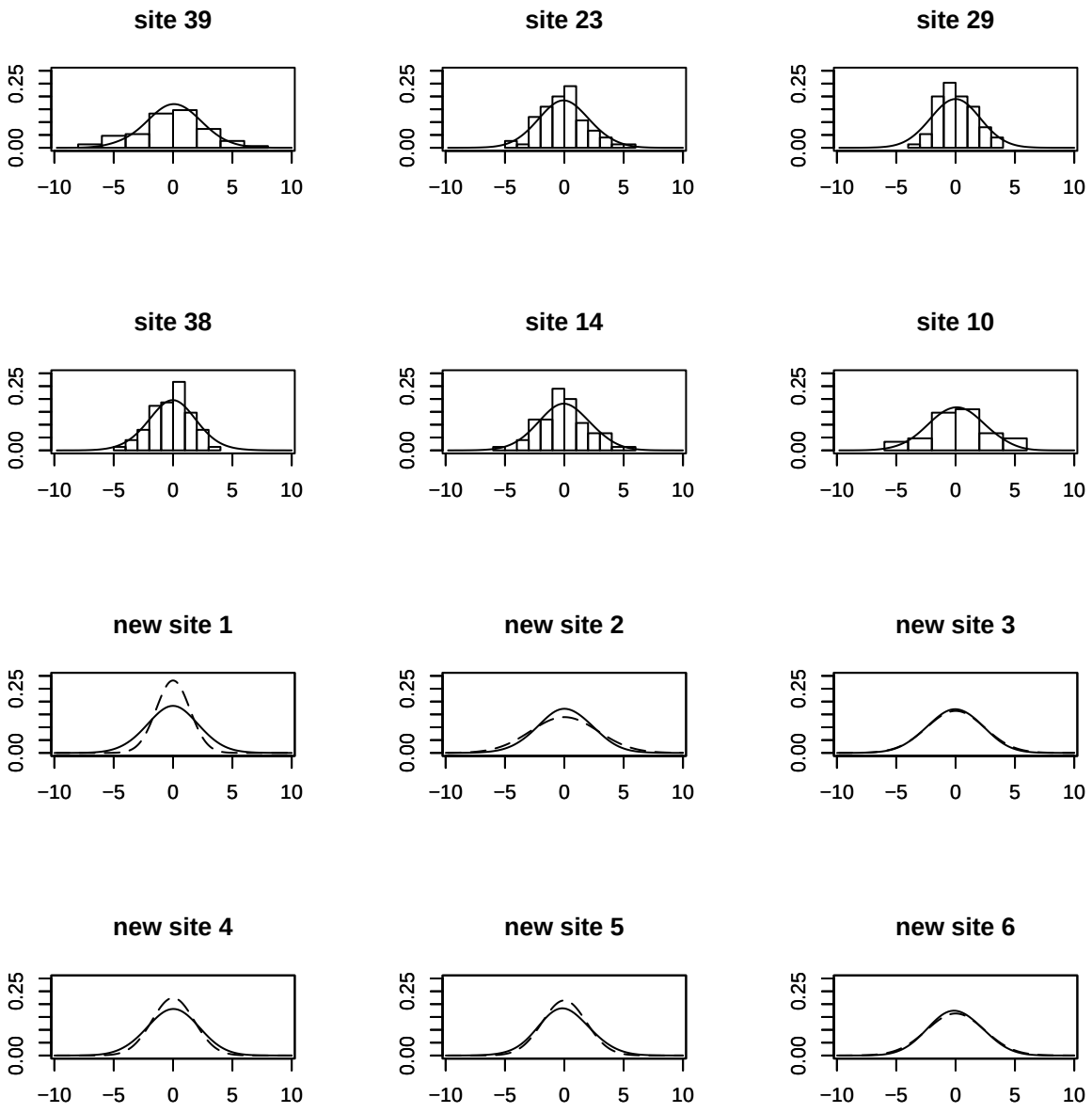



Figure 2: Simulated data, case I. Posterior predictive densities (solid lines) at sites $s_{39}$, $s_{23}$, $s_{29}$, $s_{38}$, $s_{14}$, $s_{10}$ and new sites $\tilde{s}_1$,...,$\tilde{s}_6$ (given in section 4.1). The former are overlaid on histograms of corresponding data and the latter on corresponding true densities (dashed lines).

