Nonstationary Spatial Process Models with Spatially Varying Covariance Kernels

Sébastien Coube-Sisquelle\textsuperscript{a,1} *, Sudipto Banerjee\textsuperscript{b,2}, and Benoît Liquet\textsuperscript{a,c,3}

\textsuperscript{a} Laboratoire de Mathématiques et de leurs Applications, Université de Pau et des Pays de l’Adour, E2S-UPPA, Pau, France
\textsuperscript{b} Department of Biostatistics, University of California, Los Angeles, United States of America
\textsuperscript{c} School of Mathematical and Physical Sciences, Macquarie University, Sydney, Australia

\textsuperscript{1} sebastien.coube@univ.pau.fr \textsuperscript{2} sudipto@ucla.edu \textsuperscript{3} benoit.liquet-weiland@mq.edu.au

Abstract

Spatial process models for capturing nonstationary behavior in scientific data present several challenges with regard to statistical inference and uncertainty quantification. While nonstationary spatially-varying kernels are attractive for their flexibility and richness (see, e.g., Paciorek, 2003), their practical implementation has been reported to be overwhelmingly cumbersome because of the high-dimensional parameter spaces resulting from the spatially varying process parameters. Matters are considerably exacerbated with the massive numbers of spatial locations over which measurements are available. With limited theoretical tractability offered by nonstationary spatial processes, overcoming such computational bottlenecks require a synergy between model construction and algorithm development. We build a class of scalable nonstationary spatial process models using spatially varying covariance kernels. We present some novel consequences of such representations that benefit computationally efficient implementation. More specifically, we operate within a coherent Bayesian modeling framework to achieve full uncertainty quantification using a Hybrid Monte-Carlo with nested interweaving. We carry out experiments on synthetic data sets to explore model selection and parameter identifiability and assess inferential improvements accrued from the nonstationary modeling. We illustrate strengths and pitfalls with a data set on remote sensed normalized difference vegetation index with further analysis of a lead contamination data set in the Supplement.

Keywords: Bayesian hierarchical models; Hybrid Monte-Carlo; Interweaving; Nearest-Neighbor Gaussian processes; Nonstationary spatial modeling.

*The authors gratefully acknowledge E2S: Energy and Environment Solutions
1 Introduction

Bayesian hierarchical models for analysing spatially and temporally oriented data are widely employed in scientific and technological applications in the physical, environmental and health sciences (Cressie and Wikle, 2015; Banerjee et al., 2014; Gelfand et al., 2019). Such models are constructed by embedding a spatial process within a hierarchical structure,

\[
[\text{data} \mid \text{process, parameters}] \times [\text{process} \mid \text{parameters}] \times [\text{parameters}], \tag{1}
\]

which specifies the joint probability law of the data, an underlying spatial process and the parameters. The process in (1) is a crucial inferential component that introduces spatial and/or temporal dependence, allows us to infer about the underlying data generating mechanism and carry out predictions over entire spatial-temporal domains. Point-referenced spatial data, which will be our focus here, refer to measurements over a set of locations with fixed coordinates. These measurements are assumed to arise as a partial realization of a spatial process over the finite set of locations. Stationary Gaussian processes are conspicuous in spatial models. Stationarity imposes simplifying assumptions on the dependence structure of the process, which are unlikely to hold in most scientific applications.

Nonstationary spatial models attempt to relax such assumptions and deliver wide-ranging benefits to inference. For example, when variability in the data is a complex function of space composed of multiple locally varying processes, the customary stationary covariance kernels may be inadequate. Here, richer and more informative covariance structures in nonstationary processes, while adding complexity, may be more desirable by improving smoothing, goodness of fit and predictive inference. Nonstationary spatial models have been addressed by a number of authors (Higdon, 1998; Fuentes, 2002; Paciorek, 2003; Banerjee et al., 2008; Cressie and Johannesson, 2008; Lemos et al., 2009; Guhaniyogi et al., 2013; Yang and Bradley, 2021; Risser and Calder, 2015; Risser, 2016;
The richness sought in nonstationary models have been exemplified in a number of the above references. Paciorek (2003) and Kleiber and Nychka (2012) introduce nonstationarity by allowing the parameters of the Matern class to vary with location, yielding local variances, local ranges and local geometric anisotropies. While some of these approaches are related to ideas in modeling covariance matrices using covariates (see, e.g., Pourahmadi, 1999) and have been extended and further developed in a number of different directions including regularized inference, they have not been devised for implementation on massive spatial data sets in the order of $10^5+$. While recent developments in nonstationary models have addressed data sets in the order of hundreds (Risser and Calder, 2015; Ingebrigtsen et al., 2015; Heinonen et al., 2016) or thousands (Guhaniyogi et al., 2013; Fuglstad et al., 2015a) of locations, these are modest with respect to the size of commonly encountered spatial data (see the examples in Datta et al., 2016; Heaton et al., 2019; Katzfuss and Guinness, 2021). Bakka et al. (2019) proposed a novel “barrier” nonstationary model that achieves computational effectiveness by using simultaneous autoregression represented through a stochastic partial differential equation. Nonstationary models also suffer from unidentifiable parameters arising from complex space-varying covariance kernels, which yield weakly identifiable models that are difficult to estimate and interpret. This also complicates model evaluation and selection as inference is very sensitive to the specifications of the model.

We devise a new class of nonstationary spatial models for massive data sets that build upon Bayesian hierarchical models based on directed acyclic graphs (DAGs) such as the Nearest Neighbor Gaussian Process (NNGP) (Datta et al., 2016) and, more generally, Vecchia processes (Katzfuss and Guinness, 2021), which exploit their attractive computational and inferential properties (Finley et al., 2019). We focus upon spatially-varying
Matérn covariance kernels (e.g., Paciorek, 2003; Kleiber and Nychka, 2012) by endowing the space-varying parameters with a low-dimensional Gaussian process. This achieves sparser representations than Paciorek (2003) or Risser and Calder (2015) and is a natural extension of the usual logarithmic prior for positive parameters such as the marginal variance, the noise variance, and the range when it is not elliptic. We embed these well-defined stochastic processes within a coherent hierarchical modeling framework as in Heinonen et al. (2016), but tailor our method towards modeling large data sets and capturing local anisotropy.

A key challenge is learning about the nonstationary covariance processes. This cannot be addressed solely using theoretical results (asymptotic theory for nonstationary processes are hardly accessible) or solely relying upon sophisticated algorithms (achieving convergence in high-dimensional parameter spaces will be unfeasible). Therefore, our approach relies upon some novel consequences of sparser representations in conjunction with a Hamiltonian Monte Carlo (HMC) algorithm adapted from Heinonen et al. (2016) to overcome computational impediments and analyze spatial data at massive scales. More specifically, we hybridize the approach of Heinonen et al. (2016) with the interweaving strategies of Yu and Meng (2011); Filippone et al. (2013). Our proposed method is able to achieve fully probabilistic inference on a latent nonstationary process which captures residual spatial behavior after accounting for spatial predictors. Our framework encompasses several simpler models as special cases. While theoretical properties of model selection based upon nonstationary features are analytically inaccessible, we conduct an array of computer experiments on synthetic data to demonstrate that our devised sparsity-inducing process models do not over-fit data generated from a simpler model. Instead, they degenerate towards a state corresponding to a simpler model.

