Published in final edited form as:
Environmetrics. 2006 ; 17(5): 483–506. doi:10.1002/env.785.

Spatial Modelling Using a New Class of Nonstationary Covariance Functions

Christopher J. Paciorek and **Mark J. Schervish**
[1]Christopher Paciorek is Assistant Professor, Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115 (E-mail: paciorek@alumni.cmu.edu). Mark Schervish is Professor, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213 (E-mail: mark@stat.cmu.edu).

Abstract

We introduce a new class of nonstationary covariance functions for spatial modelling. Nonstationary covariance functions allow the model to adapt to spatial surfaces whose variability changes with location. The class includes a nonstationary version of the Matérn stationary covariance, in which the differentiability of the spatial surface is controlled by a parameter, freeing one from fixing the differentiability in advance. The class allows one to knit together local covariance parameters into a valid global nonstationary covariance, regardless of how the local covariance structure is estimated. We employ this new nonstationary covariance in a fully Bayesian model in which the unknown spatial process has a Gaussian process (GP) prior distribution with a nonstationary covariance function from the class. We model the nonstationary structure in a computationally efficient way that creates nearly stationary local behavior and for which stationarity is a special case. We also suggest non-Bayesian approaches to nonstationary kriging.

To assess the method, we use real climate data to compare the Bayesian nonstationary GP model with a Bayesian stationary GP model, various standard spatial smoothing approaches, and nonstationary models that can adapt to function heterogeneity. The GP models outperform the competitors, but while the nonstationary GP gives qualitatively more sensible results, it shows little advantage over the stationary GP on held-out data, illustrating the difficulty in fitting complicated spatial data.

Keywords

smoothing; Gaussian process; kriging; kernel convolution

1 Introduction

One focus of spatial statistics research has been spatial smoothing - estimating a smooth spatial process from noisy observations or smoothing over small-scale variability. Statisticians have been interested in constructing smoothed maps and predicting at locations for which no data are available. Two of the most prominent approaches have been kriging and thin plate splines (see Cressie (1993, chap. 3) for a review). A simple Bayesian version of kriging for Gaussian data can be specified as

$$\begin{aligned} Y_i &\sim \mathrm{N}(f(x_i), \eta^2);\ i=1,\ldots,n \\ f(\cdot) &\sim \mathrm{GP}(\mu, C(\cdot,\cdot;\theta)), \end{aligned} \tag{1}$$

where η^2 is the noise variance (the nugget), μ is a scalar mean, $\boldsymbol{x_i}$ is a spatial location, and $f(\cdot)$ is the unknown spatial process with a Gaussian process (GP) prior distribution. The covariance function, $C(\cdot, \cdot; \boldsymbol{\theta})$, of the GP is parameterized by $\boldsymbol{\theta}$ and determines the covariance between any two locations. This model underlies the standard kriging approach, in which $C(\cdot; \boldsymbol{\theta})$ is a stationary covariance, a function only of Euclidean distance (or possibly a more general Mahalanobis distance) between any two locations. The low-dimensional $\boldsymbol{\theta}$ is generally estimated using variogram techniques (Cressie 1993, chap. 2) or by maximum likelihood (Smith 2001, p. 66) after integrating the unknown process values at the observation locations out of the model. The spatial process estimate is the posterior mean conditional on estimates of μ, η, and $\boldsymbol{\theta}$. While various approaches to kriging and thin plate spline models have been used successfully for spatial process estimation, they have the weakness of being global models, in which the variability of the estimated process is the same throughout the domain because $\boldsymbol{\theta}$ applies to the entire domain.

This failure to adapt to variability, or heterogeneity, in the unknown process is of particular importance in environmental, geophysical, and other spatial datasets, in which domain knowledge suggests that the function may be nonstationary. For example, in mountainous regions, environmental variables are likely to vary much more quickly than in flat regions. Spatial statistics researchers have made some progress in defining nonstationary covariance structures; in particular, this work builds on Higdon, Swall, and Kern (1999), who convolve spatially-varying kernels to give a nonstationary version of the squared exponential stationary covariance. Fuentes and Smith (2001) and Fuentes (2001) have an alternative kernel approach in which the unknown process is taken to be the convolution of a fixed kernel over independent stationary processes with different covariance parameters; Barber and Fuentes (2004) give a discretized mixture version of the model. Wood, Jiang, and Tanner (2002) estimate the spatial process as a mixture of thin plate splines to achieve nonstationarity, while Kim, Mallick, and Holmes (2005) use mixtures of Gaussian processes defined locally on a tesselation. The spatial deformation approach attempts to retain the simplicity of the stationary covariance structure by mapping the original input space to a new space in which stationarity can be assumed (Sampson and Guttorp 1992; Damian, Sampson, and Guttorp 2001; Schmidt and O'Hagan 2003). Research on the deformation approach has focused on multiple noisy replicates of the spatial function rather than the setting of one set of observations on which we focus here.

Many nonparametric regression methods are also applicable to spatial data, but spatial modelling requires flexible two-dimensional surfaces, while many nonparametric regression techniques focus on additive models, summing one-dimensional curves. In particular, while Bayesian free-knot spline models, in which the number and location of the knots are part of the estimation problem, have been very successful in one dimension (DiMatteo, Genovese, and Kass 2001), effectively extending splines to higher dimensions is more difficult. Using different bases, Denison, Mallick, and Smith (1998) and Holmes and Mallick (2001) fit free-knot spline models for two and higher dimensions using reversible-jump MCMC. Lang and Brezger (2004) and Crainiceanu, Ruppert, and Carroll (2004) use penalized splines with spatially-varying penalties in two dimensions. While not commonly used for spatial data, neural network models can adapt to function heterogeneity (Neal 1996). Tresp (2001) and Rasmussen and Ghahramani (2002) use mixtures of stationary GPs; they show success in one dimension, but do not provide results in higher dimensions nor compare their model to other methods.

In this work, we extend the Higdon et al. (1999) nonstationary covariance function to create a class of closed-form nonstationary covariance functions, including a nonstationary Matérn covariance, parameterized by spatially-varying covariance parameters (Section 2). We demonstrate how this covariance can be used in an ad hoc nonstationary kriging approach

(Section 3.1) and in a fully Bayesian GP spatial model (Section 3.2). We compare the performance of the nonstationary GP model to alternative models on real climatological data (Section 4). We conclude by suggesting strategies for improving computational efficiency and discussing the use of the nonstationary covariance in more complicated models (Section 5).

2 A new class of nonstationary covariance functions

In this section we extend the nonstationary covariance function of Higdon et al. (1999), providing a general class of closed-form nonstationary covariance functions built upon familiar stationary covariance functions. The approach constructs a global covariance function by knitting together local covariance structures and is valid regardless of how the local covariance parameters are estimated.