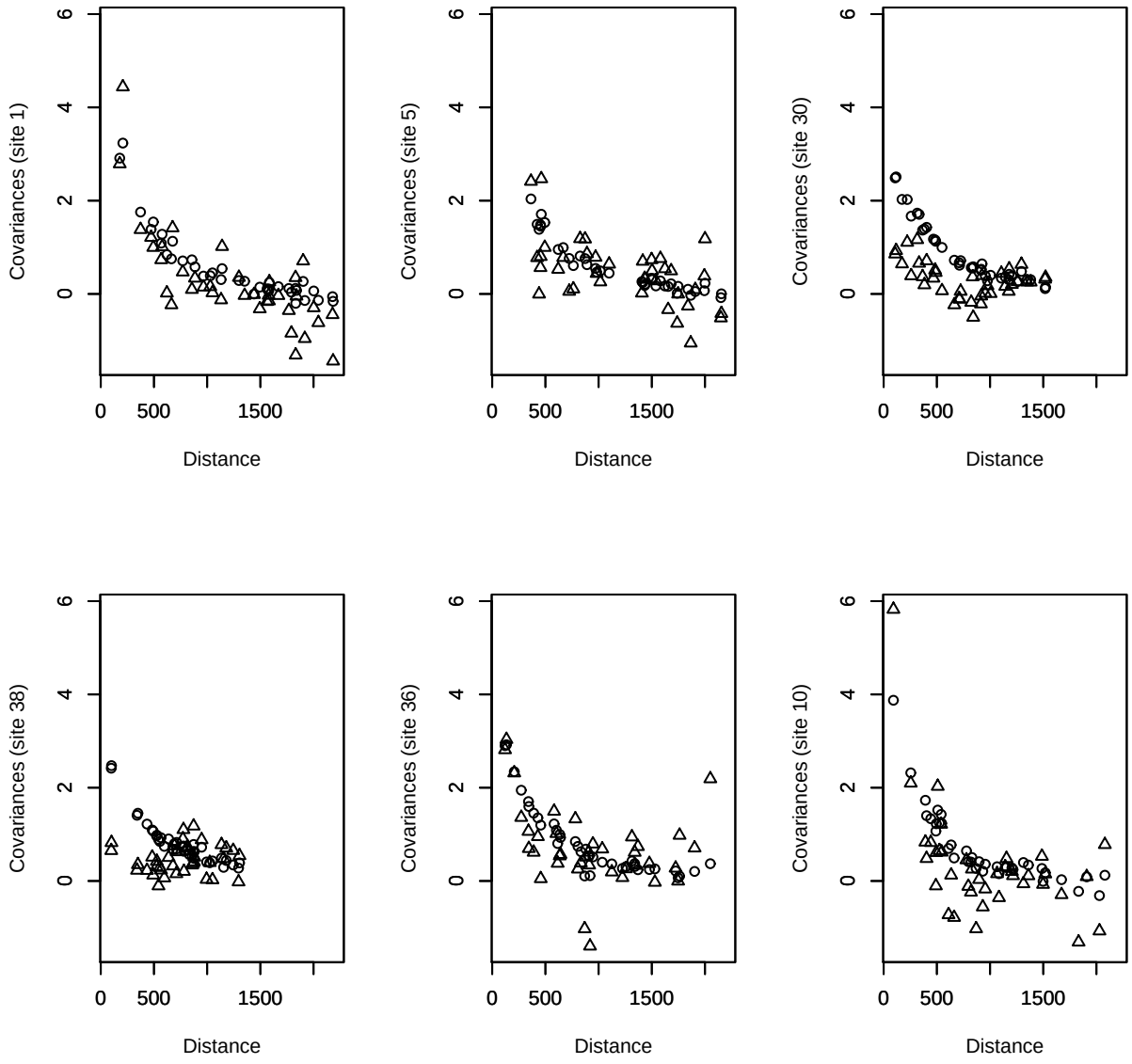



Figure 3: Simulated data, case I. Posterior mean covariances (circles) between each one of 6 sites ($s_1$, $s_5$, $s_{30}$, $s_{38}$, $s_{36}$ and $s_{10}$) and the remaining 38 sites plotted against distance. The triangles denote the corresponding sample covariances.