The balance of the article proceeds as follows. Section 2 outlines the covariance and
data models, and the properties of a nonstationary NNGP density, which are incorporated into a Bayesian hierarchical model presented in Section 3. Section 4 details the MCMC implementation of the model based upon a Gibbs sampler using interweaving in Section 4.2 and upon a Hamiltonian Monte Carlo algorithm in Section 4.3. Section 5 focuses on applications: experiments on synthetic data to test the properties of the model. We use the model to analyse a data set of lead contamination Hengl (2009) in the United States. Section 6 summarizes and points to future research. An elaborate accompanying Supplement offers key derivations, further details on algorithms and numerical results from simulation experiments.

2 Spatially-Varying Covariance Kernel Models

Let $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ be a set of $n$ spatial locations indexed in a spatial domain $\mathcal{D}$, where $\mathcal{D} \subset \mathbb{R}^d$ with $d \in \{1, 2, 3\}$. For any $s \in \mathcal{D}$ we envision a spatial regression model

$$z(s) = x(s)^T \beta + w(s) + \epsilon(s),$$

(2)

where $z(s)$ represents a measurement at $s$, $x(s)$ is a $p \times 1$ vector of predictors, $\beta$ is the corresponding $p \times 1$ vector of slopes, $w(s)$ is a latent spatial process and $\epsilon(s)$ is white noise attributed to random disturbances. In full generality, the noise will be modeled as heteroskedastic so that $\epsilon(s) \overset{\text{ind}}{\sim} \mathcal{N}(0, \tau^2(s))$ while $w(s)$ is customarily modelled using a Gaussian process over $\mathcal{D}$. Therefore,

$$w(\mathcal{S}) := (w(s_1), w(s_2), \ldots, w(s_n))^T \overset{\text{ind}}{\sim} \mathcal{N}(0, \Sigma(\mathcal{S})),$$

(3)

where the elements of the $n \times n$ covariance matrix $\Sigma(\mathcal{S})$ are determined from a spatial covariance function $K(s, s')$ defined for any pair of locations $s$ and $s'$ in $\mathcal{D}$. In full generality the covariance function can accommodate spatially varying parameters to obtain
nonstationarity, which is our current focus. Therefore, the \((i, j)\)-th element of \(\Sigma(S)\) is

\[
\Sigma(s_i, s_j) = K(s_i, s_j) = \sigma(s_i)\sigma(s_j)K_0(s_i, s_j; \alpha(s_i), \alpha(s_j)),
\]

where \(\sigma(s_1 \ldots s_n) := \{\sigma(s_i) : i = 1, \ldots, n\}\) is a collection of (positive) spatially varying marginal standard deviations, \(K_0(s, s'; \{\alpha(s), \alpha(s')\})\) is a valid spatial correlation function defined for any pair of locations \(s\) and \(s'\) in \(D\) with two spatial range parameters \(\alpha(s)\) and \(\alpha(s')\) that vary with the locations. For two sets of locations \(a \in S\) and \(b \in S\), \(\Sigma(a, b)\) denotes the submatrix comprising the rows and columns of \(\Sigma\) whose indices correspond to \(a\) and \(b\) in \(S\), respectively. We also abbreviate \(\Sigma(a, a)\) into \(\Sigma(a)\). These parameters can be either positive-definite matrices offering a locally anisotropic nonstationary covariance structure or positive real numbers specifying a locally isotropic nonstationary range. For example, Paciorek (2003) proposed a valid class of nonstationary covariance functions

\[
K_0(s, s'; A(s), A(s')) = \frac{2^{d/2}|A(s)|^{1/4}|A(s')|^{1/4}}{|A(s) + A(s')|^{1/2}}K_i \left( d_M \left( s, s', (A(s) + A(s'))/2 \right) \right),
\]

where \(A(s)\) and \(A(s')\) are anisotropic spatially-varying range matrices, \(d\) is the dimension of the space-time domain, \(d_M(\cdot, \cdot, \cdot)\) is the Mahalanobis distance and \(K_i\) is an isotropic correlation function. If \(A(\cdot)\) does not vary by location, the covariance structure is anisotropic but stationary. A nonstationary correlation function is obtained by setting \(A(s) = \alpha(s)I_d\),

\[
K_0(s, s'; \alpha(s), \alpha(s')) = \left( \frac{\sqrt{2}\alpha(s)^{1/4}\alpha(s')^{1/4}}{(\alpha(s) + \alpha(s'))^{1/2}} \right)^d K_i \left( d_E(s, s')/ ((\alpha(s) + \alpha(s'))/2) \right),
\]

where \(d_E(\cdot, \cdot)\) is the Euclidean distance (Mahalanobis distance with matrix \(I_d\)).

Spatial process parameters in isotropic covariance functions are not consistently estimable under fixed-domain asymptotic paradigms (Zhang, 2004). Therefore, irrespective of sample size, no function of the data can converge in probability to the value of the parameter from an oracle model. Irrespective of how many locations we sample, the effect of the prior on these parameters will not be eliminated in Bayesian inference. This
can be addressed using penalized complexity priors to reduce the ridge of the equivalent range-marginal variance combinations to one of its points (Fuglstad et al., 2015b). The covariance function sharply drops to 0 so the observations that inform about the covariance parameters at a location tend to cluster around the site. Nonstationary models are significantly more complex. The parameters specifying the spatial covariance are functions over \( \mathcal{D} \). These form uncountable collections and, hence, inference requires modeling them as spatial processes. This considerably exacerbates the challenges surrounding identifiability and inference for these completely unobserved processes. Asymptotic inference is precluded due to the lack of regularity conditions. Bayesian inference, while offering fully model-based solutions for completely unobserved processes, will also need to obviate the computational hurdles arising from weakly identified processes, which result in poorly behaved MCMC algorithms, and inference extensible to massive data sets.

The customary NNGP (Datta et al., 2016) specifies a valid Gaussian process in two steps. We begin with a “parent” process \( GP(0, K(s, s')) \) so that \( w(\mathcal{S}) \) has the probability law in (3). Let \( f(w(\mathcal{S}) \mid \theta) \) be the corresponding joint density, where \( \theta \) denotes parameters in \( K(\cdot, \cdot) \). First, we build a sparse approximation of this joint density. Using a fixed topological ordering of the points in \( \mathcal{S} \) we construct a nested sequence \( \mathcal{S}_{i-1} \subset \mathcal{S}_i \), where \( \mathcal{S}_i = \{s_1, s_2, \ldots, s_i-1\} \) for \( i = 2, 3, \ldots, n \). The joint density of the NNGP is given by \( \tilde{f}(w(\mathcal{S}) \mid \theta) = f(w(s_1) \mid \theta) \prod_{i=2}^{n} \tilde{f}(w(s_i) \mid w(\mathcal{S}_{i-1}), \theta) \) (also referred to as Vecchia’s approximation Vecchia, 1988; Stein et al., 2004), where

\[
\tilde{f}(w(s_i) \mid w(\mathcal{S}_{i-1}), \theta) = f(w(s_i) \mid w(pa(s_i)), \theta), \tag{7}
\]

