Modeling and emulation of nonstationary Gaussian fields. ☆

Douglas Nychka[a,*], Dorit Hammerling[a], Mitchell Krock[b], Ashton Wiens[b]

[a] *National Center for Atmospheric Research, Boulder, Colorado, USA*
[b] *Department of Applied Mathematics, University of Colorado, Boulder, Colorado, USA*

Abstract

Geophysical and other natural processes often exhibit nonstationary covariances and this feature is important for statistical models that attempt to emulate the physical process. A convolution-based model is used to represent nonstationary Gaussian processes that allows for variation in the correlation range and the variance of the process across space. We apply this model in two steps: windowed estimates of the covariance function under the assumption of local stationarity and encoded local estimates into a single spatial process model that allows for efficient simulation. We show that nonstationary covariance functions based on the Matérn family can be reproduced by the LatticeKrig (LK) model, a flexible, multi-resolution representation of Gaussian processes. Stationary models based on the Matérn covariance are fit in local windows and these estimates are assembled into a single, global LK model. The LK model is efficient for simulating nonstationary fields even at 10^5 locations. This work is motivated by the interest in emulating spatial fields derived from numerical model simulations such as Earth system models. We successfully apply these ideas to emulate fields that describe the uncertainty in the pattern scaling of mean summer (JJA) surface temperature from a series of climate model experiments. The spatial covariance structure developed in this paper is not limited to emulation, and could also be used

☆This document is a collaborative effort.
Declarations of interest for all authors: none

*Corresponding author: Computational and Information Systems Laboratory, National Center for Atmospheric Research, PO Box 3000, Boulder, Colorado, USA, 80307-3000

Email addresses: nychka@ucar.edu (Douglas Nychka), dorith@ucar.edu (Dorit Hammerling), Mitchell.Krock@colorado.edu (Mitchell Krock), Ashton.Wiens@colorado.edu (Ashton Wiens)

for spatial prediction and conditional simulation for observational data and leverages embarrassingly parallel strategies for computational efficiency.

Keywords: Nonstationary Gaussian Process, Markov Random Field, Fixed Rank Kriging, NCAR Large Ensemble Experiment

1. Introduction

In many areas of the geosciences it is natural to expect spatial fields to be nonstationary. Not accounting for how the covariance function may vary over space can result in misinterpreting the amount of spatial correlation and also lead to unrealistic emulation of the spatial fields. As spatial datasets grow in size and often have global extent, it is more likely that one would expect nonstationary fields simply because the spatial domain covers a heterogenous region. This is often the case for surface climate fields where distinct land and ocean regions might be expected to exhibit different spatial structure.

Although large spatial data sets have the advantage of making it easier to identify nonstationary covariances, they pose computational challenges when one attempts to apply standard statistical models that involve solving a linear system and finding the determinant of an $n \times n$ covariance matrix, where n is the number of spatial locations. This feature is due to the well-known increase in computational burden that grows as $\mathcal{O}(n^3)$. Currently this feature effectively prohibits fitting and simulating from Gaussian spatial process models when the number of locations exceeds several thousand. Moreover, even for sample sizes where computation is still feasible, interactive spatial data analysis will always benefit from faster computation.

Given the spatial variation of a nonstationary covariance function it is natural to focus on local modeling of the spatial field. Besides reducing bias in the estimated covariance parameters, this strategy also finesses some computational problems by converting a single large problem into many smaller ones. A local approach does have the disadvantage that it may not lead to a global model for the covariance function or may imply a covariance model that is not readily computed. This work combines efficient local covariance estimates with a global model, LatticeKrig (LK, [1]), that can incorporate the local information. The LK model is designed for statistical computations for large data sets, and in particular it is possible to simulate realizations from this model and make spatial predictions with only modest computational resources. Although local covariance estimates in aggregate require

the same order of computation, the memory demands are smaller and the computations can be easily done in parallel. As a practical matter we exploit a large computing resource (the NCAR supercomputer Cheyenne [2]) for these computations and find that the computation time scales almost linearly up 1000 processors (cores).

This work is motivated by a substantial example from impact assessment modeling and Earth system science. We have 30 derived fields from the NCAR Large Ensemble Project (NCAR-LENS) [3] that indicate the variation in local surface temperature increase due to an increase in the global average. For each model grid box we find the slopes from a simple linear regression of the local grid box temperatures on the global mean temperature. This can be done for each of the 30 members of NCAR-LENS and a simple summary is to then find the mean of these slopes for each grid box. The global pattern of the mean of the slopes is illustrated in Figure B.1 and has the interpretation that a one degree change in global mean summer temperature will result on average in a change in local temperature according to the gridded values of this field. Such fields form the basis of the pattern scaling technique in climate science. One surprise from these different realizations in the NCAR-LENS is that there is significant variability about the mean scaling pattern in Figure B.1 (e.g. see bottom row Figure B.8) among the ensemble members. The data science goal then is to quantify this variability. This is a large spatial problem; the model grid is at approximately one degree resolution and so there are more than 55,000 spatial locations (288×192 grid). Since these data cover the entire globe, even subregions exhibit nonstationary behavior.

Due the nature of climate model ensemble simulations, one can assume that the 30 fields are independent replicates from the same climate distribution. The goal is to model these fields accurately and simulate additional realizations. A larger set of realizations will be useful for quantifying the uncertainty of impact assessment modeling of climate change. Earth system models are large computer codes that can take months to run at dedicated supercomputing centers. Strategies for extending the results using fast statistical emulators is an important application to save additional computing resources. Moreover, detailed statistical models often reveal features of the simulations not obvious from basic data analysis. The specific spatial application in this paper is part of a larger statistical emulation of surface temperature fields for extending model results to other conditions [4]. This application is typical of climate model ensemble experiments, and the availability of replicated fields facilitates estimating nonstationary covariance functions.

The next section provides some background to this problem and presents the convolution process model as a basis for considering nonstationary covariances. Section 3 outlines the LK statistical model that is useful for large spatial data sets. Section 4 gives evidence to show that this model can approximate more standard covariance families such as the Matérn. The application to the pattern scaling ensemble is covered in Section 5, followed by a discussion of the results. We conclude by highlighting the novel contributions presented in this paper.

2. Background

We assume that the field of interest can be approximated as a Gaussian spatial process, $y(\boldsymbol{s})$, with $\boldsymbol{s} \in \mathcal{D} \subset \Re^2$, and for convenience $\mathcal{D}$ to be a rectangle. Furthermore, assume that this field follows the additive model

$$y(\boldsymbol{s}) = z(\boldsymbol{s})^T \boldsymbol{d} + g(\boldsymbol{s}) + \epsilon(\boldsymbol{s}), \tag{1}$$

where $z(\boldsymbol{s})$ is a low-dimensional vector of known covariates at each location, $\boldsymbol{d}$ a vector of linear parameters, $g(\boldsymbol{s})$ is a mean zero, smooth Gaussian process, and $\epsilon(\boldsymbol{s})$ a Gaussian white noise process independent of g. The parameters $\boldsymbol{d}$ represent fixed effects in this model while g and ϵ are stochastic.

There are several features of the observational model that are specific to our climate model application. Let $\{\boldsymbol{y}^m\}$ index M replicate fields that are independent realizations of the additive model (1). Given N spatial locations $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N\} \subset \mathcal{D}$, the observations are $Y_{i,m} = \boldsymbol{y}^m(\boldsymbol{s}_i)$, M independent fields observed at N locations. We also assume that the observations are complete – every replicate field is observed at all locations, which is typical for climate model output. Thus $\boldsymbol{Y}$ can be represented as an $N \times M$ matrix. Besides assuming replicate fields, we also assume that there is no measurement error in the observations and the white noise process ϵ is an approximation to a fine scale process that is uncorrelated when sampled on the scale of the observation locations. Note that in other applications, the white noise process ϵ may be interpreted as observational error. In our case we consider it intrinsic to the stochastic model for the climate model output.

2.1. Gaussian process convolution models

Under the Gaussian process assumption, the distribution of y is determined by the covariance function for g and the variance function for ϵ. In

particular we set

$$E[g(\boldsymbol{s})g(\boldsymbol{s}')] = k(\boldsymbol{s},\boldsymbol{s}') \text{ and } E[\epsilon(\boldsymbol{s})^2] = \tau(\boldsymbol{s})^2.$$

Our main concern is to model k without assuming stationarity of the process, and to this end we use a convolution representation. We postulate that the LK model is a discrete approximation of a convolution process (3.3), and for this reason we will start with a review of relevant material.

Let ψ be a continuous and square integrable function in $\Re^2$ and normalized so that

$$\int_{\Re^2} \psi\left(\|\boldsymbol{u}\|\right)^2 d\boldsymbol{u} = 1.$$

Define the spatially varying kernel function for two dimensions as

$$H(\boldsymbol{s},\boldsymbol{u}) = \frac{1}{\theta(\boldsymbol{s})}\psi\left(\frac{\|\boldsymbol{s}-\boldsymbol{u}\|}{\theta(\boldsymbol{s})}\right),$$

and

$$k(\boldsymbol{s},\boldsymbol{s}') = \sigma(\boldsymbol{s})\sigma(\boldsymbol{s}')\int_{\Re^2} H(\boldsymbol{s},\boldsymbol{u})H(\boldsymbol{s}',\boldsymbol{u})\, d\boldsymbol{u}, \tag{2}$$

where we assume that $\theta(\boldsymbol{s})$ is at least piecewise continuous and is interpreted as a range parameter varying over space. Also note that if $\sigma \equiv 1$ then k is a correlation function. Based on this form, we see that k will always be a valid covariance function, as it can be formally derived from the process

$$g(\boldsymbol{s}) = \int_{\Re^2} H(\boldsymbol{s},\boldsymbol{u})\, dW(\boldsymbol{u})$$

with $dW(\boldsymbol{u})$ a two-dimensional standard white noise process.