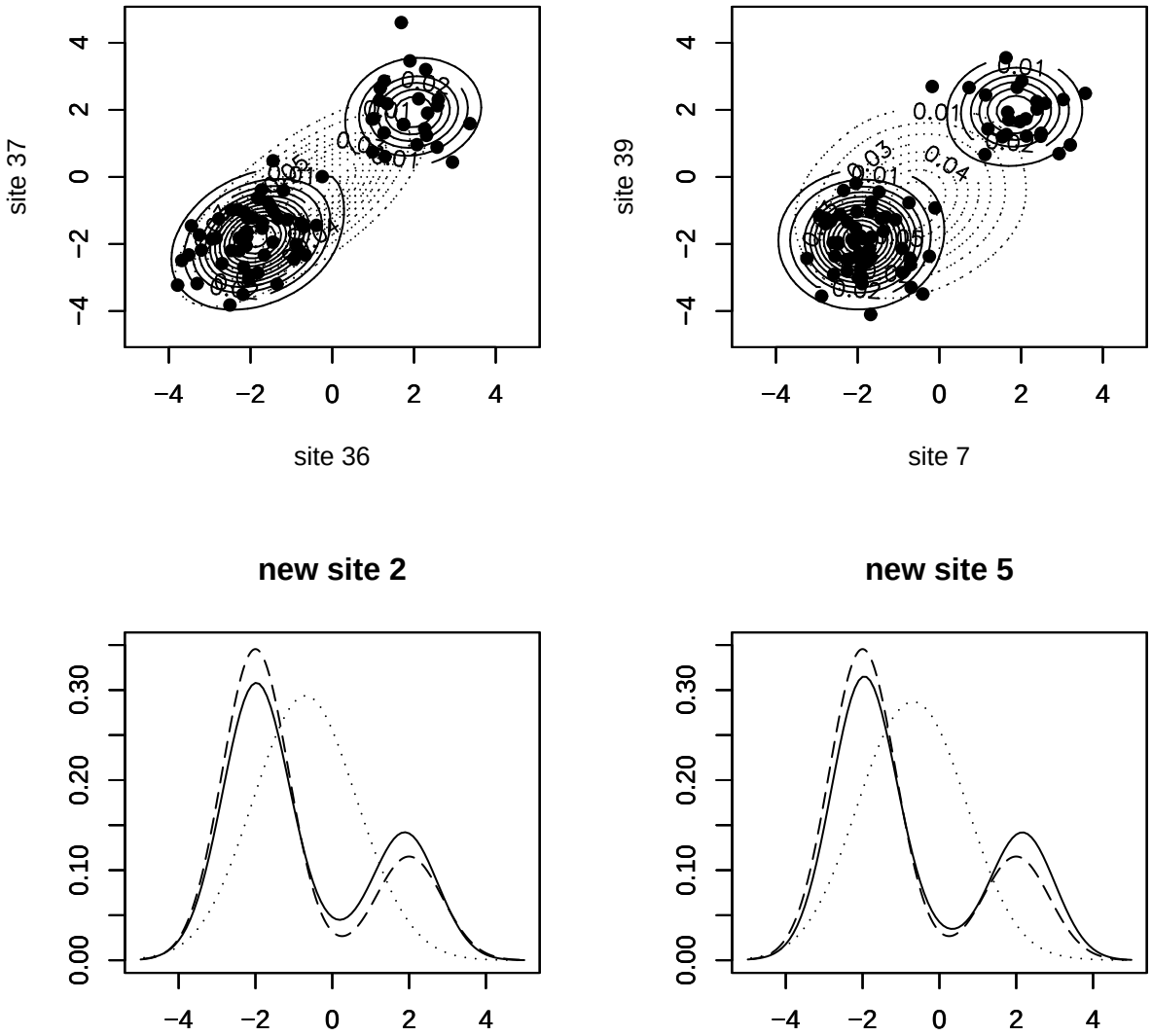



Figure 4: Simulated data, case II. The top row panels provide bivariate posterior predictive densities for pairs of sites $(s_{36}, s_{37})$ and $(s_7, s_{39})$ overlaid on corresponding plots of data. The bottom row panels include posterior predictive densities at new sites $\tilde{s}_2 = (6316, 17452)$ and $\tilde{s}_5 = (7250, 18870)$ and the associated true densities (dashed lines). The solid lines correspond to the spatial DP mixture model and the dotted lines to the parametric GP mixture model.