2.1 Review of stationary covariance functions

The covariance function is crucial in GP modelling; it controls how the observations are weighted for spatial prediction. Recent work in spatial statistics has focused on the Matérn covariance, whose stationary, isotropic form is

$$C(\tau)=\sigma^2\frac{1}{\Gamma(\nu)2^{\nu-1}}\left(2\sqrt{\nu}\frac{\tau}{\rho}\right)^{\nu}\mathcal{K}_{\nu}\left(2\sqrt{\nu}\frac{\tau}{\rho}\right),\ \rho>0;\ \nu>0$$

where τ is distance, ρ is the spatial range parameter, and $\mathcal{K}_{\nu}(\cdot)$ is the modified Bessel function of the second kind, whose order is the differentiability parameter, ν. The behavior of the covariance function of a stochastic process near the origin determines the smoothness properties of process realizations (Abrahamsen 1997; Stein 1999; Paciorek 2003, chap. 2). The Matérn form has the desirable property that Gaussian processes with this covariance have realizations (sample paths) that are $\lceil \nu - 1 \rceil$ times differentiable. As $\nu \to \infty$, the Matérn covariance function approaches the squared exponential (also called the Gaussian) form, popular in machine learning, whose realizations are infinitely differentiable. For $\nu = 0.5$, the Matérn takes the exponential form, popular in spatial statistics, which produces continuous but non-differentiable realizations; this seems insufficiently smooth for many applications. While it is not clear that process differentiability can be estimated from data, having the additional parameter, ν, allows one to choose from a wider range of realization behavior than the extremes of the exponential and squared exponential covariances provide, or, if estimated, allows for additional flexibility in spatial modelling. For example, the smoothing matrices produced by the exponential and Matérn ($\nu = 4$) correlation functions are rather different, as are realizations.

Stationary, isotropic covariance functions can be easily generalized to anisotropic covariance functions that account for directionality by using the Mahalanobis distance between locations,

$$\tau(\boldsymbol{x}_i,\boldsymbol{x}_j)=\sqrt{(\boldsymbol{x}_i-\boldsymbol{x}_j)^T\Sigma^{-1}(\boldsymbol{x}_i-\boldsymbol{x}_j)}, \quad (2)$$

where Σ is an arbitrary positive definite matrix, rather than $\Sigma = I$, which gives Euclidean distance and isotropy. The nonstationary covariance function we introduce next builds on this more general anisotropic form.

2.2 From stationarity to nonstationarity via kernel convolution

Higdon et al. (1999) introduced a nonstationary covariance function, $C^{NS}(\boldsymbol{x_i}, \boldsymbol{x_j}) = \int_{\Re^2} K_{\boldsymbol{x_i}}(\boldsymbol{u}) K_{\boldsymbol{x_j}}(\boldsymbol{u}) d\boldsymbol{u}$, obtained by convolving spatially-varying kernel functions, $K_{\boldsymbol{x}}(\boldsymbol{u})$. Here, $\boldsymbol{x_i}$, $\boldsymbol{x_j}$, and $\boldsymbol{u}$ are locations in $\Re^2$, and the kernel functions are functions such as those used in kernel density estimation. Higdon et al. (1999) motivate this construction as the covariance function of a process, $f(\cdot)$,

$$f(\boldsymbol{x}) = \int_{\Re^2} K_{\boldsymbol{x}}(\boldsymbol{u}) \psi(\boldsymbol{u}) d\boldsymbol{u}, \tag{3}$$

produced by convolving a white noise process, $\psi(\cdot)$, with the spatially-varying kernel function. One can avoid the technical details involved in carefully defining such a white noise process by using the definition of positive definiteness to show directly that the covariance function is positive definite in every Euclidean space, $\Re^p$, $p = 1, 2, \ldots$:

$$\begin{aligned} \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j C^{NS}(\boldsymbol{x_i}, \boldsymbol{x_j}) &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \int_{\Re^p} K_{\boldsymbol{x_i}}(\boldsymbol{u}) K_{\boldsymbol{x_j}}(\boldsymbol{u}) d\boldsymbol{u} \\ &= \int_{\Re^p} \sum_{i=1}^{n} a_i K_{\boldsymbol{x_i}}(\boldsymbol{u}) \sum_{j=1}^{n} a_j K_{\boldsymbol{x_j}}(\boldsymbol{u}) d\boldsymbol{u} \\ &= \int_{\Re^p} \left(\sum_{i=1}^{n} a_i K_{\boldsymbol{x_i}}(\boldsymbol{u}) \right)^2 d\boldsymbol{u} \geq 0. \end{aligned} \tag{4}$$

Note that the kernel is an arbitrary function; positive definiteness is achieved because the kernel at a location provides all the information about how the location affects the pairwise correlations involving that location. However, standard kernel functions are density functions or non-negative functions that decay monotonically from their mode. For Gaussian kernels (taking $K_{\boldsymbol{x}}(\cdot)$ to be a (multivariate) Gaussian density centered at $\boldsymbol{x}$), one can show using convolution (see Appendix 1) that the nonstationary covariance function takes the simple form,

$$C^{NS}(\boldsymbol{x_i}, \boldsymbol{x_j}) = \sigma^2 |\Sigma_i|^{\frac{1}{4}} |\Sigma_j|^{\frac{1}{4}} \left| \frac{\Sigma_i + \Sigma_j}{2} \right|^{-\frac{1}{2}} \exp(-Q_{ij}), \tag{5}$$

with quadratic form

$$Q_{ij} = (\boldsymbol{x_i} - \boldsymbol{x_j})^T \left(\frac{\Sigma_i + \Sigma_j}{2} \right)^{-1} (\boldsymbol{x_i} - \boldsymbol{x_j}), \tag{6}$$

where $\Sigma_i = \Sigma(\boldsymbol{x_i})$, which we call the kernel matrix, is the covariance matrix of the Gaussian kernel centered at $\boldsymbol{x_i}$. (5) has the form of an anisotropic squared exponential covariance function, but in place of a spatially constant matrix, Σ, in the quadratic form (2), which gives stationarity and anisotropy, we average the kernel matrices for the two locations (6). Gibbs (1997) derived a special case of (5) in which the kernel matrices are diagonal. The evolution of the kernel matrices in the domain produces nonstationary covariance, with kernels with small variances, and therefore little overlap with kernels at other locations, producing locally short correlation scales. Unfortunately, so long as the kernel matrices vary smoothly in the

input space, realizations from GPs with the covariance (5) are infinitely differentiable (Paciorek 2003, chap. 2), just as for the stationary squared exponential. Stein (1999) discusses in detail why such highly smooth realizations are undesirable and presents an asymptotic argument for using covariance functions in which the smoothness is allowed to vary.

2.3 Generalizing the kernel convolution form

To create a more general form than the squared exponential, we construct a class of nonstationary covariance functions, substituting $\sqrt{Q_{ij}}$ in place of τ/ρ in stationary correlation functions. Unfortunately, since $\sqrt{Q_{ij}}$ is not a distance metric, it violates the triangle inequality, so this cannot be done arbitrarily, but it can be done for a class of stationary correlation functions; the proof of the theorem (which is stated but not proven in Paciorek and Schervish (2004)) is given in Appendix 2.

Theorem 1—*If an isotropic correlation function, $R^S(\tau)$, is positive definite on $\Re^p$ for every $p = 1, 2, \ldots$, then the function, $R^{NS}(\cdot, \cdot)$, defined by*

$$R^{NS}(\boldsymbol{x_i}, \boldsymbol{x_j}) = |\Sigma_i|^{\frac{1}{4}} \left|\Sigma_j\right|^{\frac{1}{4}} \left|\frac{\Sigma_i + \Sigma_j}{2}\right|^{-\frac{1}{2}} R^S\left(\sqrt{Q_{ij}}\right) \tag{7}$$

with $\sqrt{Q_{ij}}$ used in place of τ, is a nonstationary correlation function, positive definite on $\Re^p$, $p = 1, 2, \ldots$.

The result applies to correlation functions that are positive definite in Euclidean space of every dimension, in particular the power exponential, rational quadratic, and Matérn correlation functions. Nonstationary covariance functions can be constructed merely by multiplying by a variance parameter, σ^2.