and \( pa(s_i) \) comprises the parents of \( s_i \) from a DAG over \( \mathcal{S} \). The parents in the DAG are often chosen as the nearest neighbours of \( s_i \) among \( s_1, \ldots, s_{i-1} \). This approximation defines a Gaussian Process whose covariance matrix can be expressed as \( (\tilde{R}^T \tilde{R})^{-1} \). Each row of \( \tilde{R} \)
is computed straightforwardly from the conditional distribution \( f(w(s_i) \mid w(pa(s_i)), \theta) \). It follows that \( \tilde{R} \) is lower triangular and very sparse with its \( i \)-th row having nonzero entries only at the indices corresponding to \( pa(s_i) \) (Pourahmadi, 1999, 2007; Datta et al., 2016; Katzfuss and Guinness, 2021). Prediction at unobserved locations is achieved by appending them sequentially as new nodes with edges connecting the new locations to its nearest neighbors from the preceding nodes already in the DAG. This results in the predictive distribution \( \tilde{Q}_{n, i} = \tilde{f}(w(\tilde{s}_i) \mid w(pa(\tilde{s}_i)), \theta) \), where \( \tilde{s}_i \) are a topologically sorted sequence of locations where predictions are sought (see, e.g., Datta et al., 2016; Finley et al., 2019; Katzfuss et al., 2020, for scalable spatial interpolation and prediction).

We obtain \( \tilde{f}(w(S) \mid \theta(S)) = f(w(s_1) \mid \theta(s_1)) \prod_{i=2}^{n} \tilde{f}(w(s_i) \mid w(S_{i-1}), \theta(S_i)) \) by applying (7) to spatially varying parameters, where \( \theta(s) \) is a vector comprising the covariance function parameters at \( s \) and \( \theta(S) = (\theta(s_1)^T, \theta(s_2)^T, \ldots, \theta(s_n)^T)^T \) is a \( (\sum_{i=1}^{n} |\theta(s_i)|) \times 1 \) vector.

We build scalable nonstationary processes by exploiting the following key property (see Section S1.1 of the Supplement for the derivation),

\[
\tilde{f}(w(s_i) \mid w(S_{i-1}), \theta(S)) = f(w(s_i) \mid w(pa(s_i)), \theta(s_i \cup pa(s_i))) ,
\]

where the NNGP density is derived using covariance kernels in (5) and (6), both of which accommodate spatially-variable parameters \( \theta(S) \). Equation (8) reveals a crucial property that the the NNGP will reduce the dimension of conditional sets from \( w(S_{i-1}) \) to \( w(pa(s_i)) \) and from \( \theta(S) \) to \( \theta(s_i \cup pa(s_i)) \).

Another useful property relates the NNGP derived from the covariance function in (4) and its corresponding correlation function. Let \( \tilde{R}_0 \) be the NNGP factor obtained from the precision matrix using the correlation function \( K_0(\cdot) \) instead of the covariance function \( K(\cdot) \) and let \( \sigma(S) = (\sigma(s_1), \sigma(s_2), \ldots, \sigma(s_n))^T \). Then, \( \tilde{R} = \tilde{R}_0 \text{diag}(\sigma(S))^{-1} \), where \( \text{diag}(\sigma(S)) \) is the diagonal matrix with elements \( \sigma(s_i) \); see the Section S1.2 of the Supplement for the
derivation. The logarithm \( N(w(S) | 0; \Sigma(S; \theta)) \) is computed up to a constant as

\[
- \sum_{i=1}^{n} \log \left( \frac{(\tilde{R}_0)_{i,i}}{\sigma(s_i)} \right) - \frac{1}{2} w^T \text{diag}(\sigma(S))^{-1} \tilde{R}_0 \tilde{R}_0 \text{diag}(\sigma(S))^{-1} w,
\]

(9)

exploiting cheap expressions for the determinant of \( \tilde{R} \). Furthermore, the topological ordering heuristics explored by Guinness (2018) hold for NNGP specifications in (5) and (6).

3 Hierarchical space-varying covariance models

We extend (2) to accommodate replicated measurements at each location. If \( z(s, j) \) is the \( j \)-th measurement at location \( s \), where \( j = 1, 2, \ldots, n_s \), and \( z(s) \) is the \( n_s \times 1 \) vector of all measurements at location \( s \), then (2) is modified to \( z(s) = X(s)^T \beta + 1_{n_s} w(s) + \epsilon(s) \),

where \( X(s)^T \) is \( n_s \times p \) with the values of predictors or design variables corresponding to each location \( s \); \( 1_{n_s} \) is the \( n_s \times 1 \) vector of ones; \( \epsilon(s) \) is an \( n_s \times 1 \) vector with \( j \)-th element \( \epsilon(s) \sim N(0, \text{diag}(\tau^2(s))) \), where \( \tau^2(s) = (\tau^2_1(s), \ldots, \tau^2_{n_s}(s))^T \) is \( n_s \times 1 \); and \( \beta \) and \( w(s) \) are exactly as in (2). Note that \( X(s)^T \) can include predictors that do not vary within \( s \), e.g., elevation, and that can vary within the spatial location, e.g., multiple technicians can record measurements at \( s \) and the technician’s indicator may be used as a covariate.

We build a hierarchical space-varying covariance model over a set of spatial locations \( S = \{s_1, \ldots, s_n\} \) so the total number of measurements available for \( z(s) \) is \( |z| = \sum_{i=1}^{n} n_{s_i} \).

We propose the following hierarchical framework,

\[
(a) \underbrace{[z \mid \beta, w(S), \tau^2(S)]}_{(a)} \times (b) \underbrace{[w(S) \mid \theta(S)]}_{(b)} \times (c) \underbrace{[\theta(S) \mid \beta_\theta, \gamma_\theta]}_{(c) \text{ and } (d)} \times (e) \underbrace{[\tau^2(S) \mid \gamma_\tau]}_{(e) \text{ and } (f)} \times (f) \underbrace{[\gamma_\theta, \gamma_\tau, \beta_\theta, \beta_\tau]}_{(g)},
\]

(10)
where \(\cdot\) denotes probability distributions. These symbolic blocks are modeled as below:

\[
\begin{align*}
(a) \quad z & \sim N(X\beta + Mw(S), \text{diag}(\tau^2(S))) ; \\
(b) \quad w(S) & \sim N(0, \tilde{\Sigma}(S; \theta(S))) ; \\
(c) \quad \log(\theta(S)) & = X_\theta(S)\beta_\theta + W_\theta(S) ; \\
(d) \quad W_\theta(S) & \sim N(0, \zeta_{\gamma_\theta}) ; \\
(e) \quad \log(\tau^2(S)) & = X_\tau(S)\beta_\tau + MW_\tau(S) ; \\
(f) \quad W_\tau(S) & \sim N(0, \zeta_{\gamma_\tau}) ; \\
(g) \quad \{\gamma_\theta, \gamma_\tau, \beta_\theta, \beta_\tau, \beta\} & \sim p(\cdot, \cdot, \cdot, \cdot, \cdot),
\end{align*}
\]

(11)

where in (11(a)) \(z\) denotes \(|z| \times 1\) vector of all measurements, \(X\) is \(|z| \times p\) obtained by stacking up \(X(s_i)\) over the \(n\) locations in \(S\), \(\beta\) is the corresponding vector of regression coefficients, and \(\text{diag}(\tau^2(S))\) is a diagonal matrix with elements of the \(|z| \times 1\) vector \(\tau^2(S) = \langle \tau^2(s_1), \tau^2(s_2), \ldots, \tau^2(s_n) \rangle\) along the diagonal. The matrix \(M\) is a \(|z| \times n\) matching matrix with elements \(M_{i,j} = 1\) if the \(i\)-th element of \(z\) corresponds to the \(j\)-th spatial location and 0 otherwise. We preclude an observation from being obtained in two spatial locations at the same time. Hence, each row of \(M\) has exactly one term equal to one. Also, since there is at least one observation in each location, each column of \(M\) has at least one term equal to one. This yields (2) with \(|z| = n\) and \(M\) as a permutation matrix.