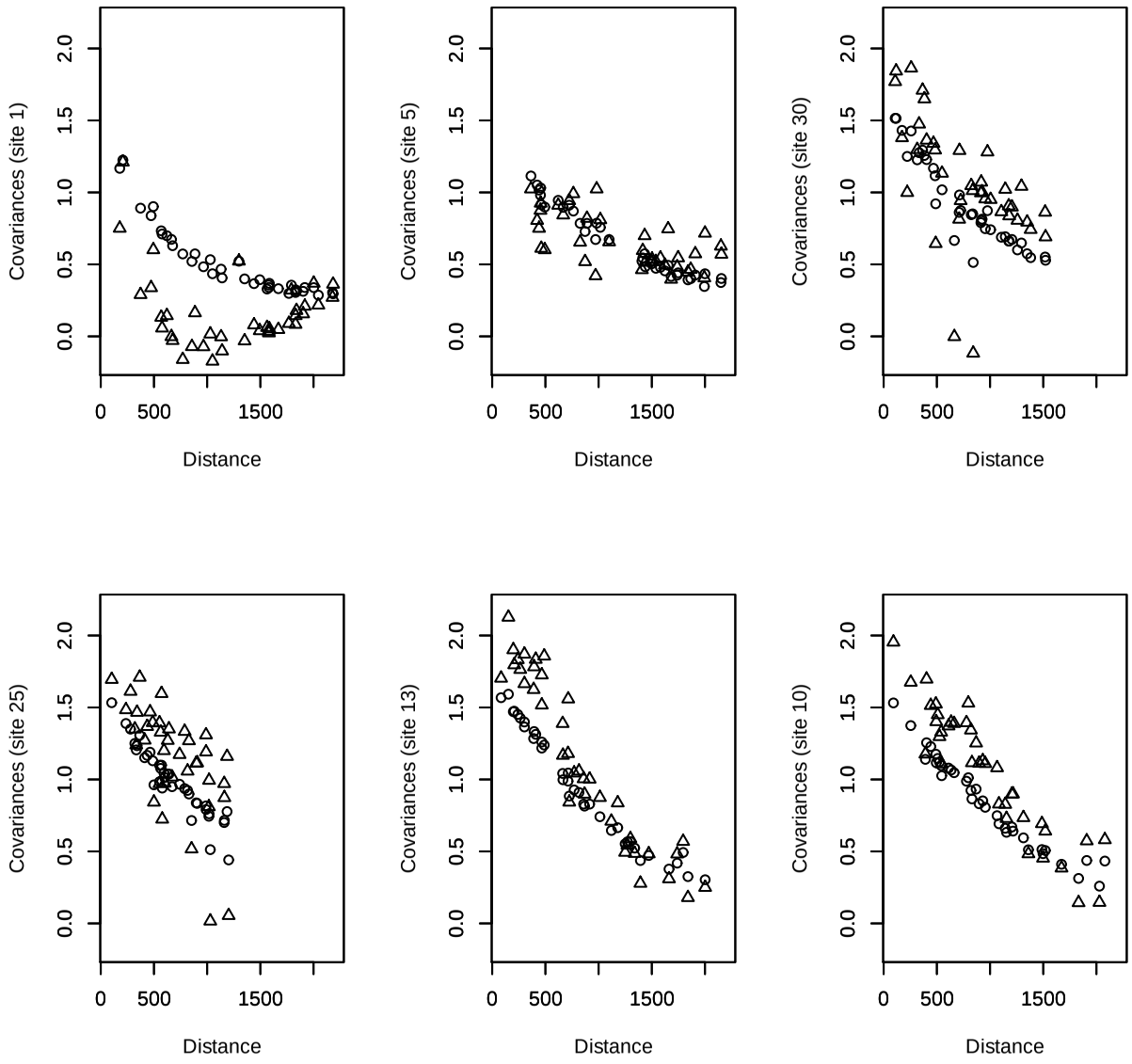



Figure 5: French precipitation data. Posterior mean covariances (circles) between each one of 6 sites ($s_1$, $s_5$, $s_{30}$, $s_{25}$, $s_{13}$ and $s_{10}$) and the remaining 