Under conditions that ensure that the elements of the kernel matrices vary smoothly over the domain, the mean square and sample path differentiability of Gaussian processes parameterized by correlation functions of the form (7) follow from the differentiability properties of Gaussian processes parameterized by the underlying stationary correlation function (Paciorek 2003, chap. 2). The precise statement of the theorems and proofs behind this result are involved and not the focus of this paper. However, the result is intuitive and best made clear as follows. If the elements of the kernel matrices vary smoothly (see Section 3.2.1 for such a construction), then in a small neighborhood of $\boldsymbol{x_0}$, the kernel matrices are essentially constant, $\Sigma(\boldsymbol{x}) \approx \Sigma(\boldsymbol{x_0})$, so the resulting local correlation structure is essentially stationary. Hence the differentiability properties, which depend on the behavior of the correlation near the origin, are the same as those for the underlying stationary correlation.

The new class of nonstationary covariance functions includes a nonstationary version of the Matérn covariance function,

$$C^{NS}(\boldsymbol{x_i}, \boldsymbol{x_j}) = \sigma^2 \frac{1}{\Gamma(\nu) 2^{\nu-1}} |\Sigma_i|^{\frac{1}{4}} \left|\Sigma_j\right|^{\frac{1}{4}} \left|\frac{\Sigma_i + \Sigma_j}{2}\right|^{-\frac{1}{2}} \left(2\sqrt{\nu Q_{ij}}\right)^{\nu} \mathcal{K}_{\nu}\left(2\sqrt{\nu Q_{ij}}\right), \tag{8}$$

which includes a nonstationary version of the exponential covariance function as the special case when $\nu = 0.5$. As with the stationary form, the sample path differentiability of Gaussian processes with nonstationary Matérn covariance increases with ν. Another form is the rational quadratic covariance function,

$$C^{NS}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sigma^2 |\Sigma_i|^{\frac{1}{4}} |\Sigma_j|^{\frac{1}{4}} \left| \frac{\Sigma_i + \Sigma_j}{2} \right|^{-\frac{1}{2}} \left(\frac{1}{1+Q_{ij}} \right)^{\nu},$$

a covariance function with a long tail. In the remainder of the paper, we focus on the nonstationary Matérn form.

3 Implementation of the new nonstationary covariance

3.1 Ad hoc nonstationary kriging

The kernel convolution nonstationary covariance provides a way to construct closed-form covariance functions based on stationary correlation functions and local covariance parameters. One advantage of the approach is that construction using arbitrary local parameters is positive definite. The nonstationary covariance structure is highly parameterized relative to a stationary covariance structure, so optimization is difficult and runs the danger of overfitting. In this section, we propose to use the nonstationary correlation (7) in a kriging framework, estimating the covariance structure and then fitting the surface conditional on that structure.

3.1.1 Estimating the nonstationary covariance structure—When distinct regions are present, one can piece together stationary regional covariances. The parameters of regional anisotropic covariances could be estimated using either a variogram fitting approach or marginal likelihood maximization in which only the data in a region are used to estimate the region-specific parameters. Then, to knit together a full covariance for the entire domain, set $\Sigma_i = \Sigma_{A(i)}$, where $A(i)$ denotes the region in which location i falls and $\Sigma_{A(i)}$ is constructed for each region from the parameters, $\boldsymbol{\theta}_{A(i)}$, of the anisotropic correlation structure estimated for the region. One can use different values for σ, μ, and η for the different regions. One could also use different values of ν, following Stein (2005), who has extended our approach to allow for spatially-varying ν. In the next section, we illustrate this approach in Colorado, splitting the state into two regions of differing topography.

Another possibility is to estimate the kernel matrices at the locations of interest in a weighted or moving window fashion. Recall that if the elements of the kernel matrices vary slowly, the nonstationary covariance is locally a nearly stationary anisotropic covariance. In a small neighborhood, for $\boldsymbol{x_j}$ near $\boldsymbol{x_i}$, $\Sigma_j \approx \Sigma_i$; our task is to estimate the parameters, $\boldsymbol{\theta_i}$, of an anisotropic covariance, from which the kernel matrix, Σ_i, will be constructed. To estimate $\boldsymbol{\theta_i}$ based on the variogram, we could use a moving window to include in the empirical variogram only pairs of locations for which either one or both of the locations are near $\boldsymbol{x_i}$. One could also assign weights to each pair of locations and estimate $\boldsymbol{\theta_i}$ based on weighted variograms. Such a fitting approach is similar to that of Barber and Fuentes (2004), who fit local variograms to time-replicated data. To estimate $\boldsymbol{\theta_i}$ using the marginal likelihood approach, one could use the marginal likelihood only for observations from locations near $\boldsymbol{x_i}$. Doing this for each location allows us to create spatially-varying kernel matrices, Σ_i, $i = 1, \ldots$.

Finally, one might parameterize the nonstationary covariance as a function of key covariates to reduce the dimension of the estimation problem. For example, in the Colorado

precipitation example in the next section, one might specify the kernels as $\Sigma_i = \rho_i I$, with the spatial ranges specified as a simple parameterized function of local elevation heterogeneity.

The spatial process at the observation locations, $\boldsymbol{f}$, and at the prediction locations, $\boldsymbol{f}^*$, can then be estimated using the mean and variance conditional on the nonstationary covariance structure constructed from the estimated kernel matrices, using standard multivariate Gaussian conditional calculations,

$$\begin{aligned} \boldsymbol{f} \mid \boldsymbol{Y}, \mu, \eta, \Sigma(\cdot) \sim & \quad \mathrm{N}\left(\mu\mathbf{1} + C_{\boldsymbol{f}}(C_{\boldsymbol{f}} + C_{\boldsymbol{Y}})^{-1}(\boldsymbol{Y} - \mu\mathbf{1}), C_{\boldsymbol{f}}(C_{\boldsymbol{f}} + C_{\boldsymbol{Y}})^{-1}C_{\boldsymbol{Y}}\right) \\ \boldsymbol{f}^* \mid \boldsymbol{Y}, \mu, \eta, \Sigma(\cdot) \sim & \quad \mathrm{N}\left(\mu\mathbf{1} + C_{\boldsymbol{f}^*\boldsymbol{f}}(C_{\boldsymbol{f}} + C_{\boldsymbol{Y}})^{-1}(\boldsymbol{Y} - \mu\mathbf{1}), C_{\boldsymbol{f}^*\boldsymbol{f}^*} - C_{\boldsymbol{f}^*\boldsymbol{f}}(C_{\boldsymbol{f}} + C_{\boldsymbol{Y}})^{-1}C_{\boldsymbol{f}^*\boldsymbol{f}}^T\right), \end{aligned} \tag{9}$$

where $C_{\boldsymbol{f}}$ is the covariance matrix of the process at the observation locations, $C_{\boldsymbol{f}^*\boldsymbol{f}}$ is the covariance matrix of the process values at the prediction locations and those at the observation locations, $C_{\boldsymbol{f}^*\boldsymbol{f}^*}$ is the covariance matrix of the process at the prediction locations, and $C_{\boldsymbol{Y}} = \eta^2 I$. Note that for prediction, we need to estimate kernel matrices for the prediction locations using covariance information based on nearby observation locations.

3.1.2 Illustration—Climatological data for the state of Colorado in the United States provide a nice illustration of a simple application of nonstationary kriging. To first approximation, Colorado is divided into a mountainous western portion, west of Denver and the I-25 highway corridor, and a flat plains portion in the east (Fig. 1a). The Geophysical Statistics Project at the National Center for Atmospheric Research (http://www.cgd.ucar.edu/stats/Data/US.monthly.met) has posted a useful subset of the United States climate record over the past century from a large network of weather stations. Areas of complex topography are a particular challenge for making predictions off the observation network, and even for the continental U.S. the observation record is quite sparse relative to the resolution needed for understanding impacts of changing climate. For this illustration, we take the log-transformed annual precipitation in Colorado for 1981 (Fig. 1b), the year for which the most stations without any missing monthly values (217) are available and compare kriging with model (1) using both stationary and nonstationary covariance functions.