In (11(b)), the distribution of \(w(S)\) is specified using (9) as described in Section 2. Equations (11(c)) and (11(e)) specify models for the space-varying parameters \(\theta(S)\) and \(\tau^2(S)\), respectively, with (11(d)) and (11(f)) providing the distributional specifications.

The dimension of \(\tau^2(S)\) for fully heteroskedastic specifications is \(|z| \times 1\) while \(\theta(S) = \langle \theta(s_1), \ldots, \theta(s_n) \rangle\) is a vector whose dimension will depend upon the dimension of each \(\theta(s_i)\) in the model specification. The quantities \(\zeta_{\gamma_\theta}\) and \(\zeta_{\gamma_\beta}\) denote covariance matrices depending on parameters \(\gamma_\theta\) and \(\gamma_\beta\) according to the particular model we pursue. Furthermore, each \(\theta(s_i)\) is a \(d \times d\) matrix in anisotropic range models, whence (11(c)) is a matrix-variate linear model. We will revisit this adaptation in greater detail below. Both \(X\) and \(X_\tau\) have \(|z|\) rows that vary with measurements, while \(X_\theta(S)\) has \(n\) rows correspond-
ing to the spatial sites. Specifying $X_{\theta}$ is necessary to ensure that $w(s)$ is well-identified, while $X_{\tau}$ accommodates modeling errors within a spatial site, e.g. to account for variation among measurements from different technicians within a single spatial location. Finally, (11(g)) completes the hierarchical specification with distributions on the hitherto unspecified parameters introduced earlier in (11).

The models in (11(c) and (e)) require some further discussion. Let $|\theta(s)|$ be the dimension of the space-varying covariance kernel parameters at location $s$. Therefore, $|\theta(S)|$ is $O(n)$ and so is $|W_\theta(S)|$, which results in an apparently insurmountable estimation problem when $n >> 10^4$. One approach is to assign an NNGP prior for $W_\theta(S)$, as was done for $w(S)$ in Section 2. This, as has been discussed in Section 2, approximates the parent Gaussian process while offering massive computational benefits and can deliver estimates of the random field for $\theta$ that are almost indistinguishable from the parent. Another option is a low rank process for each element in $W_\theta(s)$. Writing $W_\theta(s) = (W_{\theta,1}(s), W_{\theta,2}(s), \ldots, W_{\theta,|\theta(s)|}(s))^T$ as a $|\theta(s)| \times 1$ vector, we model its $j$-th element $W_{\theta,j}(s) = \sum_{k=1}^{p} B_{\theta,j,k}(s) u_k$ for $j = 1, 2, \ldots, |\theta(s)|$, where $B_{\theta,j,k}(s)$ is a basis functions and $u = (u_1, u_2, \ldots, u_p)^T$ is a $p \times 1$ vector of basis coefficients. Therefore, $W_\theta(s) = B_\theta(s)u$, where $B_\theta(s)$ is $|\theta(s) \times p|$ with elements $(B_{\theta,j,k}(s))$ and $u \sim N(0, \gamma_\theta I_p)$ is $p \times 1$. In particular, the basis functions can be constructed as in Banerjee et al. (2008) using the conditional expectation of a full-rank Gaussian Process $E[W_\theta(s) \mid W_\theta(S^*)]$, where $S^*$ is a smaller set of $p$ “knots” in the domain (see Section S6 of the Supplement). We provide a few remarks on the selection of knots. The knots can be found as the centers of a K-means clustering, thereby allowing the user to specify only the desired number of knots. Our package vignette offers a quick look at how the PP samples behave using our spatial plotting functions. We point out that Guhaniyogi et al. (2011) have undertaken a rather thorough
investigation of adaptive knot selection in predictive process models. For now, our recommend-
ination is to keep the PP as simple as possible in the interest of being able to work with
a very large number of high-level parameters. Collecting over locations in $\mathcal{S}$, we obtain

$$W_\theta(\mathcal{S}) = B_\theta(\mathcal{S})u$$

where $B_\theta(s_i)$ is the $i$-th block, $i = 1, 2, \ldots, n$. Therefore, $W_\theta(\mathcal{S}) \sim \mathcal{N}(0, \gamma_\theta B(\mathcal{S})B^T(\mathcal{S}))$. While the variance parameter is modeled as

$$\log \gamma_\theta \sim U(\cdot, \cdot),$$

where $U(\cdot, \cdot)$ represents a uniform prior, we leave the spatial range of $W_\theta$ to be specified and fixed by the modeler due to limited information about the spatial range and marginal variance of a Gaussian process (Zhang, 2004).

This framework enables learning about $\theta(s)$ and $\tau(s)$ by borrowing information from measurements of explanatory variables, $X_\theta(s)$ and $X_\tau(s)$ that drive nonstationary behavior with fixed effects $X_\theta\beta_\theta$ and $X_\tau\beta_\tau$, respectively. If such variables are absent, then $X_\theta$ and $X_\tau$ consist of an intercept only. The associated regression coefficients are customarily assigned a Gaussian prior specified by the user. The structure of the fixed effects and the random effects is, therefore, very similar with $B$ playing the same role as $X_\theta$, the difference between them being that the prior variance of $\tilde{W}_\theta$ is estimated by the model while the prior variance of $\beta_\theta$ is fixed by the user. If $\theta(s)$ is a matrix, as occurs with anisotropic range parameters in (6), the above framework is easily modified. Let $\theta(s) = A(s)$ be a $d \times d$ positive definite matrix with positive eigenvalues $\lambda_1(s), \lambda_2(s), \ldots, \lambda_d(s)$. If $A(s) = P(s)\Lambda(s)P(s)^T$ is the spectral decomposition, then we use the matrix logarithm

$$\log A(s) = P(s)\log(\Lambda(s))P(s)^T,$$

where $\log(\Lambda(s))$ is the diagonal matrix with $\log(\lambda_i(s))$ for $i = 1, 2, \ldots, d$ as the diagonal elements. The matrix logarithm maps the positive definite matrices to the symmetric matrices (but not necessarily positive definite) and $(\log(A(s))^{-1}) = -\log(A(s))$, which is convenient for parametric specifications. Let symmat(·) denote the symmetric matrix operator (or the inverse map of the half-vectorization operator vech(·)) that fills the upper
and lower triangles of a matrix from a vector of suitable length to form a symmetric matrix. The model for \( \log(\theta(s)) \) in (11) now is

\[
\log(A(s)) = \text{symmat}((X_A(s)\beta_A + W_A(s))M(d)),
\]