38 sites plotted against distance. The triangles denote the corresponding sample covariances.

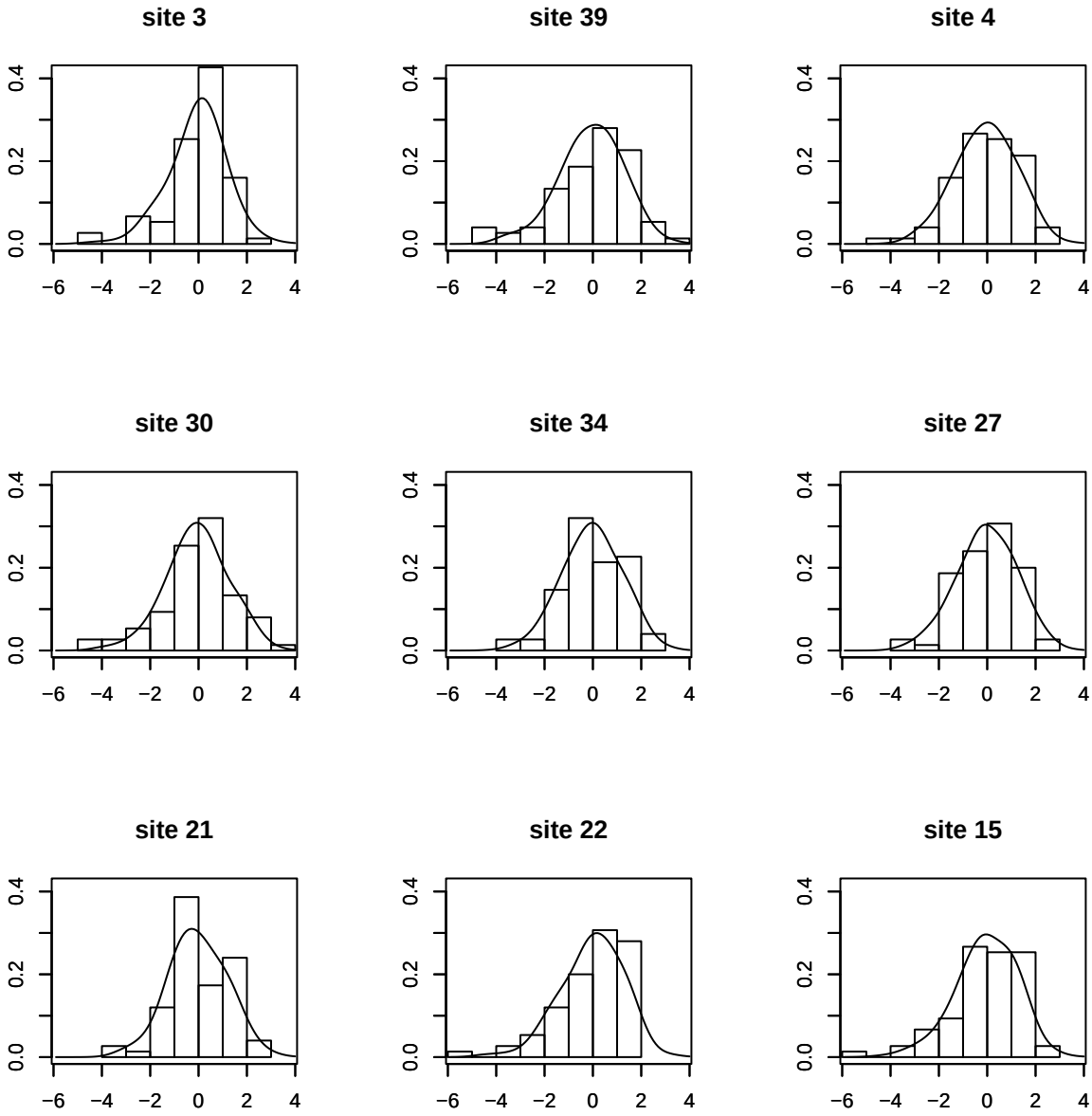
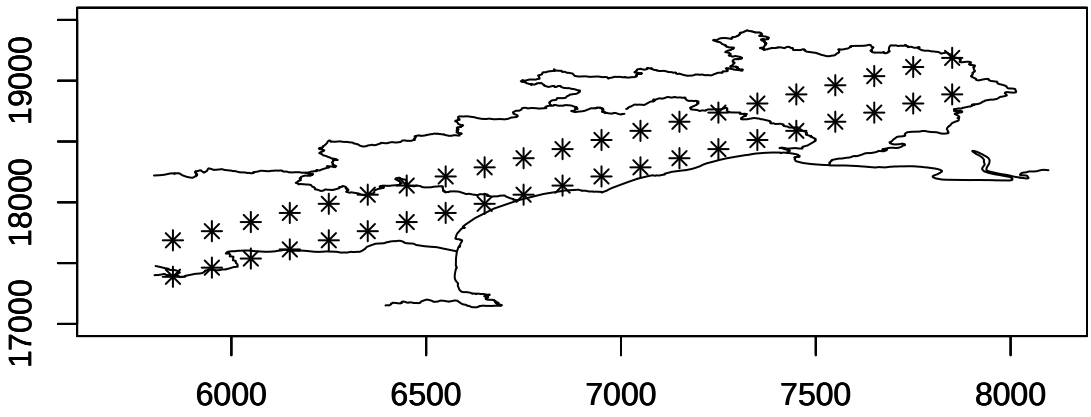
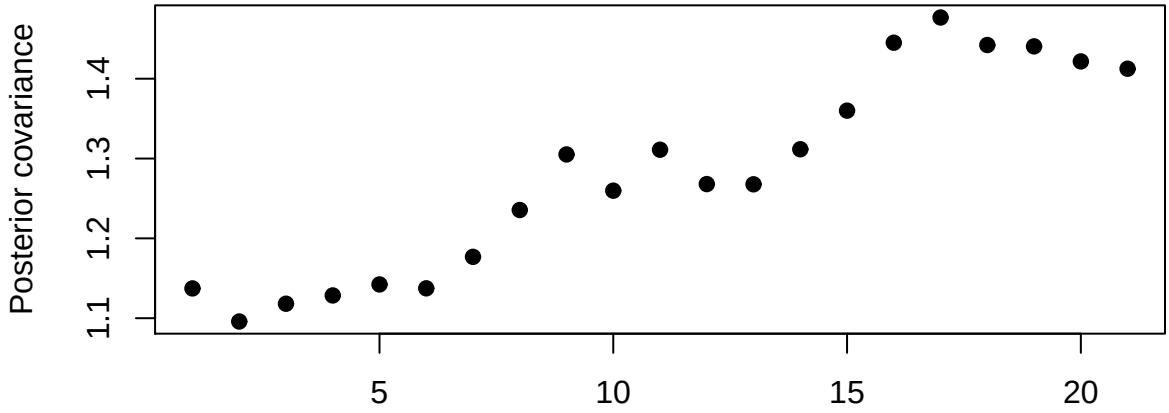