For stationary kriging, we use an anisotropic Matérn covariance function based on Mahalanobis distance (2), a function of $\Sigma = \Gamma\Lambda\Gamma^T$, with Γ an eigenvector matrix parameterized by an angle of rotation, ψ, (a Givens rotation matrix, see Anderson, Olkin, and Underhill (1987)) and Λ a diagonal eigenvalue matrix with eigenvalues (squared spatial range parameters), ρ_1^2 and ρ_2^2. After integrating the spatial function values at the observation locations out of the model, we estimate the parameters, $\{\mu, \eta, \sigma, \rho_1, \rho_2, \psi\}$, by maximizing the resulting marginal likelihood using the nlm() function in R. We fix $\nu = 4$.

For nonstationary kriging, we fit separate anisotropic Matérn covariance structures in the eastern and western regions, split at longitude 104.873° W, by maximizing the marginal likelihoods for each region with respect to the parameters. We again fix $\nu = 4$. We construct the Matérn nonstationary covariance structure (8) for the entire dataset by setting $\Sigma_i = \Sigma(\boldsymbol{\theta}_{A(i)})$, where $\boldsymbol{\theta}_{A(i)} = \{\rho_{1,A(i)}, \rho_{2,A(i)}, \psi_{A(i)}\}$.

Table 1 shows the parameter estimates for the two models. As expected for the nonstationary model, because of the topographical variability in western Colorado, the correlation ranges are much smaller than in eastern Colorado, and the estimate of the function variability, σ, is larger. The similarity of the western estimates with the statewide

stationary estimates suggests that the stationary estimates are driven by the more variable (and more dense) western data. The surface estimate from the nonstationary model shows more heterogeneity in the west than the east (Fig. 2). The standard deviations of the estimated surface for the nonstationary model suggest much more certainty about the surface values in the east, as expected (Fig. 2c,d). In the west, the complexity of the surface results in high levels of uncertainty away from the observation locations, as is the case throughout the state in the stationary approach. Both approaches estimate 137 degrees of freedom for the surface (11), but the nonstationary kriging model has a much higher log likelihood (181 compared to 143), as expected. The nonstationary model is preferred based on the AIC_C model selection criteria (Hurvich, Simonoff, and Tsai 1998), designed for smoothing parameter selection in nonparametric regression models in which the smoother is linear (as is the case here).

One drawback to having sharply delineated regions is seen in the sharp changes at the border between the regions (Fig. 2b,d). This occurs because the kernels change sharply at the border; see Gibbs (1997) and Paciorek (2003, sec. 2.2) for discussion of this effect). One simple way to remove this discontinuity would be to smooth the covariance parameters, and therefore the resulting kernel matrices, in the vicinity of the boundary. A more principled approach is a fully Bayesian model, in which the kernel matrices are constrained to vary smoothly, minimizing the effect, as seen in Section 4.4.1.

3.2 A hierarchical Bayesian model for spatial smoothing

The nonstationary kriging approach suffers from several drawbacks. First, the ad hoc estimation of the covariance structure depends on how one estimates the local covariance parameters; in the illustration this involved how to split Colorado into regions. Second, as for kriging in general, uncertainty in the covariance structure is not accounted for; in the case of nonstationary covariance with its more flexible form and larger number of parameters, this uncertainty is of much more concern than in the stationary case. To address these concerns, we construct a fully Bayesian model, with prior distributions on the kernel matrices that determine the nonstationary covariance structure.

3.2.1 Basic Bayesian model—The Bayesian model starts with the basic kriging setup (1) and sets $C(\cdot,\cdot;\theta)=C_f^{NS}(\cdot,\cdot;\Sigma(\cdot),\nu_f)$, where C_f^{NS} is the nonstationary Matérn covariance function (8) constructed from $\Sigma(\cdot)$, the kernel matrix process, described below. For the differentiability parameter, we use the prior, $\nu_f \sim U(0.5, 30)$, which produces realizations that vary between non-differentiable ($\nu_f = 0.5$) and highly differentiable. We use proper, but diffuse, priors for μ_f, σ_f^2, and η^2, and bound σ_f^2 based on the range of the observation values. The main challenge is to parameterize the kernels, since their evolution over the domain determines how quickly the covariance structure changes over the domain and therefore the degree to which the model adapts to heterogeneity in the unknown function. In many problems, it seems natural that the covariance structure would evolve smoothly, as parameterized below.

The kernel matrix process, $\Sigma(\cdot)$, is parameterized as follows. Each location, $\boldsymbol{x_i}$, has a Gaussian kernel with mean, $\boldsymbol{x_i}$, and covariance (kernel) matrix, $\Sigma_i = \Sigma(\boldsymbol{x_i})$. Since there are (implicitly) kernel matrices at each location in the domain, we have a multivariate process, the matrix-valued function, $\Sigma(\cdot)$. First, construct an individual kernel matrix using the spectral decomposition, $\Sigma_i=\Gamma_i\Lambda_i\Gamma_i^T$ where Λ_i is a diagonal matrix of eigenvalues, $\lambda_1(\boldsymbol{x_i})$ and $\lambda_2(\boldsymbol{x_i})$, and Γ_i is an eigenvector matrix constructed as described below from $\gamma_1(\boldsymbol{x_i})$ and $\gamma_2(\boldsymbol{x_i})$. We construct $\Sigma(\cdot)$ over the entire space, ensuring that each $\Sigma(\boldsymbol{x_i})$ is positive definite, by creating spatial hyperprocesses, $\lambda_1(\cdot)$, $\lambda_2(\cdot)$, $\gamma_1(\cdot)$, and $\gamma_2(\cdot)$. We will refer to these as the

eigenvalue and eigenvector processes, and to them collectively as the eigenprocesses. Let $\varphi(\cdot) \in \{\log(\lambda_2(\cdot)), \gamma_1(\cdot), \gamma_2(\cdot)\}$ denote any one of these eigenprocesses; $\lambda_1(\cdot)$ is derived from $\gamma_1(\cdot)$ and $\gamma_2(\cdot)$. We take each $\varphi(\cdot)$ to have a stationary GP prior distribution with anisotropic Matérn correlation function (Section 3.2.2); this parameterization of the processes ensures that the kernel matrices vary smoothly in an elementwise fashion by forcing their eigenvalues and eigenvectors to vary smoothly. For simplicity we assume a priori independence between the eigenprocesses, relying on the data to infer posterior dependence. Parameterizing the eigenvectors of the kernel matrices using a spatial process of angles of rotation, with an angle at each location, is difficult because the angles have range $[0, 2\pi) \equiv S^1$, which is not compatible with the range of a GP. Instead, Γ_i is constructed from the eigenvector processes,

$$\Gamma_i = \begin{pmatrix} \frac{\gamma_1(\boldsymbol{x}_i)}{d_i} & \frac{-\gamma_2(\boldsymbol{x}_i)}{d_i} \\ \frac{\gamma_2(\boldsymbol{x}_i)}{d_i} & \frac{\gamma_1(\boldsymbol{x}_i)}{d_i} \end{pmatrix},$$

where $d_i = \sqrt{\gamma_1^2(\boldsymbol{x}_i) + \gamma_2^2(\boldsymbol{x}_i)}$. In turn, $\lambda_1(\boldsymbol{x}_i)$is taken to be d_i^2, the squared length of the eigenvector constructed from $\gamma_1(\boldsymbol{x}_i)$ and $\gamma_2(\boldsymbol{x}_i)$. Any given process realization, $f(\cdot)$, will be more variable in regions in which $\lambda_1(\cdot)$ and $\lambda_2(\cdot)$ are small, as this corresponds to having small local spatial ranges.