The Matérn family is a popular choice for representing a covariance function and can also be interpreted with respect to process convolution. Let

$$\psi(d) = C(\nu)d^{\nu}K_{\nu}(d)$$

with $\nu > 0$ a parameter controlling the smoothness of the process, K_ν a modified Bessel function of the second kind, and C a constant depending on ν. Assume that $\theta(\boldsymbol{s}) \equiv \theta$ and let $H(\boldsymbol{s},\boldsymbol{u}) = \psi(\|\boldsymbol{s}-\boldsymbol{u}\|/\theta)$. Using the spectral representation of the Matérn it has been shown [5] that $k(\boldsymbol{s},\boldsymbol{s}')$ will also be a member of the Matérn family with scale parameter still θ. If ν_g is the smoothness for g, then H must have smoothness $\nu_H = \nu_g/2 - d/4$. For example, when $d = 2$ and $\nu_g = 2$, g is obtained by convolution using the

exponential covariance ($\nu_H = 1/2$). When the scale/range parameter is not constant, however, the derived covariance is not strictly Matérn and will not have a form that is readily computed.

A convolution model to represent nonstationarity of a Gaussian process has been addressed by many authors. In particular we highlight early work in this area as applied to ocean temperature data [6], [7] and the subsequent development [8]. Although not explicit, the more recent models based on stochastic partial differential equations can also be tied to this representation [9], [10]. If H is the Green's function for a partial differential operator, $\mathcal{L}$, then g can also be identified with the solution: $\mathcal{L}g(\boldsymbol{s}) = dW(\boldsymbol{s})$. An alternative to the convolution model is an explicit nonstationary covariance first proposed by Paciorek [11] and extended to include smoothness parameters [12]. Our understanding is that this model is derived as a scale mixture of Gaussian covariance convolutions and so will not be the same as the direct convolution model sketched above.

Some more recent work has addressed the computation for large datasets [5] and use of a low dimensional function for the covariance parameters [13]. Recent work by [14] also is amenable to large data sets but focuses on the Paciorek form of covariance. A common thread in this past work is an emphasis on spatial prediction rather than simulation of the unconditional, nonstationary process. Thus much of this work is not directly transferable to our application of statistical emulation.

2.2. Maximum likelihood estimates

To explain the algorithms for large data sets, we review the relevant statistical computations associated with Gaussian process inference. Although this work considers maximum likelihood for inference, we note that the extension to approximate Bayesian inference may also benefit from the computational shortcuts that we highlight. Let

$$K_{i,j} = \mathrm{Cov}(g(\boldsymbol{s}_i), g(\boldsymbol{s}_j)) = k(\boldsymbol{s}_i, \boldsymbol{s}_j),$$

and let R be a diagonal matrix with elements $R_{i,i} = \tau(\boldsymbol{s}_i)^2$. This gives the covariance matrix for $\boldsymbol{y}^m$ as $K + R$. Also let Z be a matrix with N rows where the i^{th} row is the covariate vector $z(\boldsymbol{s}_i)$. In vector/matrix form the likelihood for the complete data set, $\boldsymbol{Y} = [\boldsymbol{y}^1, \ldots, \boldsymbol{y}^M]$, is given by

$$\frac{N}{2}\log(2\pi) - \frac{M}{2}\log|K+R| - \sum_{m=1}^{M}\frac{1}{2}(\boldsymbol{y}^m - Z\boldsymbol{d})^T(K+R)^{-1}(\boldsymbol{y}^m - Z\boldsymbol{d}) \quad (3)$$

with the covariance matrices implicitly depending on the parameters (or fields of parameters) σ, τ and θ. For large data sets evaluating the likelihood poses the well-known computational hurdles of storing K, solving the linear system associated with $K+R$ and evaluating the determinant of $K+R$. In addition, for a nonstationary model, evaluating the covariance as a convolution may also involve significant computation if the integral does not have a closed form.

These features make it difficult to estimate nonstationary models. Here we take a local approach by assuming that the covariance function is approximately stationary in a small spatial neighborhood and we take the stationary parameter estimates for σ, τ and θ as representing the values of these parameter surfaces in the center of the neighborhood. This is not a new idea and has roots dating back to the early work on moving window Kriging [15], [16], [17] and is also similar to local likelihood ideas [18], [19].

In this work we add two new features to this method. First, we develop a computational framework that exploits a highly parallel approach to estimate the local parameters. Subsequently the local covariance estimates are encoded into a global spatial process model that is efficient for simulation.

3. A multi-resolution spatial process model

The process convolution model (2) is a useful nonstationary model but difficult to implement for large spatial data. Here we present an alternative, the LatticeKrig (LK) model, that is a good approximation to standard covariance families but is much more amenable to fast computation.

The LK model is one of several recent approaches to handle large spatial data in a consistent global way. The recent review [20] compares many of these methods with an emphasis on spatial prediction for a data set of 10^5 locations. The multi-resolution approximation [21], hierarchical nearest neighbor methods [22] and stochastic partial differential equation models [10] might all be alternatives to using LK for the global simulation.

The basic idea of LK is to adopt fixed rank Kriging (FRK, see [23], [24]) but model the *precision* matrix of the coefficients as a sparse matrix. This model draws on the work of FRK and stochastic PDEs but also adds a multi-resolution elaboration that greatly improves its flexibility. Moreover, the nonstationary LK model can be interpreted as a superposition of convolution-type processes at different spatial scales.

We assume that the process $g(\boldsymbol{s})$ is the weighted sum of L independent latent processes

$$g(\boldsymbol{s}) = \sum_{\ell=1}^{L} \sigma_\ell(\boldsymbol{s}) g_\ell(\boldsymbol{s}). \tag{4}$$

Here $g_\ell(\boldsymbol{s})$ is a Gaussian spatial process with mean zero and marginal variance of 1, and σ_ℓ is a nonnegative function for the marginal standard deviation of the field at level ℓ. Thus, the marginal variance for $g(\boldsymbol{s})$ is $\sum_{\ell=1}^{L} \sigma_\ell(\boldsymbol{s})^2$.

3.1. Multi-resolution basis

Each component g_ℓ is defined through a basis function expansion as

$$g_\ell(\boldsymbol{s}) = \sum_{k=1}^{m(\ell)} c_{k,\ell}\, \varphi_{k,\ell}(\boldsymbol{s}), \tag{5}$$

where $\varphi_{k,\ell}$, $1 \le k \le m(\ell)$, is a sequence of fixed basis functions and $\boldsymbol{c}_\ell$ is a vector of coefficients distributed multivariate normal with mean zero and covariance matrix, $\boldsymbol{Q}_\ell^{-1}$. Coefficients are assumed to be independent between the different levels.

To achieve a multi-resolution, the basis functions are formed from translations and scalings of a single radial function. The basis functions depend on a sequence of nested rectangular grids $\{\boldsymbol{u}_{j,\ell}\}$, where $1 \le j \le m(\ell)$ and $1 \le \ell \le L$. The grid spacing is kept at the same distance in both dimensions and decreases by a factor of 2 from ℓ to $\ell+1$. Let ϕ be a unimodal, symmetric function in one dimension, and for this work we assume that it is compactly supported on the interval $[-1, 1]$. We adopt a scale parameter δ to set the overlap of the basis functions and the basis functions are then defined as

$$\varphi^*_{j,\ell} = \phi(2^{\ell-1}\|\boldsymbol{s} - \boldsymbol{u}_{j,\ell}\|/\delta). \tag{6}$$

Here the * indicates that these are not exactly the final versions of the basis functions but will be normalized as described in the Appendix. Although an important detail for implementation, the normalization is not crucial for understanding the main features of this model.

3.2. Spatial Autoregressive Model

In the LK model, the spatial covariance for $\boldsymbol{c}_\ell$ at a given level ℓ is modeled as a nonstationary Markov random field. The coefficient vector $\boldsymbol{c}_\ell$ at a single level follows a spatial autoregression (SAR) and is organized by the node

points. Each coefficient $c_{k,\ell}$ is associated with a node point $u_{k,\ell}$ and will have up to four first-order neighbors. Denote this set $\mathcal{N}_k$. We assume that for a parameter field, $a(\boldsymbol{s})$, and $v_{k,\ell}$, *iid* $N(0,1)$ random variables, the coefficient fields satisfy

$$a(u_{k,\ell})\, c_{k,\ell} - \sum_{k^* \in \mathcal{N}_k} c_{k^*,\ell} = v_{k,\ell}, \tag{7}$$

where $a(\boldsymbol{s}) > 4$ for the process to be stable. Note that in this work we enforce the restriction that the a parameter field does not vary between levels and we discuss this point in the last section. The parameters $a(\boldsymbol{s})$ control the dependence of the GMRF within a level and for constant a are related to the range parameter. Specifically this GMRF has been studied in [10] and approximates a Matérn covariance with smoothness $\nu = 1.0$ and an approximate range parameter given by $1/\sqrt{a-4}$.