Figure 6: French precipitation data. Posterior predictive densities at 9 sites, $s_3$, $s_{39}$, $s_4$, $s_{30}$, $s_{34}$, $s_{27}$, $s_{21}$, $s_{22}$ and $s_{15}$, (among the ones where observations are recorded) overlaid on histograms of data observed at the corresponding sites.




(a)




(b)

Figure 7: French precipitation data. Stationarity investigation: 7(a) shows 21 pairs of sites, all pairs having the same separation vector. 7(b) provides associated posterior mean covariances.

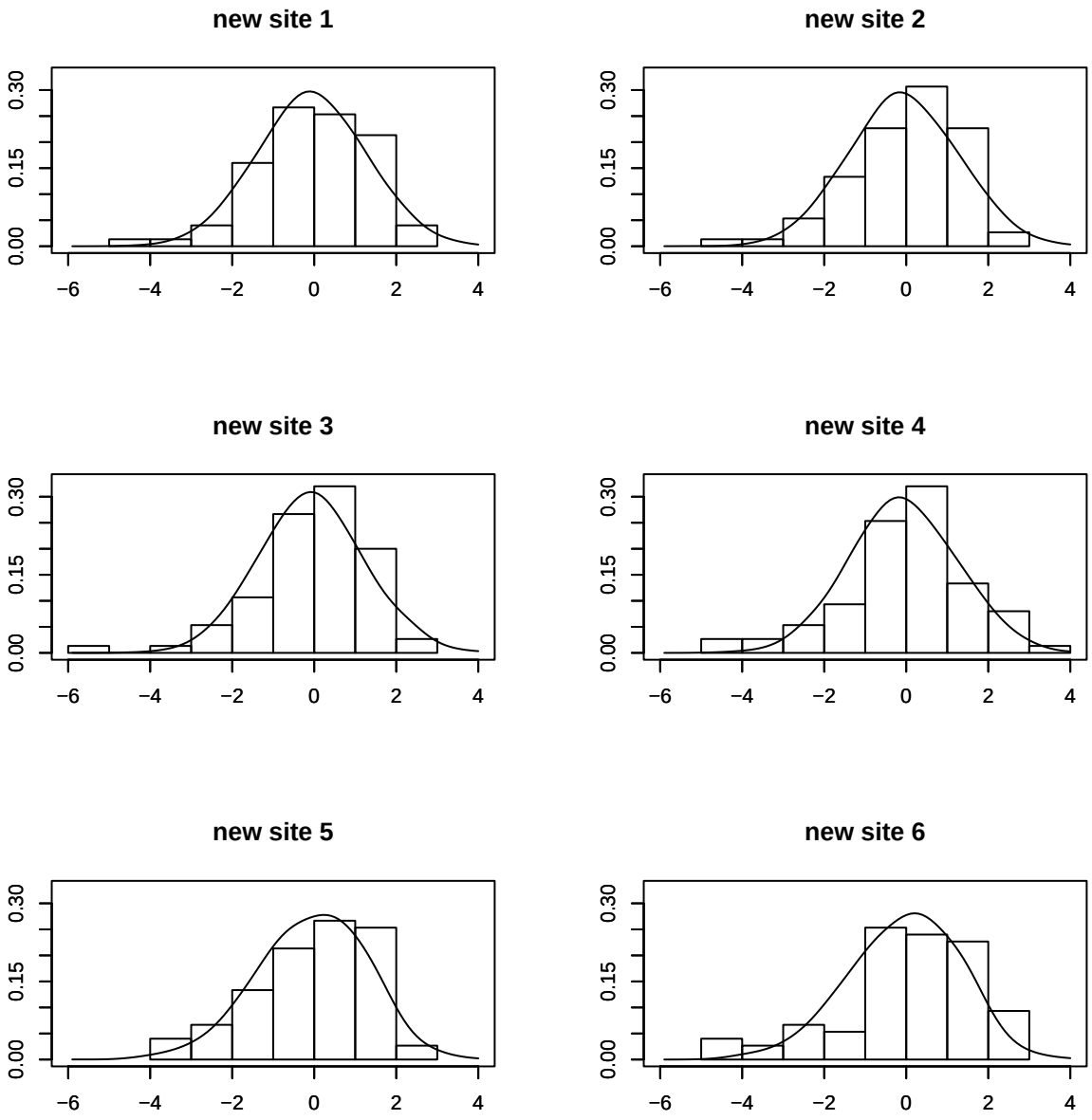



Figure 8: French precipitation data. Posterior predictive densities at new sites $(\tilde{s}_1, ..., \tilde{s}_6) = (s_4, s_{35}, s_{29}, s_{30}, s_{13}, s_{37})$, based on model fitted to data from 33 sites (after removing sites $s_4$, $s_{35}$, $s_{29}$, $s_{30}$, $s_{13}$ and $s_{37}$). The histograms are based on the data observed at the corresponding sites in the full dataset (that includes all 39 sites).