An alternative to the eigendecomposition parameterization is to represent the Gaussian kernels as ellipses of constant probability density, parameterized by the focus and size of the ellipse, and to have the focal coordinates and ellipse sizes vary smoothly over space (Higdon et al. 1999). However, Higdon et al. (1999) fixed the ellipse size at a constant value common to all locations, and Swall (1999, p. 94) found overfitting and mixing problems when the ellipse size was allowed to vary, although we also noticed slow mixing in our parameterization. Also, the eigendecomposition approach extends more readily to higher dimensions, which may be of interest for spatial data in three dimensions and more general nonparametric regression problems (Paciorek and Schervish 2004).

3.2.2 Representation of stationary GPs in the hierarchy—One can represent the stationary GPs, $\varphi(\cdot) \in \{\log(\lambda_2(\cdot)), \gamma_1(\cdot), \gamma_2(\cdot)\}$, used to construct the nonstationary covariance structure in a straightforward way, working with the Cholesky decompositions of the stationary covariance matrices for each of the processes (Paciorek and Schervish 2004), but the MCMC computations are slow. Instead, we represent each using a basis function approximation to a stationary GP (Kammann and Wand 2003). The vector of values of a spatial process, $\varphi(\cdot)$, at the observation locations, $\boldsymbol{\varphi} = \{\varphi(\boldsymbol{x}_1), \ldots, \varphi(\boldsymbol{x}_n)\}$, is a linear combination of basis functions,

$$\begin{aligned} \varphi &= \mu_\varphi \mathbf{1} + \sigma_\varphi Z_\varphi \boldsymbol{u}_\varphi, \\ Z_\varphi &= \Psi_\varphi \Omega_\varphi^{-\frac{1}{2}}. \end{aligned} \tag{10}$$

The basis matrix, Z_φ, is constructed using radial (i.e., isotropic) basis functions, where Ψ_φ contains the pairwise isotropic Matérn covariances, $C(\cdot; \rho_\varphi, \nu_\varphi)$, between the observation locations and pre-specified knot locations, $\boldsymbol{\kappa}_k$, $k = 1, \ldots, K$, while $\Omega_\varphi^{-1/2}$ is calculated by singular value decomposition from the matrix, Ω, which contains pairwise covariances amongst the knot locations. Knots at all the data locations would give an exact representation of a stationary GP with mean, μ_φ, standard deviation, σ_φ, spatial range, ρ_φ,

and differentiability parameter, ν_φ. Prediction for $\varphi(\cdot)$ uses Ψ^*_φ, which contains the pairwise covariances between the prediction locations and the knots. We use a relatively coarse 8 by 8 grid of $K = 64$ knots, because the eigenprocesses are in the hierarchy of the model and we do not expect them to be very variable.

To limit the number of parameters involved and improve mixing, we place some constraints on the hyperparameters of the stationary GPs, while still allowing the eigenprocesses to be flexible. These constraints are informed by the fact that the values of $\lambda_1(\cdot)$ and $\lambda_2(\cdot)$ play the role of locally-varying spatial range parameters so we know what values of the eigenprocesses and their hyperparameters are plausible and when they are so extreme that the data cannot distinguish between the values. In particular, we fix σ_φ, letting the variability of the GP be determined by ρ_φ (see Zhang (2004) for asymptotic justification). We also fix μ_φ, thereby centering the eigenprocesses at reasonable values. Since we are working in geographic space, in which distances in different directions are measured in the same units, for the eigenvector processes, we use a single ρ_γ as a spatial range parameter common to the two processes. For all of the eigenprocesses, we fix $\nu = 5$, because it should have minimal impact on the spatial surface estimate and is not well-informed by the data. We take $\boldsymbol{u}_\varphi \sim \mathrm{N}(0, I)$ and $\log(\rho_\varphi) \sim U(\log(0.1), \log(3.85))$ and force $\lambda_1(\cdot)$ and $\lambda_2(\cdot)$ to lie in $(1/(n^2), 16)$. These constraints ensure propriety (discussed in Berger, De Oliveira, and Sansó (2001)) and prevent the Markov chain from getting stuck in regions with a flat likelihood.

Marginally, the prior covariance structure is stationary because the locations are exchangeable, but any particular realization from the prior, conditioning on sample eigenprocesses, gives a nonstationary covariance structure. The posterior produces a nonstationary covariance structure, induced by the posterior distribution of the eigenprocesses.

Since it is difficult to encapsulate prior knowledge about the spatial surface directly into the GP priors for the eigenprocesses, one could also place an additional prior on the complexity of the posterior mean spatial surface, conditional on the eigenprocesses. This induces a prior on the covariance structure (Hodges and Sargent 2001). Since the GP model is a linear smoother, the complexity can be estimated by the trace of the smoothing matrix,

$$df = \mathrm{tr}(C_{\boldsymbol{f}}(C_{\boldsymbol{f}} + C_{\boldsymbol{y}})^{-1}) + 1 \tag{11}$$

(Hastie and Tibshirani 1990, p. 52; Hodges and Sargent 2001; Ruppert, Wand, and Carroll 2003, p. 256), where we add one to account for the degree of freedom used to estimate μ_f. Note that in the case where $\eta = 0$ (no nugget effect), the smoother exactly interpolates the data, a saturated model, and *df* is equivalent to the number of observations, which is the sensible upper limit on the complexity it is possible to estimate with a dataset of size n. Our results are not based on this additional prior, because the nonstationary model did not tend to use more *df* than the stationary model, presumably because of the natural Bayesian penalty on model complexity (Denison, Holmes, Mallick, and Smith 2002, p. 20).

3.2.3 MCMC sampling—One can integrate $\boldsymbol{f}$, the spatial process evaluated at the observation locations, out of the GP model, leaving a marginal posterior whose marginal likelihood is,

$$\boldsymbol{Y} \sim \mathrm{N}(\mu_f, C_{\boldsymbol{f}}^{NS} + \eta^2 I). \tag{12}$$

where C_f^{NS} is the nonstationary covariance matrix of the spatial process at the observation locations. In the stationary GP model, the marginal posterior contains a small number of hyperparameters to either optimize or sample via MCMC. In the nonstationary case, the dependence of C_f^{NS} on the kernel matrices precludes straightforward optimization; instead we use MCMC. We sample the parameters at the first level of the prior hierarchy, μ_f, σ_f, ν_f, and η, via Metropolis-Hastings. Sampling the eigenprocesses and their hyperparameters is more involved. For a given eigenprocess, $\varphi(\cdot) \in \{\log(\lambda_2(\cdot)), \gamma_1(\cdot), \gamma_2(\cdot)\}$, we choose to sample, via Metropolis-Hastings, ρ_φ and (as a vector) $\boldsymbol{u}_\varphi$. $\boldsymbol{\varphi}$ is not sampled directly, but is determined by the representation (10), thereby involving the eigenprocess hyperparameters directly in the marginal likelihood through their effect on $\boldsymbol{\varphi}$ and therefore on C_f^{NS} in (12). This is analogous to the uncentered parameterization discussed in Gelfand, Sahu, and Carlin (1996), in contrast to the centered parameterization, which in this case would involve sampling $\boldsymbol{\varphi}$ rather than $\boldsymbol{u}_\varphi$ and in which acceptance of hyperparameter proposals would depend only on their priors and the GP distributions of the eigenprocesses. In our experience, the uncentered approach, in which the hyperparameters are informed directly by the data, mixes faster than the centered approach. Christensen, Roberts, and Sköld (2006) discuss on-the-fly reparameterizations, but their focus is on spatial processes that determine mean structure, unlike this situation in which the eigenprocesses are involved in parameterizing the nonstationary covariance structure. Furthermore their reparameterizations are computationally intensive and may involve computations with numerically singular matrices when the Matérn covariance is used.