Let B_ℓ be the SAR matrix that is square with the same dimension as $\boldsymbol{c}_\ell$. The diagonal elements of B_ℓ are $a(u_{k,\ell})$, the off-diagonal elements are -1 at the positions of the nearest neighbors, and the remaining entries are zero. With this construction $B_\ell \boldsymbol{c}_\ell = \boldsymbol{v}_\ell$, and simple linearity implies that the precision matrix for $\boldsymbol{c}_\ell$ is $Q_\ell = B_\ell^T B_\ell$. Given this model for each level the basis functions are normalized so that the marginal variance of $g_\ell(\boldsymbol{s})$ is one (see Appendix) and the levels are assumed to be stochastically independent. The net result is an overall precision matrix for the coefficients that is block diagonal according to Q_ℓ. Note that because Q is formed from the SAR both the precision matrix and covariance matrix for $\boldsymbol{c}$ will be positive definite.

3.3. An approximate convolution process

We can also conjecture how this model behaves as a discretized convolution process. Let Φ_ℓ be the matrix with (i,j) elements $\varphi_{j,\ell}(\boldsymbol{s}_i)$. For a given level $1 \leq \ell \leq L$, a realization of $\boldsymbol{g}_\ell$ at the observations has the representation

$$\boldsymbol{g}_\ell = \Phi_\ell B_\ell^{-1} \boldsymbol{v}_\ell \tag{8}$$

where the matrix multiplications in this expression are sums over the lattice points. Given that the lattice is equally spaced and the support of the basis functions is calibrated to overlap several lattice points, this expression may approximate integrals over the spatial domain. From the discussion in [5] (see Table 1) B_ℓ^{-1} can be associated with a Matérn kernel with smoothness 0 and is denoted as K_0, the modified Bessel function of the second kind of

order 0 as in [5]. We conjecture that $\Phi_\ell B_\ell^{-1}$ is approximated by

$$\int_{\Re^2} \phi(2^\ell \|\boldsymbol{x}-\boldsymbol{w}\|/\delta) K_0(\|\boldsymbol{w}-\boldsymbol{u}\|/\kappa(\boldsymbol{s}))\, d\boldsymbol{w}.$$

with $\kappa(\boldsymbol{s}) = 1/\sqrt{a(\boldsymbol{s})-4}$. Although K_0 has a singularity at 0, the convolution with the Wendland basis functions is smooth at zero and will result in a bounded kernel H.

Computational efficiency and simulation

To simulate from the LK model it is enough to simulate a realization of the coefficients since the basis is fixed. Also note that the coefficients between levels are independent. Focusing on the ℓ^{th} level, B_ℓ is a sparse matrix with at most 5 nonzero entries per row. Thus $Q_\ell = B^T B$ will also be sparse with at most 13 nonzero entries. Accordingly let $Q_\ell = AA^T$ be the sparse Cholesky decomposition with A a sparse, lower triangular matrix and let $\boldsymbol{v}$ be a vector of *iid* $N(0,1)$ random variables. Then $\boldsymbol{c}_\ell^*$ is simulated by solving the sparse linear system

$$A^T \boldsymbol{c}_\ell^* = \boldsymbol{v}. \tag{9}$$

g_ℓ is now evaluated using (5) and the components are added. Note that evaluating g_ℓ in (5) will also be efficient due to the compact support of the basis functions. In addition, computational burden is not increased by introducing nonstationarity through varying the a and σ_ℓ parameter fields.

4. Approximating the Matérn family as a multi-resolution

We first provide some results to demonstrate how the LK model approximates a stationary Matérn model and then generalize this connection with two examples for a nonstationary convolution process.

4.1. Stationary approximation

The theory in [25] proves an asymptotic result that indicates that the LK model can approximate the smoothness of members of the Matérn family as the number of levels becomes infinite. The main finding is that σ_ℓ should be chosen to decay as $2^{-\ell\nu}$ to approximate a Matérn process with smoothness ν. To build the best approximation over a limited number of resolution levels, however, it is more accurate to optimize the LK parameters numerically.

For an approximation criterion we focus on the accuracy of simulating a random field. For an integer K we generate a grid of $(2K+1)^2$ locations at unit spacing for the spatial domain $[-K, K] \times [-K, K]$. Given smoothness and range parameters, the Matérn correlation matrix is evaluated at these locations. Denote this matrix by $\Gamma(\theta, \nu)$ and let $\omega_M(\theta, \nu)$ denote the center row of the symmetric square root matrix for $\Gamma(\theta, \nu)$. In an analogous way for LK parameters a and $\{\sigma_\ell\}$ with $\ell \in \{1, 2, \ldots, L\}$, let $\omega_{LK}(a, \{\sigma_\ell\})$ denote the center row for the symmetric square root of the LK correlation matrix. We determine the best representation for the Matérn as the values for a and $\{\sigma_\ell\}$ that minimize

$$\left(\|\omega_M(\theta, \nu) - \omega_{LK}(a, \{\sigma_\ell\})\|_2\right)^2 \tag{10}$$

with $\|\cdot\|_2$ being the usual Euclidean vector norm, the set $\{\sigma_\ell\}$ nonnegative and summing to 1, and $a > 4$. The motivation for this criterion is based on simulating the Matérn process at the grid locations using $\Gamma(\theta, \nu)^{1/2}\boldsymbol{v}$, where $\boldsymbol{v}$ is a vector of independent $N(0, 1)$ random variables. The weights, $\omega_M(\theta, \nu)$, are applied to this random vector to obtain the simulated field value at the center of the domain. Thus $[(\omega_M(\theta, \nu) - \omega_{LK}(a, \{\sigma_\ell\})]^T \boldsymbol{v}$ will be the difference between this center value simulated under the Matérn model and that simulated under the LK model. The sum of squares in (10) will be the variance of this difference. We believe that this criterion is appropriate for approximating the simulated random field and we focus on the central grid location to minimize any boundary effects. Also we focus on the simulation error because this project is concerned with accurate emulation of the model output.

This approximation strategy is applied for $K = 10$ (a 21×21 grid) and the parameters are optimized using the R `optim` function. We consider 3 levels of multi-resolution with the coarsest having 2.0 unit spacing for the nodes. Figure B.2 summarizes this approximation for different sets of Matérn parameters. For example, for θ in the interval $[1, 12]$ and for a smoothness of $\nu = 1.0$ the LK model can approximate the Matérn to within a few percent of relative root mean squared error. We obtain approximations with less than 6 percent relative root mean squared error for the case $\nu = 2.0$ over the interval $[1, 8]$. Moreover, we expect the approximation to improve if additional levels are added.

These results can simply be scaled to other domain sizes by adjusting the spacing of the basis function grid at the coarsest level of the LK model. In

this way our results form the basis of a general approximation table to map Matérn covariance models into ones in the LK framework. The significance of these results is that given a locally stationary Matérn process one can identify parameters a and $\{\sigma_\ell\}$ for the LK model that give an accurate approximation. This approximation is tested under nonstationary specifications in the next section.

4.2. Approximating nonstationary fields

We use two informative test cases to explore the properties of the LK approximation to nonstationary, convolution-type processes. The spatial domain is taken to be $[-24, 24] \times [-24, 24]$. For the first case, we divide the spatial domain vertically into regions with two different correlation ranges:

$$\theta(\boldsymbol{s}) = \begin{cases} 5 & \text{for } s_1 \leq 0 \\ 1.9 & \text{for } s_1 \geq 0 \end{cases}$$

while fixing $\sigma(\boldsymbol{s}) \equiv 1$ and $\tau(\boldsymbol{s}) \equiv 0$. Here $\mathbf{s} = (s_1, s_2)$ and this can be considered an idealized test case of an land/ocean boundary from the climate model example.

The nonstationary field was defined by this range parameter and convolving exponential kernels according to (2). Based on the properties of the Matérn we expect a stationary Matérn covariance function with smoothness 2 when θ is constant.

Figure B.3 illustrates the exact correlation functions of the field at two locations along the the transect where the y-coordinate is 0 ($s_2 = 0$). Each evaluation of the correlation function now requires a numerical integration and so the comparison was restricted to a horizontal transect was done in order to limit the amount of computation. For reference, superimposed are the Matérn covariance functions assuming local stationarity. Note that these tend to track the nonstationary curves except at the boundary where θ is discontinuous. Also note the surprising lack of monotonicity in the correlation function at the location $(7, 0)$. Figure B.4 reproduces these true nonstationary correlation functions and superimposes the correlation functions from the LK approximation. Here the LK model is encoded to be a locally stationary Matérn approximation with a spatially varying a parameter. The precise value of a is found by interpolating $\theta(\boldsymbol{s})$ to the node points and then converting θ to a using the stationary approximation described in the previous section. Overall the LK model appears to capture the general features of

the nonstationarity and the transition from $\theta = 5$ to 1.9 on the line $s_1 = 0$. The LK model makes a smoother transition, however, across this boundary tending to overestimate correlations for locations closer to the discontinuity. The LK nonstationary model also misses the departure from monotonicity in the correlation function. The second nonstationary test case is setup similar to the first except that $\theta(\boldsymbol{s})$ is set to vary as a linear ramp function in s_1 decreasing from 6 at the left boundary to 1 at the right. Figure B.4 compares the true correlation functions to those approximated by the LK model. In this case the agreement is good and we attribute this to the smoothly varying choice for the $\theta(\boldsymbol{s})$ field.