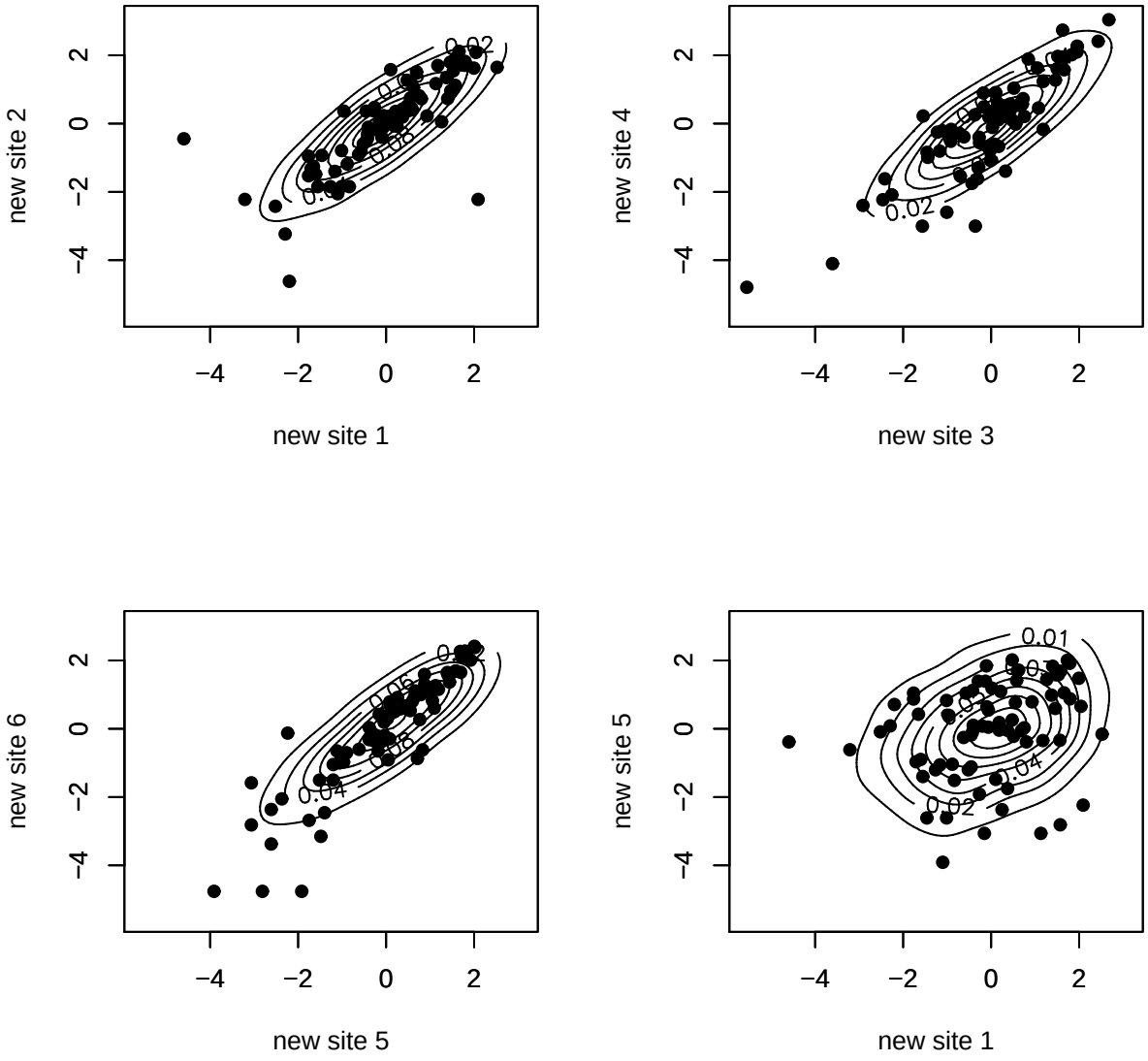



Figure 9: French precipitation data. Bivariate posterior predictive densities for pairs of new sites $(\tilde{s}_1, \tilde{s}_2) = (s_4, s_{35})$, $(\tilde{s}_3, \tilde{s}_4) = (s_{29}, s_{30})$, $(\tilde{s}_5, \tilde{s}_6) = (s_{13}, s_{37})$ and $(\tilde{s}_1, \tilde{s}_5) = (s_4, s_{13})$, based on model fitted to data from 33 sites (after removing sites $s_4$, $s_{35}$, $s_{29}$, $s_{30}$, $s_{13}$ and $s_{37}$). Plots of data observed at the corresponding pairs of sites in the full dataset are also included.