where \( X_A(s)\beta_A + W_A(s) \) is a vector of length \( d \times (d + 1)/2 \), \( s \in \mathbb{R}^d \), \( M \) is a square matrix of size \( d(d + 1)/2 \), and \( \beta_A \) is \( n_{X_A} \times d(d + 1)/2 \). The symmetry of \( \log(A(s)) \) leads to the positive-definiteness of \( A(s) \). For the sake of interpretation, we apply a change of basis to the half-vectorization of \( \log(A(s)) \) using \( M(d) \). The first row of \( M(d) \) is \( \text{vech}(I_{d(d+1)/2}/\sqrt{d \times (d+1)/2}) \) and the other rows are obtained by applying \( \text{vech}(\cdot) \) to any completion of basis of the symmetric matrices from \( I_{d(d+1)/2}/\sqrt{d \times (d+1)/2} \). The first coefficient of \( X_A(s)\beta_A + W_A(s) \) controls the determinant of \( A(s) \), which yields the product of the range ellipse axis lengths, and will inflate and deflate the range of the ellipse without changing its shape. The other components will rotate and squash the ellipse without changing the product of the lengths of its axes. Further, we specify \( W_A(s) = B_A(s)\tilde{W}_A \), where \( \tilde{W}_A \) is \( p \times d(d + 1)/2 \). The variance parameter \( \gamma_A \) is \( d(d + 1)/2 \times d(d + 1)/2 \) with \( \text{vec}(\tilde{W}_A) \sim \mathcal{N}(0, \gamma_A \otimes I_p) \). We can now persist with the hyperpriors presented earlier. A Gaussian prior is assigned to \( \beta_A \), and a Uniform prior to the eigenvalues of \( \gamma_A \). If only the first columns of \( \beta_A \) and \( \tilde{W}_A \) are non-null, then this model is effectively equal to the locally-isotropic nonstationary range model presented earlier. If \( \tilde{W}_A \) is null and only the first row of \( \beta_A \) is non-null, then the model is anisotropic, but stationary. If only the upper-left coefficient in \( \tilde{W}_A \) is non-null, then the model is stationary and isotropic. An illustration of the kind of distributions obtained thus is presented in Figure 1. Figure 1(a) presents range ellipses generated with (12), while Figure 1(b) presents range ellipses generated with 11(d). Figures 1(c) and 1(d) represent one of their respective Gaussian process sample paths obtained with (11)(b).
(a) Ellipses obtained with matrix log GP
(b) Circles obtained with scalar log GP
(c) GP samples corresponding to the ellipses
(d) GP samples corresponding to the circles

Figure 1: Range ellipses and GP samples induced by the log-GP and matrix log-GP priors
We conclude this section with a few remarks. There is little to argue about the complexity encoded in (10) and (11). Inference problems in even simpler stationary models is encumbered by well-known issues of model identifiability (Stein, 1999; Zhang, 2004; Tang et al., 2021). These issues are exacerbated in nonstationary models with almost no theoretical tractability. It is, therefore, crucial to model the latent processes with reasonable priors. For example, while a non-informative prior on $\beta$ is reasonable, more informative priors for $\beta_\theta$ and $\beta_\tau$ would be desirable on the basis of information in $X_\theta$ and $X_\tau$. However, in the absence of such information, even with only intercepts in these matrices, we are able to demonstrate fairly robust inference (see Sections 5.1 and 5.2) with regard to the nonstationary processes. The fixed effects and the spatially-driven effects for the range and the field’s marginal variance cannot be independent. Indeed, some spatial incoherence of these effects is very likely to cause identifiability issues between the parameter field and the latent field. We refer the reader to the supplement and vignette for details.

4 Markov chain Monte Carlo algorithms

4.1 Outline

We devise a Gibbs sampler, where each parameter is updated conditionally on the rest, which is further improved by introducing appropriate updating methods built from other works or developed in this article. The latent field $w(\mathcal{S})$ is sampled using the blocked update (Datta et al., 2016, Section 4.1) in order to mitigate spatial auto-correlation. The regression coefficients $\beta$ are updated using the “Interweaving” method proposed in Coube-Sisqueille and Liquet (2021). The low-rank formulation of $W_\theta$ and $W_\tau$ ensures that only the basis coefficients need to be updated. These are block-sampled with their respective counterparts.
\( \beta_\theta \) and \( \beta_\tau \). The basis coefficients for \( W_\theta \) are updated using an interweaving strategy with respect to the latent field \( w(\cdot) \) inspired from (Filippone et al., 2013); see Section 4.2. We sample the basis coefficients for \( W_\tau \) from their full conditional distributions. In both cases, Hybrid Monte Carlo or HMC (see Section 4.3) is used yielding, respectively, a “HMC-within-interweaving” for \( W_\theta \) and “HMC-within-Gibbs” for \( W_\tau \). The higher-level parameters \( \{\gamma_\theta, \gamma_\tau\} \), are successively updated using an interweaving strategy on their respective latent fields. In the case of \( \gamma_\theta \), this interacts with the interweaving on \( w \). The resulting algorithm is outlined in Section S3.3 of the Supplement.

4.2 Ancillary-Sufficient Interweaving Strategy

The problem of high correlation between latent fields and higher-level parameters is familiar in stationary spatial models, and several solutions exist such as blocking (Knorr-Held and Rue, 2002), collapsing (Finley et al., 2019) or Interweaving (Filippone et al., 2013). Interweaving exploits the discordance between two parametrizations of a latent field to sample high-level parameters. When these two parametrizations are an ancillary-sufficient couple, we have an Ancillary-Sufficient Interweaving Strategy (AS-IS); see Yu and Meng (2011) and also Section S3.1 in the Supplement.

In our application, we interweave the whitened and natural parametrizations of the latent field \( w(S) \) from (11)(a) in order to update the higher-level parameters \( W_\theta(S) \) and \( \beta_\theta \) impacting the covariance structure decomposed in (11)(c) and (12). The so-called natural parametrization of the latent field is found in the decomposition (11)(a) and is a sufficient parametrization. The whitened parametrization of the latent field is ancillary and is obtained by multiplying the natural parametrization with the right NNGP factor of its prior precision matrix from (11)(b): \( w^*(S) = \tilde{R}w(S) \) so that the prior distribution (11)(b) be-
comes a standard normal distribution. The covariance parameters have no effect anymore on the prior distribution of the latent field. In turn, they acquire a role in the decomposition of the data. In (11)(a), \( w(s_i) \) is replaced by the \( i \)-th element of \( \tilde{R}^{-1}w^*(S) \), while 11(b) is replaced by \( w^*(S) \sim N(0, I_{|S|}) \). Further developments concerning the behavior of \( w^* \) are available in Coube-Sisqueille (2021).

The parameter of interest is sampled in two steps, one for each parametrization of the latent field. These individual steps can be full conditional draws, random walk Metropolis steps or, as in our case, Hybrid Monte-Carlo (HMC) steps. Hence, two potentials will be derived for the HMC steps of each parameter. Note that \( \gamma_\theta \) and \( \gamma_\tau \) are themselves covariance parameters for the latent fields \( W_\theta \) (11(d)) and \( W_\tau \) (11(f)). In order to update \( \gamma_\theta \) and \( \gamma_\tau \), interweaving is used by treating \( W_\theta \) and \( W_\tau \) as latent fields and using their respective whitened parametrizations. In the case of \( \gamma_\theta \), *nested interweaving* allows us to use the whitened parametrizations of \( \tilde{W}_\theta(S) \) and \( w(S) \).