Figure B.6 is a realization of the LK approximation for the first test case and gives a qualitative impression of the variation in the correlation scale across the discontinuity in $\theta(\boldsymbol{s})$ (grey vertical line). The three previous plots only depict the correlation along the transect $s_2 = 0$, indicated by the black line in this figure. To simulate the true field in this first case would not be difficult because $\theta(\boldsymbol{s})$ is piecewise constant. In general the simulation would be computationally intensive, however, requiring a separate convolution kernel computation for each location in the field. Even if the nonstationary matrix could be assembled there is still the challenge of computing the Cholesky factor for a large and dense covariance matrix. In contrast, the LK realization is found for a 129×129 grid and took under 20 seconds on a MacBook Air laptop (Intel Core i7, 2.2 GHz, 8 GB memory) using serial code and the `LatticeKrig` package in R.

5. Simulating variation in pattern scaling fields

As outlined in the introduction our application is to model the spatial variation among the patterns derived from the NCAR-LENS project. We will focus on a North and South America sub-domain to streamline presentation comprising $13,056 = 102 \times 128$ grid boxes. We found that a Matérn with $\nu = 1.0$ was a reasonable choice for smoothness across the domain and an isotropic Matérn covariance was fit locally using several sizes of moving square windows. Here we report estimates based on 11×11 grid box windows with the maximum likelihood estimates registered to the center grid box. Although this domain is a section of the sphere we develop the spatial model as if the model grid points are on a rectangular grid. This is a reasonable approximation because the domain is small enough where the periodicity in longitude and the singularity at the poles are not issues. Also because

we allow the correlation range to vary over the domain it will provide some adjustment to the longitude distance being a function of latitude.

5.1. Local covariance estimates

Even with 30 replicate fields, estimating the θ and σ parameters was not robust and we often obtained very large values over the ocean. This sensitivity is expected for large correlation ranges but we note that τ is still adequately estimated and is small over the ocean reflecting a smooth spatial field. Let σ_{obs} be the sample standard deviation for the replicates and for each grid box. A simple adjustment to the variance parameters is for $\hat{\tau} < .003$ and $\hat{\sigma} > \sigma_{obs}$ we take take $\hat{\sigma} \equiv \sigma_{obs}$. For $\hat{\theta} > 15$ we set $\theta \equiv 15$. Admittedly, these are crude adjustments but they respect the basic assumptions of more spatial coherence over the ocean and also the fact that correlation ranges beyond 15 degrees ($>$ 1600 km at the equator) are not likely and will not influence the simulation of the fields.

Previous work (e.g. [5], [14]) has treated the local covariance estimates as being under-smoothed and applied a second smoothing step to the estimated parameter field to improve its accuracy. We investigated this issue for these data by fitting an approximate thin plate spline (the function `fastTps` from the `fields` package [26]) to the log of the estimated θ field. The smoothing parameter was found by maximum likelihood and given 13,056 observations the effective degrees of freedom for the spline was over 3,500. This result does not suggest the need for additional smoothing of the local θ estimates. We also fit a thin plate spline model with the land/ocean mask added as a linear covariate and this did not change the results. Given this data analysis we concluded that there was little benefit in adding a second modeling step in representing the range parameter field.

5.2. Simulation of the pattern scaling uncertainty

Figure B.7 reports the Matérn estimates based on the above discussion. Perhaps the most important aspect of these data is the striking nonstationarity in all three parameter fields and the clear land/ocean demarcations along much of the coast line. We believe that this clear signal between land/ocean in the parameters suggests that our choice of window size is appropriate and overall the parameter fields are reasonable. The higher variability (σ) in the spatial process (g) in the center of North America and over the land area near Argentina is reasonable, along with a larger white noise component (τ)

over land. Although not shown, the ratio of white noise to smooth process variance (τ^2/σ^2) is small but tends to be larger over land.

The implementation of the LK model is available in the R `LatticeKrig` package [25]. These parameters were encoded into a LK model with three levels of resolution where the coarse grid spacing is 2.5 degrees. The fields were simulated on the grid of the model and took under 60 seconds on a MacBook Air system. Almost all of that time was in setting up the matrices Φ_ℓ and the computing the Cholesky decompositions of $\{Q_\ell\}$ and there is little overhead for generating more than one realization.

Figure B.8 shows four realizations of the LK process on the top row, and for reference the first four ensemble members from the spatial data set are given on the bottom row. Qualitatively, simulated and true cases have similar spatial coherence and variability. We note that the emulation, however, does have some modest deficiencies. For example, the anisotropy over the Equatorial Pacific is not well represented. In the model there appears to be longer correlation scales in the East-West direction as compared to the North-South. Of course, this is not a failing of the LK approximation but rather the use of an isotropic covariance function. As a contrast to the nonstationary model we also generated stationary realizations. The top row of Figure B.9 gives four realizations of a stationary field using the median of the parameter estimates over land. The bottom row is the same except the medians over the ocean are used. To aid in this comparison, we use the same white noise vectors for generating the land and ocean field in each column. The differences between these two choices of stationary models are striking and it is clear that neither would provide an accurate emulation of the model output.

5.3. Parallel implementation

This example was computed using a parallel strategy and the R language [27]. Fitting the spatial model for each window is an *embarrassingly parallel* operation and moreover the ensemble data set fields are relatively small (about 12Mb). We took the approach of using a supervisor R session and then spawning many R worker sessions. The supervisor session assigns tasks (i.e. specific local windows) to each worker based on balancing the work load. When a worker is done with a specific task the information from the fitting is passed back the supervisor. The complete set of results are assembled as an output list in the supervisor session and in our case this output list has as many components as grid boxes in the spatial domain and each

contains the results of the local fitting. Creating the workers, broadcasting the spatial data set, and assigning the tasks is all done in R through the `Rmpi` package [28]. In using R we have leveraged the stable and rich set of spatial analysis tools that are available to the community. In particular, the maximum likelihood estimates are found using the `spatialProcess` function from the `fields` package for the Matérn model and the `LatticeKrig` function from the `LatticeKrig` package, and these functions are called in exactly the same way as on a laptop. We have used this approach on the NCAR supercomputer *Cheyenne* [2] and found it exhibits excellent (strong) scaling. An example of timing is given below in Figure B.10. In this test case, a one level LK model limited to a 1000 grid boxes is fit directly to the data rather than the Matérn covariance. Here we see linear scaling in the time with up to 1000 parallel R worker sessions. As expected the time to spawn workers shows a linear increase (orange points) but is an order of magnitude smaller than the time spent in computation (blue points). Note that this scaling has attractive practical implications. Using 1000 cores will result in nearly a factor of 1000 speedup in the analysis and can potentially convert a lengthy serial analysis into one that is almost interactive.

6. Discussion

Combining local covariance estimation with a global model provides a practical route for modeling and simulating large spatial data sets. We have shown that the LK model can reproduce abrupt nonstationarity in a process where the range parameter has a discontinuity, and as expected, also does well when the range parameter varies smoothly across a spatial domain. Moreover, in places where the process is locally stationary we see that there is close agreement between the Matérn correlation function and the approximate one from the LK representation. The advantage of the LK representation is the ability to generate an unconditional realization of the process at large number of locations. One can also use the LK model for spatial prediction and inference [20] although that role is not needed for climate model emulation.

Most data analysis represents a compromise between model complexity and realism and the need to estimate the model accurately from data. In our application of this model, we used data with spatial replicates, which makes parameter estimation much more stable. We do not believe this data set to be an isolated example, as ensemble climate experiments are now the norm

in climate science. The local covariance models could be improved by adding anisotropy and also covariates for the land/ocean regions. Allowing the correlation range to vary independently in the vertical and horizontal dimensions should accommodate the changing distances in the spherical geometry. The addition of anisotropic parameters, however, will not involve substantially more computational resources and only requires the modification of the SAR weights to be different in the horizontal and vertical coordinates. Because the LK likelihood can be evaluated for the complete data set, there is the opportunity to fit parameters that have a global extent, such as land/ocean effects, along with local covariance parameters. The estimation strategy in this case would have the flavor of back-fitting in additive models where one would cycle among fitting different components of the model and, when appropriate, exploit embarrassing parallel computation.

The use of embarrassingly parallel steps, such as local covariance fitting or local simulation, is a computational strategy that merits more attention. Here we have developed code mainly in R to manage this process and so this framework is accessible to any accomplished R user. Indeed, the framework we use on the supercomputing system is the same that we use on a laptop except for several lines of batch scripting and changing directory pathnames. We also believe that this style of computation may drive alternative models and algorithms as the number of processors/cores available for routine spatial data analysis grows.