Note that in sampling the spatial process conditional on the nonstationary covariance, pivoting (e.g., see the R chol() function) is sometimes necessary because the conditional posterior variance (9) is numerically singular. This occurs because the unknown spatial process is generally estimated to be locally smooth and therefore highly correlated at locations close in space.

4 Case Study

Simulation results suggest that the nonstationary GP model performs as hoped on relatively simple datasets, detecting and adjusting to heterogeneity in the unknown function (Paciorek and Schervish 2004, 2005).

Here we apply the model to real data expected to exhibit nonstationarity, the Colorado precipitation data introduced in Section 3.1.2, where we applied nonstationary kriging.

To quantitatively assess the performance of the nonstationary model, we compare its performance to several alternatives, most importantly a stationary GP model, described in Section 4.3.

4.1 Data

We fit the model to the full set of data from 1981 ($n = 217$) to qualitatively assess the performance of the nonstationary model and analyze the degree and nature of the nonstationarity in the data. We then compare the fit of the nonstationary model to the alternative methods based on held-out data. To this end, we use the replication in time available in the data archive to create 47 datasets of annual precipitation (1950–1996) in Colorado with 120 training locations and 30 test locations for each year, fitting each year separately. Note that both training and test locations differ by year and that more than 150 locations are available for most years, but that we use only 150 to keep the amount of information in each dataset constant.

4.2 Evaluation criteria

We use several criteria to assess the quality of fit based on held-out data. We compute the proportion of variance explained, $R^2=1-\sum_{m=1}^{M}(y_m^*-\widehat{f}_m)^2/\sum_{m=1}^{M}(y_m^*-\bar{y}^*)^2$, where $\widehat{f}_m(y_m^*)$ is the estimated surface (held-out value) at test locations, $\boldsymbol{x}_m^*$, $m = 1, \ldots, M$. This assesses the surface point estimate (the posterior mean for the Bayesian models). To assess the model as a whole (for the Bayesian estimators only), we report the log predictive density, $h(\mathbf{y}^*|\mathbf{y})$ (Carlin and Louis 2000, p. 220), on test data using the full posterior, averaging over the MCMC samples, $t = 1, \ldots, T$:

$$\mathrm{LPD}=-\frac{1}{2}\log(2\pi)+\frac{1}{M}\log\frac{1}{T}\sum_{t=1}^{T}\frac{1}{\eta_{(t)}^{M}}\exp\left(-\frac{1}{2\eta_{(t)}^2}\sum_{m=1}^{M}(y_m^*-f_{m,(t)})^2\right),$$

where $\eta_{(t)}$ is the sampled error standard deviation and $f_{m,(t)}$ is the sampled function value. Larger values of LPD indicate better estimates. Finally, we estimate the coverage and average length of prediction intervals at the test locations.

For Bayesian methods, adequate mixing and convergence are important and determine the number of MCMC samples needed and therefore the computational speeds of the methods. We compare the iterations of the sample log posterior density (Cowles and Carlin 1996) and key parameters between the methods to get a general sense for how many iterations each needs to run, examining the autocorrelation and effective sample size (Neal 1993, p. 105),

$$\mathrm{ESS}=\frac{T}{1+2\sum_{d=1}^{\infty}\rho_d(\theta)},$$

where T is the number of MCMC samples and $\rho_d(\theta)$ is the autocorrelation at lag d for the quantity/parameter θ.

4.3 Alternative methods

The first group of alternatives includes standard methods that can be easily implemented in R, most using library functions. The second group comprises Bayesian methods that in theory can adapt to heterogeneity in the function by sampling the basis functions within the MCMC. The abbreviations used in the results are given parenthetically here in the text. The Bayesian nonstationary GP model is abbreviated ‘nsgp’.

4.3.1 Standard spatial methods—The first method (sgp) is a stationary, anisotropic version of the nonstationary Bayesian GP model. After integrating the spatial process values at the training locations out of the model, the parameters, $\{\mu, \eta, \sigma, \rho_1, \rho_2, \psi, \nu\}$, are sampled via MCMC. The second method (krig) is likelihood-based kriging as described in Section 3.1.2, but we also estimate ν in the numerical optimization. The third method (tps) fits the surface as a thin plate spline (Green and Silverman 1994), using the Tps() function in the fields library in R, in which the smoothing parameter is chosen automatically by generalized cross-validation (GCV). The fourth method (gam) also fits the spatial surface using a thin plate spline and GCV, but in a computationally efficient way (Wood 2003), coded in the gam() function in the mgcv library in R. One advantage of gam() over Tps() that arises in applications is that gam() allows the inclusion of additional covariates, including additional smooth terms, using an algorithm that can optimize multiple penalty terms (Wood 2000).

Since these methods all rely on a small number of smoothing parameters/penalty terms that do not vary with location, none are designed to handle nonstationarity. Also note that only the kriging and the stationary Bayesian GP approaches are designed to handle anisotropy.

4.3.2 Nonparametric regression models—There are many nonparametric regression methods, with much work done in the machine learning literature as well as by statisticians. These methods can be used for spatial smoothing; we restrict our attention to a small number of methods with readily available code. Other potential methods include wavelet models, mixtures of GPs or thin plate splines, and regression trees. The first two methods considered are free-knot spline models that, by allowing the number and location of the knots to change during the fitting procedure, can model nonstationarity. Denison et al. (1998) created a Bayesian version of the MARS algorithm (Friedman 1991), which uses basis functions that are tensor products of univariate splines in the truncated power basis, with knots at the data points. The second method (mls) uses free-knot multivariate linear splines (MLS) where the basis functions are truncated linear planes, which gives a surface that is continuous but not differentiable where the planes meet (Holmes and Mallick 2001). In the simulations and case study, we report numerical results only for the MLS basis, because it performed better than the MARS basis. The final method (nn) is a neural network model, in particular a multilayer perceptron with one hidden layer, with the spatial surface modelled as $f(\boldsymbol{x})=\beta_0+\sum_{k=1}^{K}\beta_k g_k\left(\boldsymbol{u_k^T x}\right)$, where the $g_k(\cdot)$ functions are tanh functions and the $\boldsymbol{u_k}$ parameters determine the position and orientation of the basis functions. This is very similar to the MLS model, for which $g_k(\cdot)$ is the identity function. We use the Bayesian implementation of R. Neal (http://www.cs.toronto.edu/~radford/fbm.software.html) to fit the model, fixing $K = 200$ to allow for a sufficiently flexible function but minimize computational difficulties.