### 4.3 Hybrid Monte-Carlo

Hybrid Monte-Carlo (HMC) (Neal et al., 2011) has been implemented successfully by Heinonen et al. (2016) for nonstationary Gaussian processes. Our approach differs in two aspects. First, we use NNGP instead of full GP and, hence, differentiate the NNGP-induced potential, which is non-trivial. Second, we use an “HMC within AS-IS” algorithm and must find the gradients of the potential for the covariance parameters using both ancillary and sufficient parametrizations. We jointly update the Predictive Process coefficients and the regression coefficients \( \beta_\tau \) and \( \tilde{W}_\tau \) in (11)(e), \( \beta_\theta \) and \( \tilde{W}_\theta \) in (11)(c), or \( \beta_A \) and \( \tilde{W}_A \) in (12). This method is especially useful for the range parameters since it avoids an unfeasible Metropolis-within-Gibbs sweep over \( \beta_\alpha \). The method applies to \( \alpha, \sigma \) or \( \tau \), so we
use “λ” as a portmanteau. We differentiate $H = -g(\beta_\lambda, W_\lambda) - \log(p(\beta_\lambda)) - \log(p(\bar{W}_\lambda|\gamma_\lambda))$.

Here, $g(\cdot, \cdot)$ depends on the role of $\lambda$ in the model and the chosen parametrization of the latent field in interweaving; $p(\cdot)$ denotes prior distributions whose parameters are either set by the user in the case of $\beta_\lambda$ or estimated by the model in the case of $\bar{W}_\lambda$. We express $g(\beta_\lambda, W_\lambda) = h \circ \log(\lambda)(\beta_\lambda, W_\lambda)$, $\log(\lambda)(\cdot, \cdot)$ being a function that links $\beta_\lambda, \bar{W}_\lambda$ and the logarithm of $\lambda$ following (11)(c), (11)(e), or (12). Using the Jacobian chain rule on $g(\beta_\lambda, \bar{W}_\lambda)$, the gradient can then be written as

$$\nabla_{\beta_\lambda, \bar{W}_\lambda} H = -\nabla_{\beta_\lambda} \log(p(\beta_\lambda)) - \nabla_{\bar{W}_\lambda} \log(p(\bar{W}_\lambda|\gamma_\lambda)) - J^T_{\beta_\lambda, \bar{W}_\lambda} \log(\lambda) \cdot \nabla_{\log(\lambda)} h(\log(\lambda)).$$

Recognizing that $\log(\lambda)$ varies linearly with respect to $\beta_\lambda$ and $\bar{W}_\lambda$, we get $J^T_{\beta_\lambda, \bar{W}_\lambda} \log(\lambda) = (X_\lambda|B)^T$. The details for computing the gradient of $h_\lambda$ are supplied in Section S4.1 of the Supplement. The cost of gradient computation depends on the specific parameter. While the noise variance boils down to a product of independent densities, the range and marginal variance of the latent field involve $\bar{R}$ for multiplication and solving. These operations, which would be unfeasible with dense matrices, remain efficient because of the triangular and sparse $\bar{R}$ (Datta et al., 2016). Furthermore, (9) allows for efficient computation of the gradient for parameters related to the field’s variance. For the range, (8) is critical to establish the fact that one range parameter only affects its children in the DAG, allowing for sparse differentiation, in synergy with the computation of the coefficients of $\bar{R}$. Overall, the cost of the gradient computation depends linearly on the number of observations.

5 Applications of the model

Our model and codes are available from the public repository https://github.com/SebastienCoube/Nonstat-NNGP and includes an extensive vignette for practitioners.
5.1 Synthetic experiments

In our synthetic experiments, we generate data sets ranging from stationary specifications to those exhibiting varying degrees of nonstationarity using (11). We randomly sample 12000 locations on a square with side length 5 using an Uniform distribution. Then, the regression coefficients in (11) (a), (c), and (e) are fixed. The only predictor in \( X \) is an intercept. We fix \( \beta = 0 \) and \( \beta_\tau = 0 \) in (11)(a) and (11)(e), respectively. In the stationary configuration, the noise and the latent field have same marginal variances equal to \( \exp(0) = 1 \). For \( \beta_\theta \) in (11)(c), the intercept coefficient for the range is set to \( \log(0.1) \approx -2.3 \) while that for the smoothness is 1.5. In the anisotropic cases, the intercept is set to 0. These parameters are the same for all runs, including our non-stationary examples.

The log-scale parameters \( \gamma_\theta, \gamma_\tau \) from (11)(g) change. In a stationary model, they are set to 0. In a nonstationary model, they are set to 1. After this, all the parameters are generated randomly, following their respective prior distributions (d), (f), (b) and (a) in (11). We subsequently fit several candidate models to the synthetic data we generated. We consider four situations: (i) “under-modeling” when the model is less complex than the data generating process; (ii) “over-modeling” which is the opposite of (i); (iii) the so-called “correct” (or “right”) modeling, when the specifications of the model and the data are the same; and (iv) models that fall in none of the above three categories. Several indicators were monitored: the Deviance Information Criterion (DIC) (Spiegelhalter et al., 1998), the mean square error (MSE) of the predicted field at unobserved sites, the MSE of the smoothed field at observed sites, and the MSE of the smoothed covariance parameters. We did not include predictions of the covariance parameters due to the smoothing of the PP. Further details are supplied in Section S5 of the Supplement.

Since the stationary models emerge as special cases of nonstationary models, we can
informally detect over-modeling from the MCMC chains without requiring to wait for full convergence by looking at the samples of the log-GP variance $\gamma$. This is done using trace plots, presented in Figures 2 and 3, where each curve corresponds to one Markov chain. When the data set is stationary while the model is nonstationary, $\gamma$ drops to very small values, such as in Figures 2(a) and 3(a), effectively inducing a stationary latent field. When the range is locally isotropic, but nonstationary, we observe the behavior of Figure 2(b), where the parameter controlling the range’s log-variance takes high values. In Figure 3, the first component, labeled “determinant”, controls the variance of the determinant of the range ellipses, allowing them to inflate or deflate over space, such as the parameters monitored in Figure 2. Two other parameters, labeled “anisotropy”, allow the range ellipses to squash (with area preservation) and change orientation. In Figure 3(b), only “determinant” is high. Hence, the ellipses may change their size, but not their shape, practically retrieving the locally isotropic covariance. However, here a simpler model is preferable.

The locally isotropic range and the heteroskedastic marginal variance of the process seem to be poorly identified. If we apply a misspecified model with one nonstationary parameter to a data set generated with the other parameter as nonstationary, then we still obtain inference competitive with the true model. The same finding holds for data sets where both parameters are nonstationary and we specify only one as nonstationary. However, if none are specified as nonstationary when at least one ought to be, the performance deteriorates. See Section S5.4 of the Supplement for details.

It is clear that the loss of performance caused by using a stationary model (the bottom rows in the tables of Section S5 in the Supplement) is usually lesser than the loss caused by a poorer model formulation (in red in the tables) or other under-modeling cases (in gray). From these results, we proffer that a nonstationary formulation should always accommodate
Figure 2: MCMC samples of the range PP log-variance parameter $\log(\gamma_\alpha)$ following the type of range model used in the data.