Conclusion

In this work, we have developed a new method to model second-order nonstationarity in spatial data. First, local spatial Matérn parameters are estimated through local fitting by maximum likelihood. The local Matérn MLE's are then encoded into a global LK model and we have given numerical evidence that this translation is an accurate representation of a convolution type nonstationary process. This approach yields advantages in both modeling and computation. The interpretation of LK as a discrete approximation to a convolution process (3.3) allows for efficient simulation of realizations at a large number of locations.

To our knowledge, this two step approach of local fitting followed by global simulation is new, and our application is a large data set relative to other examples in the literature. Figures B.7, B.8 and B.9 collectively give a clear and convincing message that the nonstationarity in the data is being captured

by the LK model. Stationary models fail to identify the covariance structure of the data, and produce unrealistic emulations as shown in Figure B.9. In contrast, the nonstationary model incorporates interpretable parameters in a global model that produces realistic emulations of the climate model fields, as shown in the top row of Figure B.8.

In closing we emphasize that modeling nonstationary data in this manner is not limited to climate model applications. The `LatticeKrig` package is a freely available through CRAN, easy to use, and one of the few nonstationary models for large spatial problems.

Acknowledgements

This work was supported in part by the National Center for Atmospheric Research (NCAR) and also the National Science Foundation Award 1406536. NCAR is sponsored by the National Science Foundation and managed by the University Corporation for Atmospheric Research. We also acknowledge high-performance computing support from Cheyenne (doi:10.5065/D6RX99HX) provided by NCAR's Computational and Information Systems Laboratory, sponsored by the National Science Foundation.

Appendix A. Wendland radial basis kernel

The Wendland functions are compactly supported on $[0, 1]$ and are also positive definite. Below is the version of the Wendland valid up to 3 dimensions and belonging to C^4:

$$\phi(d) = \begin{cases} (1-d)^6(35d^2 + 18d + 3)/3 & \text{for } 0 \leq d \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

This is implemented as the function `WendlandFunction` in the `LatticeKrig` R package.

Appendix B. Normalization to approximate stationarity

Because of the discrete nature of the SAR the marginal variance of the LatticeKrig process will not be constant in the spatial domain. This can cause artifacts in the estimated surface and compromise its ability to approximate stationary covariance functions. To adjust the marginal variances we compute the unnormalized variance, across space, and divide by this quantity to give a constant variance at any location.

Based on the model and notation from Section 3, let $\boldsymbol{C} = \boldsymbol{Q}_\ell^{-1}$ and at multi-resolution level ℓ,

$$\mathrm{Var}(g_\ell(\boldsymbol{s})) = \sum_{j,k} \varphi_{j,\ell}(\boldsymbol{s}) \boldsymbol{C}_{j,k} \varphi_{k,\ell}(\boldsymbol{s})$$

Accordingly, let $\omega(\boldsymbol{s}) = \sqrt{\mathrm{Var}(g_\ell(\boldsymbol{s}))}$ and normalize the basis functions as

$$\varphi_{j,\ell}(\boldsymbol{s}) = \frac{\varphi^*_{j,\ell}(\boldsymbol{s})}{\omega(\boldsymbol{s})}$$

These are the actual basis functions used in the LK model and for spatial analysis.

References

[1] D. Nychka, S. Bandyopadhyay, D. Hammerling, F. Lindgren, S. Sain, A multiresolution gaussian process model for the analysis of large spatial datasets, Journal of Computational and Graphical Statistics 24 (2015) 579–599.

[2] Computational and Information Systems Laboratory, Cheyenne: SGI ICE XA System (Climate Simulation Laboratory), National Center for Atmospheric Research, Boulder, Colorado, USA (2017).

[3] J. Kay, C. Deser, A. Phillips, A. Mai, C. Hannay, G. Strand, J. Arblaster, S. Bates, G. Danabasoglu, J. Edwards, et al., The community earth system model (cesm) large ensemble project: A community resource for studying climate change in the presence of internal climate variability, Bulletin of the American Meteorological Society 96 (2015) 1333–1349.

[4] S. E. Alexeeff, D. Nychka, S. R. Sain, C. Tebaldi, Emulating mean patterns and variability of temperature across and within scenarios in anthropogenic climate change experiments, Climatic Change (2016) 1–15.

[5] Z. Zhu, Y. Wu, Estimation and prediction of a class of convolution-based spatial nonstationary models for large spatial data, Journal of Computational and Graphical Statistics 19 (2010) 74–95.

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

[7] D. Higdon, et al., Space and space-time modeling using process convolutions, Quantitative methods for current environmental issues 3754 (2002) 37–56.

[8] M. Fuentes, Spectral methods for nonstationary spatial processes, Biometrika 89 (2002) 197–210.

[9] F. Lindgren, H. Rue, Explicit construction of GMRF approximations to generalised Matrn fields on irregular grids, Technical Report, [Publisher information missing], 2007.

[10] F. Lindgren, H. Rue, J. Lindström, 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 (2011) 423–498.

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

[12] M. L. Stein, Nonstationary spatial covariance functions, Unpublished technical report (2005).

[13] G.-A. Fuglstad, D. Simpson, F. Lindgren, H. Rue, Exploring a new class of non-stationary spatial gaussian random fields with varying local anisotropy, Statistica Sinica 25 (2015) 115–133.

[14] F. Fouedjio, N. Desassis, J. Rivoirard, A generalized convolution model and estimation for non-stationary random functions, Spatial Statistics 16 (2016) 35–52.

[15] T. C. Haas, Lognormal and moving window methods of estimating acid deposition, Journal of the American Statistical Association 85 (1990) 950–963.

[16] T. C. Haas, Kriging and automated variogram modeling within a moving window, Atmospheric Environment. Part A. General Topics 24 (1990) 1759–1769.

[17] J. M. V. Hoef, N. Cressie, R. P. Barry, Flexible spatial models for kriging and cokriging using moving averages and the fast fourier transform (fft), Journal of Computational and Graphical Statistics 13 (2004) 265–282.

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

[19] E. Anderes, M. Stein, Local likelihood estimation for nonstationary random fields, Journal of Multivariate Analysis (2011).

[20] M. J. Heaton, A. Datta, A. Finley, R. Furrer, R. Guhaniyogi, F. Gerber, R. B. Gramacy, D. Hammerling, M. Katzfuss, F. Lindgren, et al., Methods for analyzing large spatial data: A review and comparison, arXiv preprint arXiv:1710.05013 (2017).

[21] M. Katzfuss, A multi-resolution approximation for massive spatial datasets, Journal of the American Statistical Association 112 (2017) 201–214.

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

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

[24] M. Katzfuss, N. Cressie, Spatio-temporal smoothing and em estimation for massive remote-sensing data sets, Journal of Time Series Analysis 32 (2011) 430–446.

[25] D. Nychka, D. Hammerling, S. Sain, N. Lenssen, Latticekrig: Multiresolution kriging based on markov random fields, 2016. R package version 6.6.

[26] Douglas Nychka, Reinhard Furrer, John Paige, Stephan Sain, fields: Tools for spatial data, 2017. R package version 9.3.

[27] R. C. Team, R language definition, Vienna, Austria: R foundation for statistical computing (2000).

[28] H. Yu, Rmpi: Parallel statistical computing in r, R News 2 (2002) 10–14.

Figures

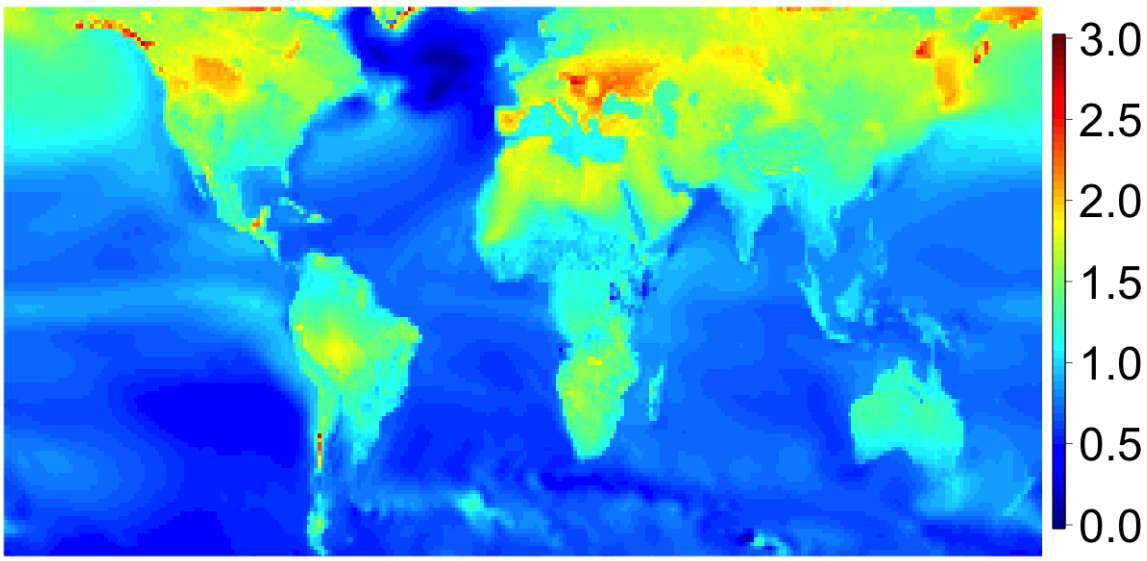

Figure B.1: Pattern scaling field for mean surface temperature (Centigrade) for the months of June, July and August (JJA). The field is an estimate of the local response to global warming. For example a value of 2.5 estimated at a particular grid box implies that that 1 degree change in global average JJA temperature will result in a change of 2.5 degrees at that location. This field is the sample mean of the 30 individual patterns found for the 30 ensemble members generated from the NCAR LENS. The simulation period 1920-2080 and uses the greenhouse gas scenario RCP8.5.