4.4 Results

4.4.1 Qualitative performance—We first consider the stationary and nonstationary GP models fit to the 1981 data. Figure 3 shows the posterior mean and pointwise posterior standard deviation of the spatial surfaces from the two models. The results match our intuition, with features that follow the topography of Colorado (Fig. 1). The nonstationary surface is smoother in the east than the stationary surface, while both are quite variable in the mountainous west. The posterior standard deviations are much smaller for the nonstationary model in the east than for the stationary model and generally somewhat larger in the west, particularly at locations far from observations. The stationary model, in trying to capture the variability in western Colorado, infers what appears to be too little smoothness and too little certainty in eastern Colorado. With either approach (as we will see with other smoothing methods as well) it appears very difficult to estimate a precipitation surface in western Colorado based on such a small number of weather stations. The posterior means of the degrees of freedom (11), averaging across the iterations, are 194 for the stationary model and 140 for the nonstationary model for a dataset with only 217 observations. The best model selection criteria for distinguishing between these Bayesian models is unclear, but if we consider the AIC_C criterion (Hurvich et al. 1998) applied to each iteration of the chains, the nonstationary model outperforms the stationary model, with the larger residual variance of the nonstationary model offset by the penalty on the larger degrees of freedom of the stationary model.

The correlation structure for the nonstationary model is shown in Figure 4 by ellipses of constant density representing the Gaussian kernels, $K_{\boldsymbol{x_i}}(\cdot)$; $i = 1, \ldots,$ used to parameterize the nonstationary covariance structure. Analogous ellipses are also shown for the kriging models and the stationary GP model. The kernels for the nonstationary model are much larger in the east than in the west, as expected, but also increase in size in the extreme west

of the state where the topography is less extreme. The posterior standard deviations for the surface correspond (Fig. 3d) to the size of the kernels (Fig. 4d). The model imposes smoothly varying kernels, in contrast to the kernels used in the ad hoc nonstationary kriging approach (Fig. 4b), thereby removing the east-west boundary discontinuity seen with nonstationary kriging (Fig. 3). The substantive result that the surface is smoother and more certain in the east remains qualitatively similar to splitting the state into two regions. The nonstationary model has the advantage of giving shorter pointwise credible intervals for the unknown function (Section 4.4.2 suggests there is no loss of coverage), with an average interval length of 0.79 for the nonstationary model and 1.14 for the stationary model over a regular grid of 1200 locations.

The Matérn differentiability parameter exhibits drastically different behavior in the stationary and nonstationary models, with the posterior mass concentrated near 0.5 for the stationary model ($E(\nu|\mathbf{y}) = 0.7$; $P(\nu < 1|\mathbf{y}) = 0.92$), while being concentrated at large values for the nonstationary model ($E(\nu|\mathbf{y}) = 16$; $P(\nu > 1|\mathbf{y}) = 0.99$). Because differentiability concerns behavior at infinitesimal distances, we suspect that when ν is estimated in the model it does not provide information about the differentiability of the surface, but rather about the correlation structure at the finest scale resolved by the data. In the stationary case, ν seems to account for inadequacy in model fit, reflecting local variability that the model would otherwise be unable to capture because of the global correlation structure. In the nonstationary model, the varying kernels are able to capture this behavior. Small values of ν are unnecessary, but the model has little ability to distinguish large values of ν. Paciorek (2003, chap. 4) found a similar difference between stationary and nonstationary models in one-dimension for a simple, highly-differentiable function with a sharp bump. We suspect that the popularity of the exponential covariance in spatial statistics can be explained in part by the fact that small ν compensates for stationary model inadequacy and allows for local adaptation when the underlying function is highly variable with respect to the resolution of the observation locations.

Mixing with these complicated real data was slow. For the year 1981, we ran the stationary model for 10,000 iterations (5 hours in R for 217 observations and 1200 prediction locations) and saved every tenth iteration, while running the nonstationary model for 220,000 iterations (in R, several days run time), again saving every tenth. Based on the log posterior density of the models from these runs, the effective sample size for the stationary model was 809 while for the nonstationary model it was only 140 (12 based on only the first 1000 subsampled observations to match the stationary model), with particularly slow mixing for the eigenprocess hyperparameters. The picture is somewhat brighter for the spatial surface estimates; averaging over the estimates at 10 test locations, the effective sample size based on the 1000 subsampled iterations was 334 with a standard deviation of 7 for the stationary model and, based on the 22,000 subsampled observations, 2950 (488 from the first 1000) with a standard deviation of 1890 (228) for the nonstationary model. While we believe the estimates from the nonstationary model are reasonable for calculating posterior means and standard deviations, albeit not for quantiles, we remain concerned about mixing and computational efficiency, but note that computational approaches mentioned in Section 5 may help mixing. Note that the MCMC performance of the free-knot spline and neural network models was also poor, suggesting that nonstationary methods are generally difficult to fit.

4.4.2 Comparison based on multiple datasets—Based on the 47 years of data, Colorado precipitation appears to be difficult for any method to predict, with an average of 20–40 percent, and for some datasets zero percent, of the variability in held-out data explained by the spatial surface (Fig. 5a). Based on this test R^2, the stationary and nonstationary GP models outperform the other methods ($p < 0.0002$), but the nonstationary

model does no better than the stationary model. This lack of improvement mirrors the lack of improvement found when comparing between the simple kriging, thin plate spline and GAM approaches and those same approaches with Colorado divided into two regions, with the regional approaches providing little improvement (Fig. 5a).

Turning to the log predictive density comparison, the nonstationary GP is significantly better than the stationary GP and MLS ($p < 0.007$), marginally better than the neural network ($p = 0.10$), and significantly worse than the likelihood-based kriging estimates, interpreted as Bayesian estimates ($p < 0.0032$) (Fig. 5c).

The coverage of prediction intervals on the test data, averaged over the 47 years, is 0.949 and 0.945 for the stationary and nonstationary models, respectively. Both coverages are good, but the nonstationary model performs slightly better because it has shorter intervals on average, 1.14, compared to 1.21 for the stationary model. The relatively small difference in intervals compared to the difference in Section 4.4.1 reflects the fact that the test locations are more concentrated in the mountainous areas than the 1200 grid locations considered previously, more of which lie in eastern Colorado. Both models have long intervals in the mountainous areas, whereas the nonstationary intervals are shorter than the stationary intervals in eastern Colorado.

Given the poor mixing of the hyperparameters for the eigenprocesses in the nonstationary GP model and the potential to borrow strength across the 47 years of data, we reran the model with common hyperparameter values for all 47 years, fixed based on the runs reported above, but found little difference in the results.

5 Discussion

We have introduced a class of nonstationary covariance functions, generalizing the kernel convolution approach of Higdon et al. (1999). The class includes a nonstationary Matérn covariance function with parameterized sample path differentiability. When there is a nugget effect, we do not think that one can reliably estimate the differentiability parameter and interpret the value as reflecting the differentiability of the underlying surface, but our generalized kernel convolution approach allows us to avoid specifying infinitely differentiable realizations, which is the case under the original Higdon et al. (1999) form. Stationarity is a special case of the nonstationary covariance function. The model is built upon spatially varying covariance parameters; if these parameters vary smoothly, the nonstationary covariance function can be thought of as being locally stationary. Building nonstationarity from local stationarity is an appealing approach; Haas (1995), Barber and Fuentes (2004), and Kim et al. (2005) consider nonstationary models with subregions of stationarity. We demonstrate an ad hoc fitting approach for a nonstationary kriging model and develop a fully Bayesian model, which we have shown elsewhere to perform well in simulations (Paciorek and Schervish 2004, 2005). On a challenging real data example, the model produces qualitative results that are more sensible than a stationary model. The lengths of prediction intervals vary sensibly in accord with climatological considerations, and the model gives good coverage with smaller intervals than a stationary model. However, the model improves little upon the stationary model based on several predictive criteria evaluated with held-out data. The nonstationary model also provides a graphical approach for assessing spatial correlation structure based on the posterior kernel ellipses, which may be of interest for either replicated or nonreplicated data. With sufficient replicated data, one could plot the spatial correlation contours for individual locations; our approach allows one to visually judge the spatial correlation structure for all locations in one plot.