(a) Data with stationary range $(\gamma_\alpha = 0)$

(b) Data with nonstationary range $(\gamma_\alpha = 1)$

5.2 Scientific data analysis

We worked on a normalized difference vegetation index (NDVI) data set drawn from Zhang and Banerjee (2022), featuring over one million observations that exhibit anisotropy and also indicate spatially-varying (heteroskedastic) noise. We tested several models combining homoskedastic and heteroskedastic noise in (11)(e); and stationary and nonstationary, isotropic and anisotropic latent field as in (11)(c). The effective dimension of the data set
Figure 3: MCMC samples of the diagonal coefficients of the range PP log-variance parameter \( \log(\gamma_A) \), following the type of range model used in the data.

was reduced through rounding of coordinates: the model was trained on 925,119 observations on 97,294 distinct locations since several locations had multiple observations mapped to them after the rounding of coordinates. This reduction in the latent field’s dimension allows, similarly to triangulation of the spatial domain (Lindgren et al., 2011) allows more elaborate analysis while retaining all observations. The remaining \( \sim 10\% \) of the observations (spatially spread using a max-min heuristic) were used for validation. We executed an NNGP with 10 neighbors and the max-min ordering with smoothness parameter \( \nu = 1.5 \).

The nonstationary parameters were modeled using spatial effects as in equations (11)(e) and (11)(f), specified through predictive processes with Matérn smoothness 1.5 and range 0.2 with 50 knots. We also implemented an INLA-SPDE algorithm (Rue and Held, 2005; Lindgren and Rue, 2015; Bakka et al., 2019) as a benchmark, although the open version of R-INLA cannot model anisotropy (see, e.g., Fuglstad et al., 2015a, for some internal version implementations). Table 1 synthesizes the performances of the runs.
Table 1: Model comparisons for the NDVI analysis. For example, the second line refers to inference from an NNGP model with a stationary but anisotropic latent field range, and heteroskedastic noise variance. Other rows are similarly defined.

All NNGP runs were executed on the Pyrene computation cluster (32 threads and 64 GB RAM). Their MCMC convergence was monitored using univariate Gelman-Rubin-Brooks diagnostics (Gelman et al., 1992; Brooks and Gelman, 1998) and Effective Sample Sizes (ESS) for the high-level parameters (all the parameters of the models except the latent field $w(S)$). Table 1 presents the minimum ESS. The number of iterations ranged between 1000 and 1500, incurring an effective running time ranging from two to sixteen hours depending upon the specific covariance structure in use. For software compatibility reasons, the INLA runs were done on an Acer Predator Helios laptop, with 12 threads, 32 GB RAM and 100 GB swap. In spite of its lower number of threads, the laptop is no slower than a cluster node. While the stationary INLA delivered, as expected, very expeditious output, it struggled with the nonstationary models and was unable to deliver even after 24 hours with the exhaustion of the 132 GB of RAM and swap. This, too, is not unexpected and can be attributed to the larger number of hyper-parameters in our models. To summarize, INLA outperforms our proposed MCMC-based implementation in stationary and isotropic settings, while our method is a much more viable option for nonstationary modeling with

| method | aniso. | nonstat. | heterosk. | DIC     | train log-dens | pred. log-dens | min ESS | n. iter. | time (mn) | machine |
|--------|--------|----------|-----------|---------|----------------|----------------|---------|---------|-----------|---------|
| NNGP   | Yes    | Yes      | Yes       | -25845  | 0.03           | -0.20          | 87      | 1500    | 1498      | Pyrene  |
| NNGP   | Yes    | No       | Yes       | -18862  | 0.03           | -0.23          | 68      | 1500    | 1526      | Pyrene  |
| NNGP   | Yes    | No       | No        | 259491  | -0.11          | -0.35          | 97      | 1000    | 989       | Pyrene  |
| NNGP   | No     | No       | Yes       | -15707  | 0.03           | -0.31          | 49      | 1000    | 192       | Pyrene  |
| NNGP   | No     | No       | No        | 262442  | -0.11          | -0.41          | 92      | 1000    | 177       | Pyrene  |
| INLA   | No     | No       | No        | 255804  | cpo=-0.14      |                |         |         |           | Helios  |
complex covariance patterns (e.g., accommodating anisotropy and heterokedastic noise).

The models are compared using the DIC, the mean log-density of the observations, and the mean log-density of the predictions in Table 1. Nonstationary noise clearly boosts the model’s performance. The improvement brought by the elliptic covariance is more subdued, but the first three rows reveal that introducing more complexity in the anisotropy of models that are equivalent in every other aspect improves the DIC and the predicted log-density. The DIC of the INLA-fit model is coherent with the remainder of the non-heteroskedastic DICs. As for the validation score, it is computed with a leave-one-out log-density score. This method removes one observation of a spatial site, leaving the rest of the observations in the same spatial site, resulting in values between the smoothing and prediction scores. The nonstationary anisotropic model is also able to recover the spatially varying range and anisotropy of the observed data. Figure 4 summarizes the results from the full nonstationary model. Its first component, Figure 4(a) presents the raw observations. Figure 4(b) depicts the mean ellipses along with the mean latent field and shows that the ellipses fit the rays of the top-right corner, while they turn into small circles in the rougher left-hand side. The ellipses are inflated in the smoother regions reflecting increased range. Figure 4(c) maps the mean log variance of the noise revealing that the fuzzier left-hand side of the map is predicted with significantly more noise than the rest. Finally, Figure 4(d) presents the logarithm of the standard deviation of the summed latent field and fixed effects. The clearer dots correspond to locations left out for validation that incurred higher variance. Predictive variance and smoothing are clearly impacted by the covariance parameters and, in particular, by the noise variance. A final interesting observation is that the PP variance of the nonstationary range takes smaller values for the anisotropy components than for the determinant component. This is quite coherent with the map of the response variable 4(a):
we can see that there is an overall bottom left to top right anisotropy with little deviation, while the correlation range itself varies much more between, for example, the calm region of the center and the more troubled region on the upper left corner.

6 Summary and future directions

We have developed a class of hierarchical nonstationary NNGP models that overcome the bottlenecks of computational costs while offering nonstationary behavior through space-varying parameters in the spatial covariance kernel. We derive and exploit a closed form of the derivatives of the NNGP density with respect to spatially-indexed covariance parameters. These derivatives can also be used in HMC, MAP or other maximum likelihood approaches. The sparsity of NNGP factors along with dimension reduction using low-rank processes will allow scaling up the analysis to data sets with $\sim 10^7$+ locations. Despite the challenges of complex scientific data sets, we have shown that nested AS-IS is a viable strategy for multi-layered hierarchical models with large data augmentations. Regardless of linking environmental predictors and spatial basis functions to the covariance structure, further investigation on the behavior of the matrix-log GP model is warranted.

The proposed approach is applicable to any stationary kernel and should be explored beyond the Matérn kernels. Another possible extension is modeling in more than two dimensions. In particular, elliptic covariances in 3 dimensions may prove useful to quantify drifts such as rainfall moving across a territory. Ellipses in higher dimensions incur further differentiation, since the matrix logarithm of the range parameters will have 6 coordinates instead of 3. A computational scale up may be necessary. Given that we have derived the gradients of the model density, implementing Maximum A Posteriori (MAP) estimation is straightforward. Here, a key observation is that the high-level parameters appear to
Figure 4: Maps generated using the nonstationary anisotropic, heteroskedastic model on the NDVI data set.
have unimodal distributions. The MAP could be attained by adapting the Gibbs sampler presented here into a coordinate descent algorithm. While an empirical Bayes approach would likely suffice for applications such as prediction of the response variable or smoothing, finding credible intervals around the MAP can present challenges. The computational effort in flops and RAM saved from MCMC can be reallocated to implementing NNGP models with richer variants of Vecchia’s approximation Katzfuss et al. (2020). Finally, while the approach is applicable to non-Gaussian data by modeling the data using exponential families, further investigations in these lines are needed to evaluate the effectiveness of estimating weakly identifiable parameters.