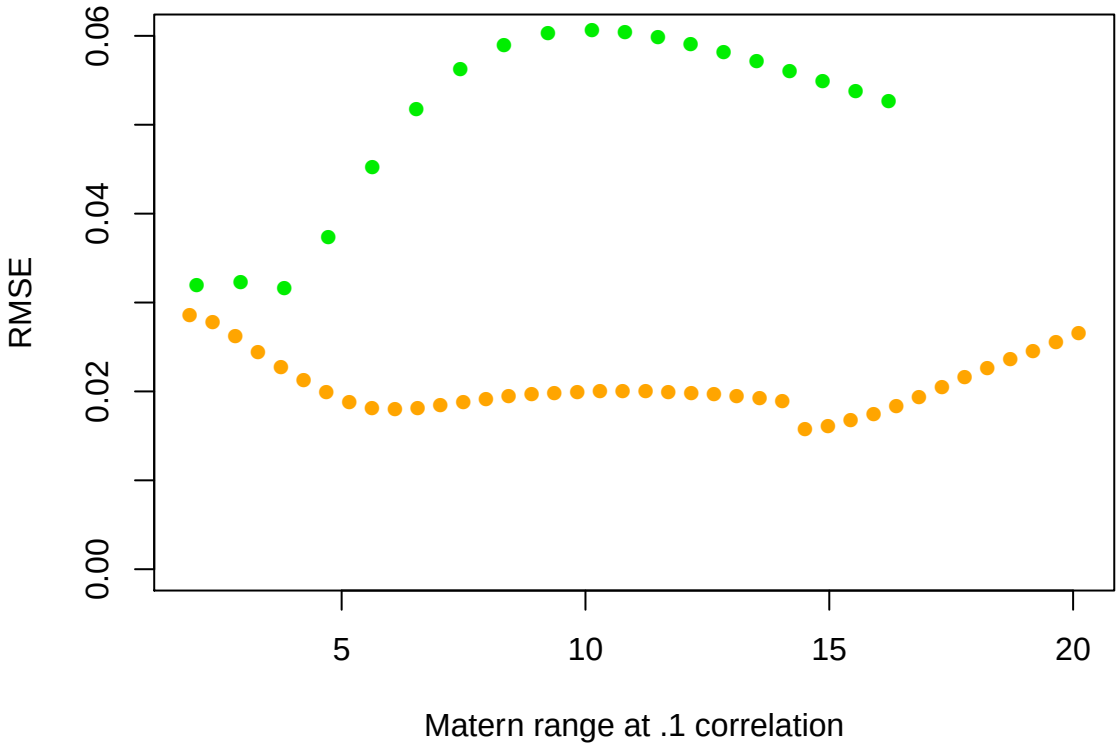

Figure B.2: Root mean squared error (RMSE) approximation error between weights for the Matérn covariances and the LatticeKrig model. The plotted points (green $\nu = 2.0$ and orange $\nu = 1.0$) are the RMSE values from minimizing the criterion in (10) over the LK parameters. The range parameter is scaled so that the correlation is .5 at a distance of the range. The break in the points is due to switching the leading resolution level for the LK model.

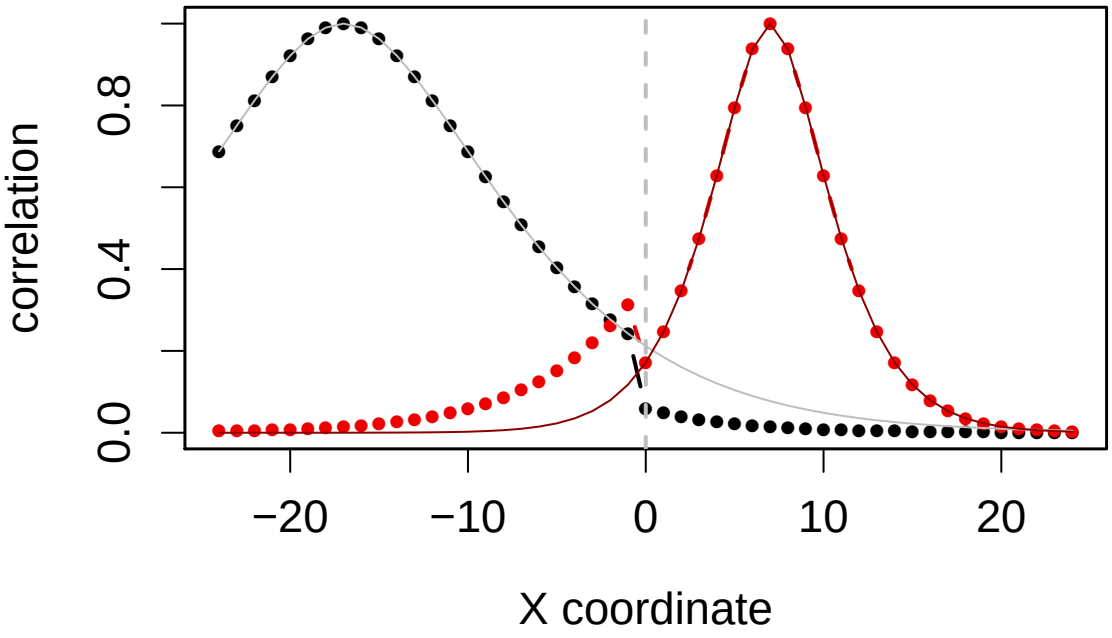

Figure B.3: Correlation curves for the data generating process illustrating the nonstationary first test case. Superimposed are the local stationary correlation functions. The spatial domain for this example is the square $[-24, 24] \times [-24, 24]$ but the correlation function is evaluated along the transect with the y-coordinate equal to zero. Plotted are the correlation functions for the location $(-17, 0)$ in black and $(7, 0)$ in red with points. The grey lines are the stationary correlation functions using the range parameter at these locations.

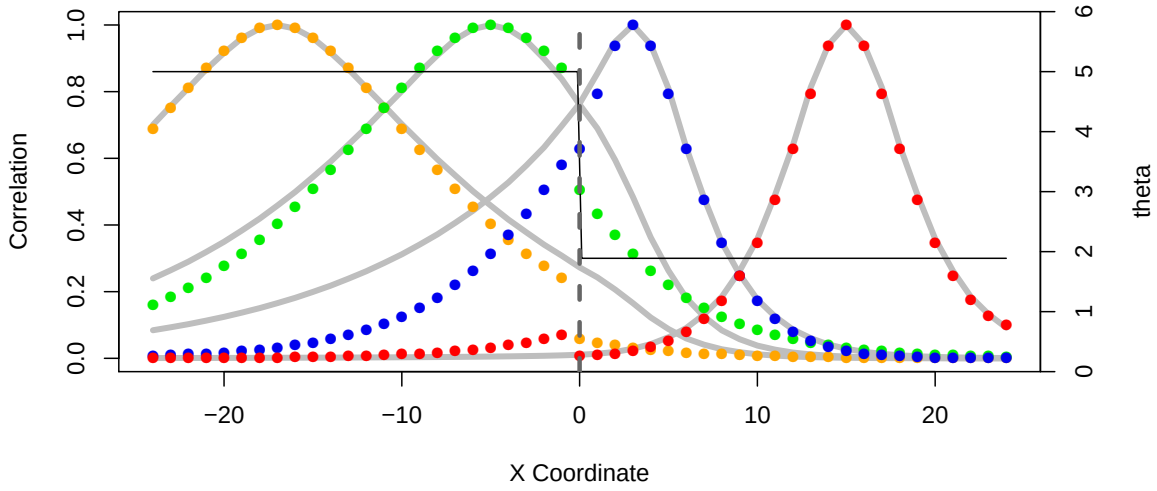

Figure B.4: Comparison of LatticeKrig approximation (grey lines) and true nonstationary correlations (points) for the first nonstationary test case, with a discontinuous range parameter. The superimposed black line gives the values for $\theta(\boldsymbol{s})$ as a function of the x-coordinate and corresponds to the axis on the right hand side of the plot. The correlation functions are with respect to the locations and colors: $(-17, 0)$ orange, $(-5, 0)$ green, $(3, 0)$ blue, and $(15, 0)$ red.