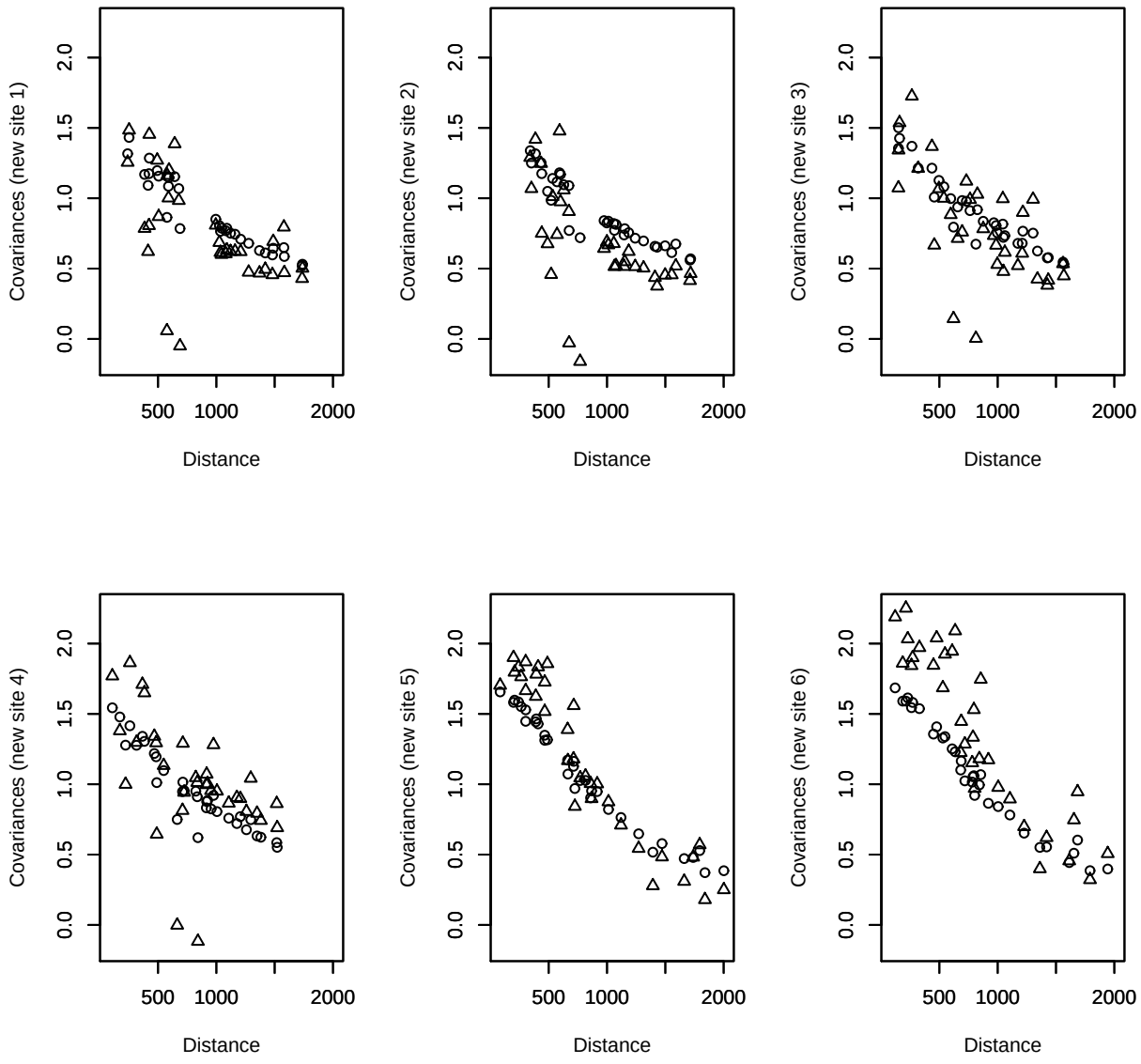



Figure 10: French precipitation data. Posterior mean covariances (circles) between each one of the new sites $(\tilde{s}_1, ..., \tilde{s}_6) = (s_4, s_{35}, s_{29}, s_{30}, s_{13}, s_{37})$ and the 33 sites used to fit the model (after removing sites $s_4$, $s_{35}$, $s_{29}$, $s_{30}$, $s_{13}$ and $s_{37}$) plotted against distance. The corresponding sample covariances (triangles) based on the full dataset are also included.

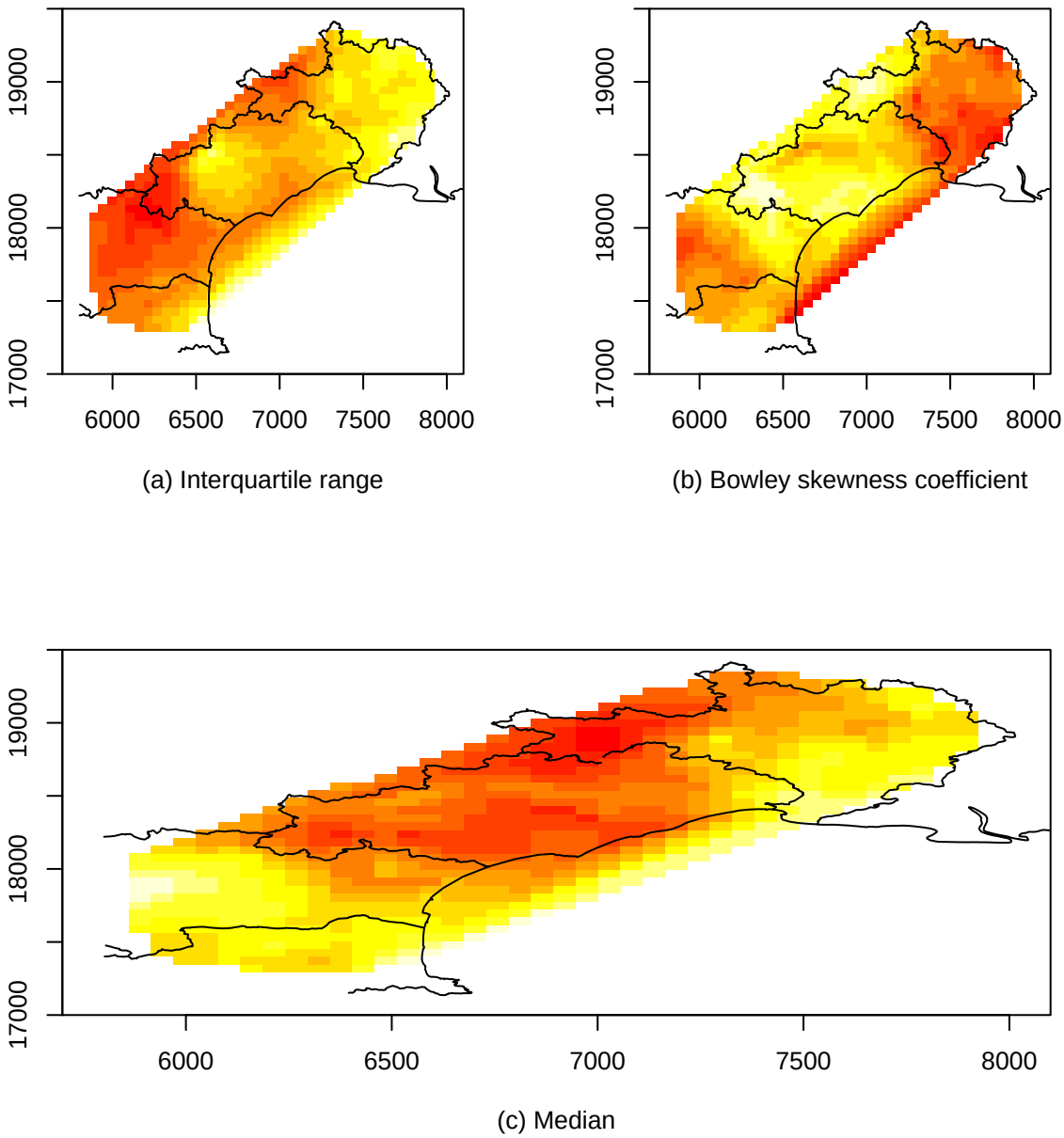



Figure 11: French precipitation data. Posterior predictive interquartile range, Bowley skewness coefficient, and median surfaces. The ranges of values are $(4.73, 5.64)$ for panel (a), $(-0.15, 0.01)$ for panel (b), and $(-0.11, 0.05)$ for panel (c), with darker colors corresponding to smaller values.