References

Bakka, H., Vanhatalo, J., Illian, J. B., Simpson, D., and Rue, H. (2019). Non-stationary gaussian models with physical barriers. *Spatial Statistics*, 29:268–288.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. CRC press.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):825–848.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455.

Coube-Sisqueille, S. (2021). *MCMC algorithms and hierarchical architectures for spatial modeling using Nearest Neighbor Gaussian Processes*. PhD thesis, Université de Pau et des Pays de l’Adour.
Coube-Sisqueille, S. and Liquet, B. (2021). Improving performances of mcmc for nearest neighbor gaussian process models with full data augmentation. *Computational Statistics & Data Analysis*, page 107368.

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):209–226.

Cressie, N. and Wikle, C. K. (2015). *Statistics for spatio-temporal data*. John Wiley & Sons.

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111(514):800–812.

Filippone, M., Zhong, M., and Girolami, M. (2013). A comparative evaluation of stochastic-based inference methods for gaussian process models. *Machine Learning*, 93(1):93–114.

Finley, A. O., Datta, A., Cook, B. D., Morton, D. C., Andersen, H. E., and Banerjee, S. (2019). Efficient algorithms for bayesian nearest neighbor gaussian processes. *Journal of Computational and Graphical Statistics*, 28(2):401–414.

Fuentes, M. (2002). Spectral methods for nonstationary spatial processes. *Biometrika*, 89(1):197–210.

Fuglstad, G.-A., Simpson, D., Lindgren, F., and Rue, H. (2015a). Does non-stationary spatial data always require non-stationary random fields? *Spatial Statistics*, 14:505–531.

Fuglstad, G.-A., Simpson, D., Lindgren, F., and Rue, H. (2015b). Interpretable priors for hyperparameters for gaussian random fields. *arXiv preprint arXiv:1503.00256*. 28
Gelfand, A. E., Diggle, P., Guttorp, P., and Fuentes, M. (2010). *Handbook of Spatial Statistics (Chapman & Hall CRC Handbooks of Modern Statistical Methods)*. Chapman & Hall CRC Handbooks of Modern Statistical Methods. Taylor and Francis.

Gelfand, A. E., Fuentes, M., Hoeting, J. A., and Smith, R. L. (2019). *Handbook of environmental and ecological statistics*. CRC Press.

Gelman, A., Rubin, D. B., et al. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472.

Guhaniyogi, R., Finley, A. O., Banerjee, S., and Gelfand, A. E. (2011). Adaptive gaussian predictive process models for large spatial datasets. *Environmetrics*, 22(8):997–1007.

Guhaniyogi, R., Finley, A. O., Banerjee, S., and Kobe, R. K. (2013). Modeling complex spatial dependencies: Low-rank spatially varying cross-covariances with application to soil nutrient data. *Journal of Agricultural, Biological and Environmental Statistics*, 18:274–298.

Guinness (2018). Permutation and grouping methods for sharpening gaussian process approximations. *Technometrics*, 60(4):415–429.

Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., Katzfuss, M., et al. (2019). A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics*, 24(3):398–425.

Heinonen, M., Mannerström, H., Rousu, J., Kaski, S., and Lähteenmäki, H. (2016). Non-stationary gaussian process regression with hamiltonian monte carlo. In *Artificial Intelligence and Statistics*, pages 732–740.
Hengl, T. (2009). A practical guide to geostatistical mapping.

Higdon, D. (1998). A process-convolution approach to modelling temperatures in the north atlantic ocean. *Environmental and Ecological Statistics*, 5(2):173–190.

Ingebrigtsen, R., Lindgren, F., Steinsland, I., and Martino, S. (2015). Estimation of a non-stationary model for annual precipitation in southern norway using replicates of the spatial field. *Spatial Statistics*, 14:338–364.

Katzfuss, M. and Guinness, J. (2021). A general framework for vecchia approximations of gaussian processes. *Statistical Science*, 36(1):124–141.

Katzfuss, M., Guinness, J., Gong, W., and Zilber, D. (2020). Vecchia approximations of gaussian-process predictions. *Journal of Agricultural, Biological and Environmental Statistics*, 25(3):383–414.

Kleiber, W. and Nychka, D. (2012). Nonstationary modeling for multivariate spatial processes. *Journal of Multivariate Analysis*, 112:76–91.

Knorr-Held, L. and Rue, H. (2002). On block updating in markov random field models for disease mapping. *Scandinavian Journal of Statistics*, 29(4):597–614.

Lemos, R. T., Sansó, B., Higdon, D., Mendelssohn, R., Wikle, C. K., Milliff, R. F., Lemos, R. T., and Sansó, B. (2009). A spatio-temporal model for mean, anomaly, and trend fields of north atlantic sea surface temperature [with comments]. *Journal of the American Statistical Association*, 104(485):5–25.

Lindgren, F. and Rue, H. (2015). Bayesian Spatial Modelling with R - INLA. *Journal of Statistical Software*, 63(19).
Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498.

Neal, R. M. et al. (2011). Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2.

Paciorek, C. J. (2003). Nonstationary Gaussian processes for regression and spatial modelling. PhD thesis, Citeseer.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika*, 86(3):677–690.

Pourahmadi, M. (2007). Cholesky Decompositions and Estimation of A Covariance Matrix: Orthogonality of Variance–Correlation Parameters. *Biometrika*, 94(4):1006–1013.

Risser, M. D. (2016). Nonstationary spatial modeling, with emphasis on process convolution and covariate-driven approaches. *arXiv preprint arXiv:1610.02447*.

Risser, M. D. and Calder, C. A. (2015). Regression-based covariance functions for nonstationary spatial modeling. *Environmetrics*, 26(4):284–297.

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*. Monographs on statistics and applied probability 104. Chapman & Hall/CRC, 1 edition.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van der Linde, A. (1998). Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. Technical report, Research Report, 98-009.
Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Series in Statistics. Springer-Verlag New York, 1 edition.

Stein, M. L., Chi, Z., and Welty, L. J. (2004). Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(2):275–296.

Tang, W., Zhang, L., and Banerjee, S. (2021). On identifiability and consistency of the nugget in gaussian spatial process models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(5):1044–1070.

Vecchia, A. V. (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):297–312.

Yang, H.-C. and Bradley, J. R. (2021). Bayesian inference for big spatial data using non-stationary spectral simulation. *Spatial Statistics*, 43:100507.

Yu, Y. and Meng, X.-L. (2011). To center or not to center: That is not the question—an ancillarity–sufficiency interweaving strategy (asis) for boosting mcmc efficiency. *Journal of Computational and Graphical Statistics*, 20(3):531–570.

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99(465):250–261.

Zhang, L. and Banerjee, S. (2022). Spatial factor modeling: A bayesian matrix-normal approach for misaligned data. *Biometrics*, 78(2):560–573.