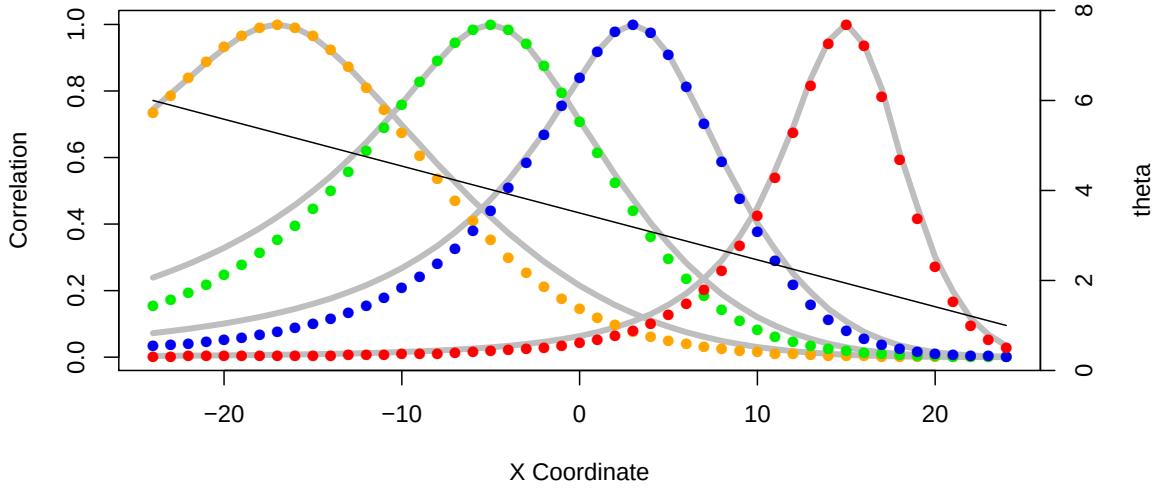

Figure B.5: Comparison of LatticeKrig approximation (grey lines) and true nonstationary correlations (points) for the second nonstationary test case, with a linearly varying range parameter. As in figure B.4 the black line indicates the value of the range parameter. The correlation functions are with respect to the locations and colors: $(-17, 0)$ orange, $(-5, 0)$ green, $(3, 0)$ blue, and $(15, 0)$ red.

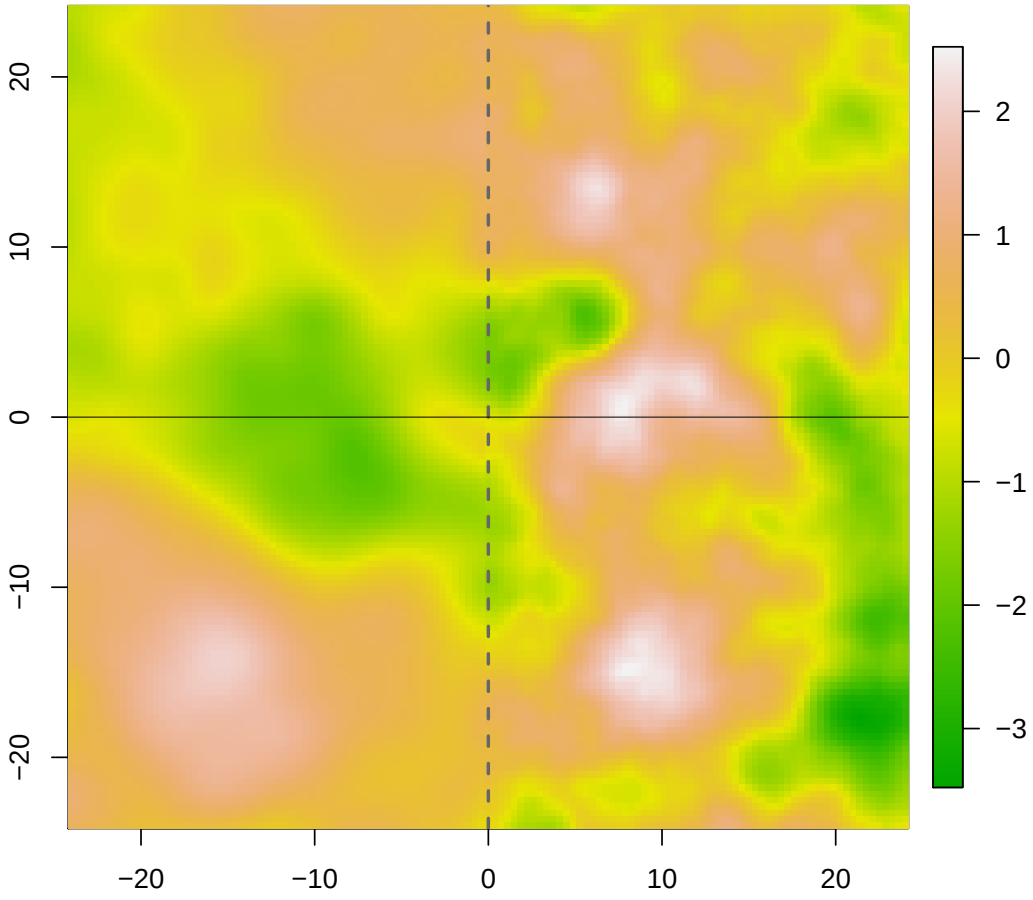

Figure B.6: Simulated field from the LatticeKrig approximation from the first nonstationary case. The vertical line is where the range parameter changes from 5 to 1.9. One can discern smaller scale structure in this field on the right side and larger scale features on the left side. For reference, the horizontal line is the transect used to evaluate the correlation functions in the previous figures.

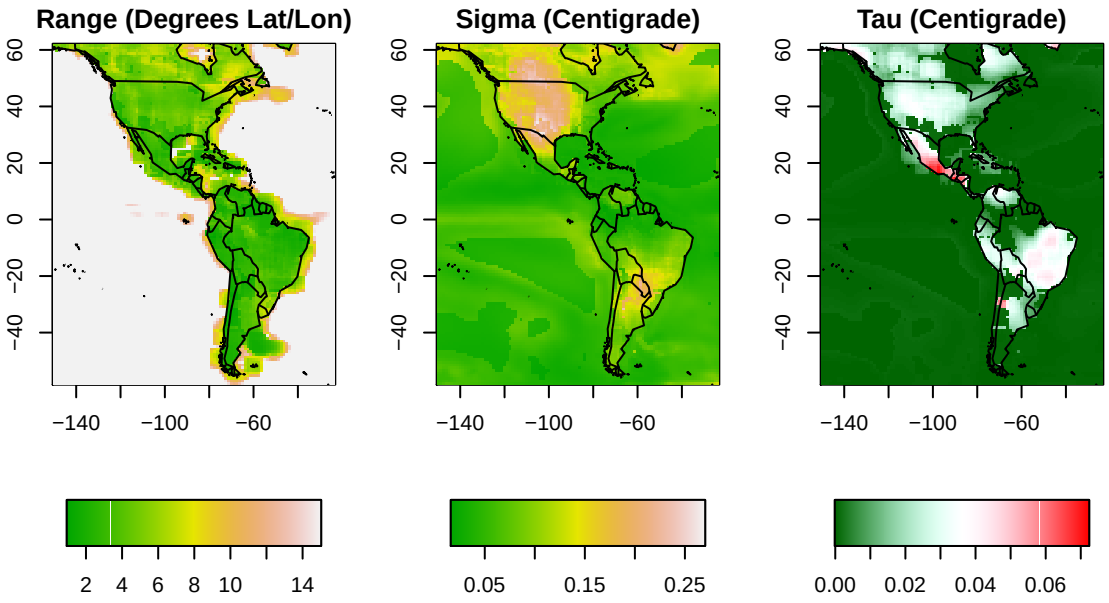

Figure B.7: Matérn parameter fields based on a 11×11 grid box moving window. At the equator this window width is 13.75 degrees or 1526 km. Parameters are found by maximum likelihood in these local windows, however, the σ and θ fields have been truncated for large values over the ocean.

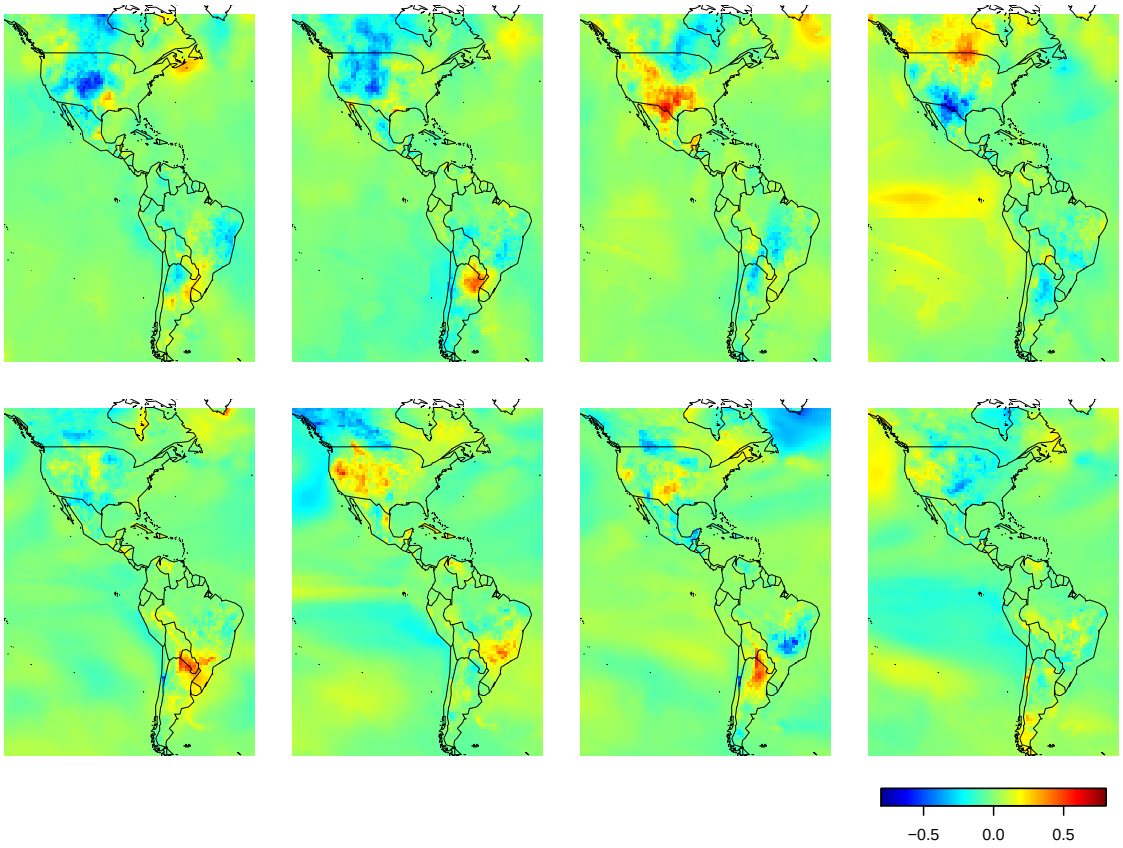

Figure B.8: Simulated and true fields for the pattern scaling data set. The top row shows four realizations from the LK Gaussian process model, and the bottom plots are the first four data fields based on the climate model output.

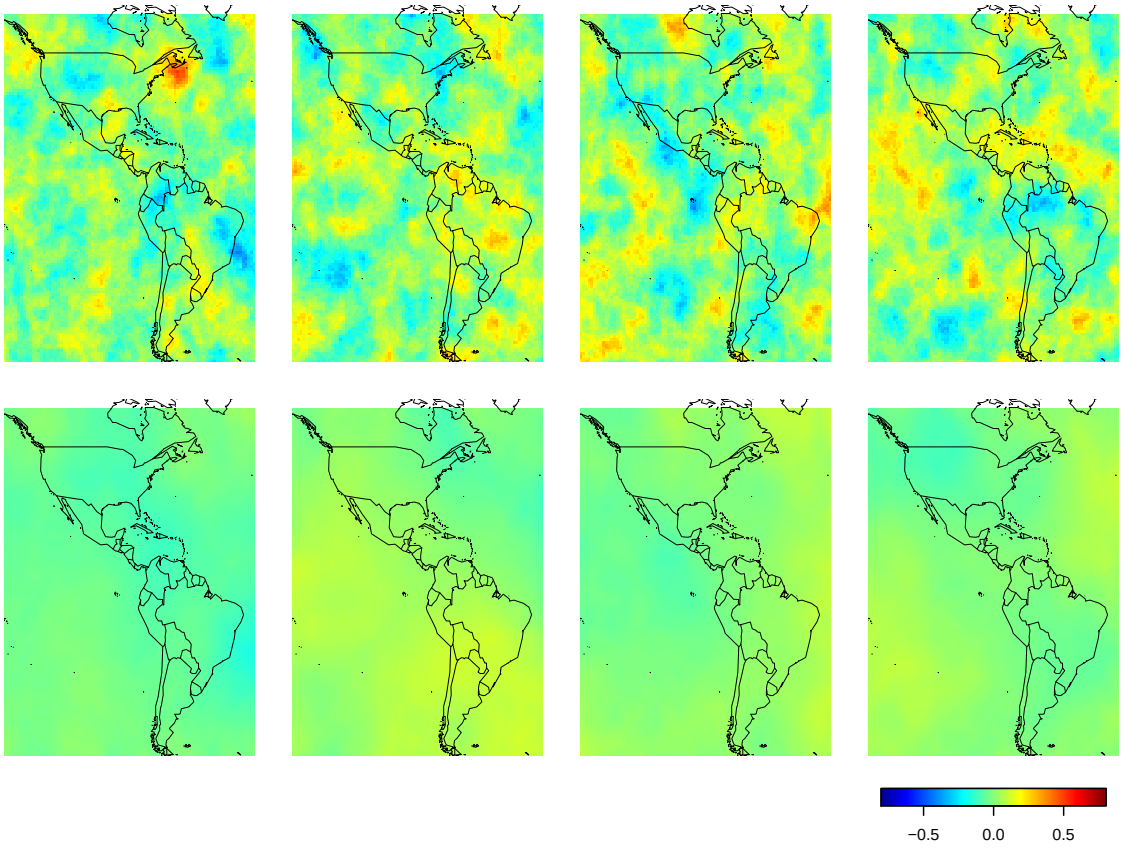

Figure B.9: Simulated stationary fields following the pattern scaling data set. The top row shows four realizations from the LK Gaussian process model using the median covariance parameters over land. The bottom row shows the corresponding realizations using median parameters over ocean.

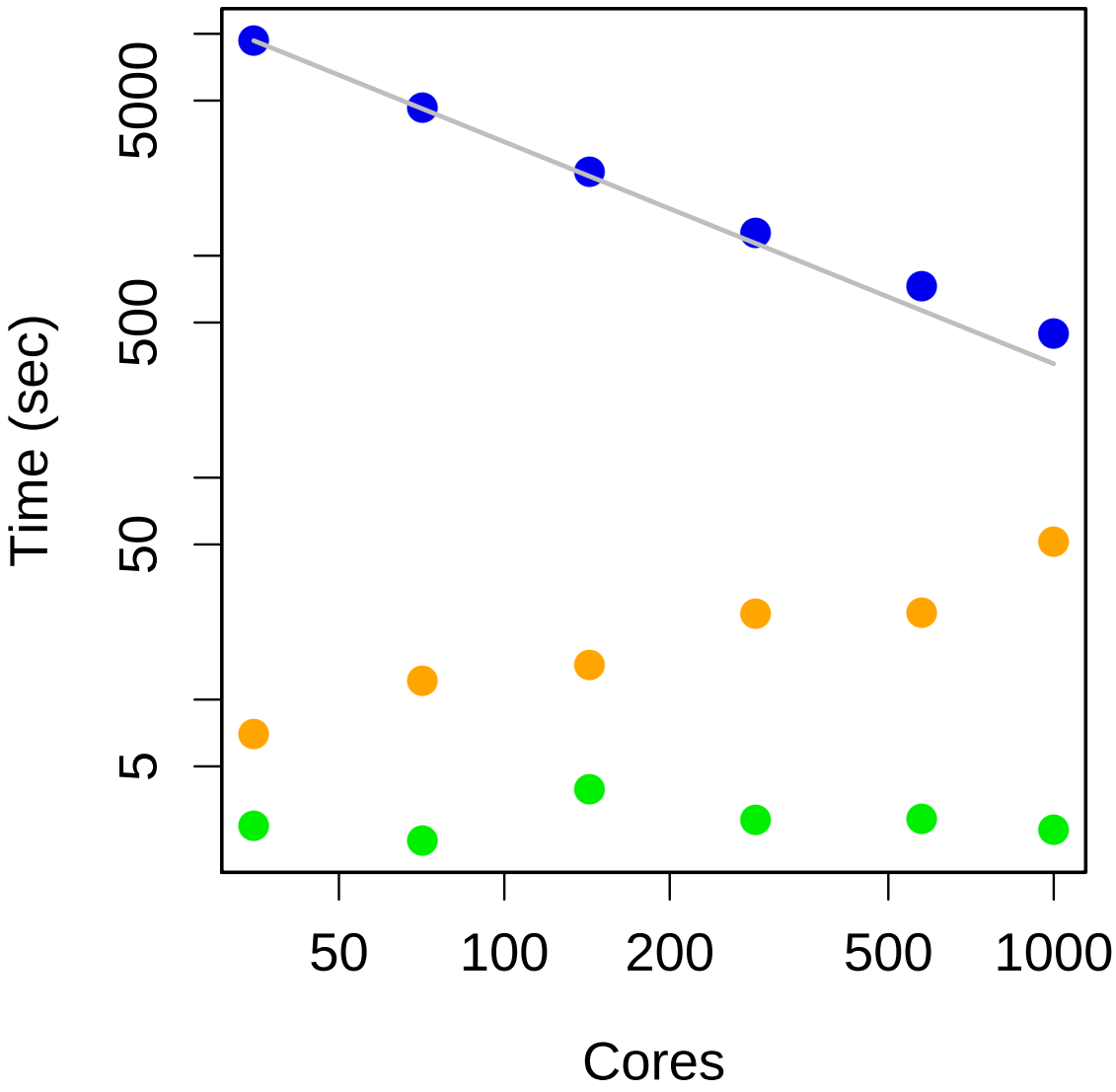

Figure B.10: Timing results for fitting local stationary covariances to 1000 grid boxes as a function of the cores. In this case the number of cores is equal to the number of worker R sessions. The parallel sessions were managed by the `Rmpi` package and done on the Cheyenne supercomputer managed by NCAR. Blue is the time fitting the model, green is the time broadcasting data to workers, and orange is the time spawning workers.