In defense of the method relative to other nonstationary models, we have seen little quantitative evidence in other work that nonstationary models outperform stationary models for prediction or coverage in spatial settings. The limited improvement of the nonstationary model over the stationary model and the lack of predictive improvement seen when stationary models are applied separately to eastern and western Colorado relative to a stationary model for the whole state indicate the difficulty in fitting a complicated, locally highly-variable surface with relatively few observations even though substantive knowledge of Colorado and model selection criteria (Section 3.1.2) suggest that a stationary model is not appropriate. Part of this may be due to a bias-variance tradeoff, with the bias of the stationary method offset by the high variance of the nonstationary method, although this does not explain why fitting just two stationary models applied separately to eastern and western Colorado does not improve matters. One difficulty with spatial data, as opposed to regression data, is that local heterogeneity may be much more important than spatially uncorrelated noise in any decomposition of the nugget effect (the decomposition is discussed in Cressie (1993, sec. 3.2.1)). With sparse data and most of the heterogeneity truly spatial in character, we may nearly interpolate the data; note the large number of degrees of freedom in Section 4.4.1. In this situation, the exact weighting involved in the local averaging performed by smoothing methods may not be all that important. The key to success for a stationary model may be merely that it avoid oversmoothing; in areas in which the spatial surface is relatively smooth, a stationary method that has a single small spatial range estimate and relies more on local averaging than a ‘correct’ nonstationary method may perform just as well because of the lack of uncorrelated noise that, in a regression setting, would hurt the performance of a stationary smoother. This may be the case for the point predictions in the Colorado example.

5.1 Computational improvements

While the Bayesian nonstationary model nicely ties the local covariance parameters together in a smooth way, the computational demands are a drawback. The slowness of model fitting arises because of the $O(n^3)$ matrix calculations involved in the marginal posterior, after integrating the function values at the observation locations out of the model, as well as the calculations involved in calculating the kernel matrices that determine the nonstationary covariance. We have provided a basis function representation of the processes in the hierarchy of the model that determine the kernel matrices at the locations of interest. This approximation speeds the calculations, but other representations may be faster and may produce faster mixing. One possibility is to use a thin plate spline basis as in Ngo and Wand (2004). Alternatively, Wikle (2002) and Paciorek and Ryan (2005) use a spectral basis representation of stationary Gaussian processes, which allows use of the FFT to dramatically improve speed, while also showing mixing benefits by a priori orthogonalization of the basis coefficients. The circulant embedding approach, which also relies on FFT calculations, is another possibility (Wood and Chan 1994). Relatively coarse resolutions are likely to be sufficient given that the processes in the hierarchy in the model should not be complicated functions.

GP models, stationary or nonstationary, are relatively slow to fit because of the marginal likelihood computations. One computational strategy would be to use a knot-based approach similar to that of Kammann and Wand (2003) (Section 3.2.2), representing the function at the observation locations as $\boldsymbol{f} = \mu_f \mathbf{1} + \sigma_f \Psi \Omega^{-1/2} \boldsymbol{u}_{\boldsymbol{f}}$, where Ψ is a matrix of nonstationary correlations between the n observation locations and K knot points based on the nonstationary covariance given in this paper and Ω is a similar matrix with pairwise nonstationary correlations between knot locations. While this approach requires one to sample the vector of coefficients, $\boldsymbol{u}_{\boldsymbol{f}}$, rather than integrating $\boldsymbol{f}$ out of the model, it replaces the $O(n^3)$ matrix calculations of the marginal posterior model with $O(K^3)$ calculations.

Williams, Rasmussen, Schwaighofer, and Tresp (2002) and Seeger and Williams (2003) use a similar computational approach, with Ω based on a subset of the training locations. Higdon (1998) uses a discrete representation of the kernel convolution (3) to avoid matrix inversions, representing nonstationary processes as linear combinations of kernel smooths of discrete white noise process values. While computationally efficient, we have had difficulty in getting this approach to mix well (Paciorek 2003, chap. 3).

In this work, we achieve nonstationary by letting the range and directionality parameters of stationary, anisotropic correlation functions vary over the space of interest. In light of Zhang (2004)'s result that the range and variance parameters in a GP with stationary Matérn covariance cannot be simultaneously estimated in a consistent fashion, one might consider achieving nonstationarity by instead letting the variance parameter vary over the space of interest, $\sigma^2(\cdot)$, taking $f(\cdot) \sim GP(\mu, \sigma^2(\cdot)R^S(\cdot; \rho, \nu))$, with $R^S(\cdot)$ a stationary correlation function. This has a positive definite covariance function and is simpler to model because one needs only one hyperprocess, for the variance, instead of the various processes determining the kernels in the approach presented in this paper. Such an approach would be similar to the penalized spline models of Lang and Brezger (2004) and Crainiceanu et al. (2004).

5.2 Extensions

We have focused on Gaussian responses and simple smoothing problems, but the nonstationary covariance structure may be useful in other settings, and, provided the computational challenges are overcome, as the spatial component in more complicated models, such as spatio-temporal and complicated hierarchical Bayesian models. When the domain is a large fraction of the earth's surface, nonstationary models can adapt to the distortion of distances that occurs when the spherical surface of the earth is projected. One could easily use the nonstationary model within an additive model with additional covariates. Finally, the Bayesian model in Section 3.2.1 simplifies easily to one-dimension and has been extended to three dimensions in practice (Paciorek and Schervish 2004) and in principle can be extended to higher dimensions. Given the success and variety of nonparametric regression techniques for one-dimensional smoothing, the model may be of most interest for higher dimensional smoothing problems, such as three-dimensional spatial models.

The nonstationary covariance and spatial model proposed here can be easily extended in principle to non-Gaussian responses using standard link functions, as demonstrated in Paciorek and Schervish (2004) for Bernoulli data in a one-dimensional domain. However, even stationary models for non-Gaussian data are slow to fit and mix (Paciorek and Ryan 2005; Christensen et al. 2006), so estimating nonstationarity in practice may be difficult, particularly for non-Gaussian data with limited information per observation, such as binary data.

The nonstationary covariance may also be used for replicated data with the advantage of the additional covariance information provided by the replications. However, when we tried this for the multi-year Colorado dataset, using common kernel matrices across time, we found that the poor mixing prevented us from fitting the Bayesian model in reasonable time. Ad hoc approaches might involve estimating the kernels at each location based on the replications and local information (e.g., Barber and Fuentes 2004). Stein (2005) has extended our class of nonstationary covariance functions to allow for spatially-varying ν by allowing the scale parameter, S, defined in Appendix 2, to have a distribution that varies spatially. This extension which would be of most interest with replicated data.

Acknowledgments

The authors thank Doug Nychka for helpful comments and suggestions and acknowledge financial support for the first author through an NSF VIGRE grant to the Department of Statistics at Carnegie Mellon University. The project was also supported by grant number 5 T32 ES007142-23 from the National Institute of Environmental Health Sciences (NIEHS), NIH to the Department of Biostatistics at Harvard School of Public Health. The contents are solely the responsibility of the authors and do not necessarily represent the official views of NIEHS, NIH.

References

Abrahamsen P. A review of Gaussian random fields and correlation functions. Technical Report 917, Norwegian Computing Center. 1997

Anderson TW, Olkin I, Underhill LG. Generation of random orthogonal matrices. SIAM Journal on Scientific and Statistical Computing. 1987; 8:625–629.

Barber JJ, Fuentes M. Nonstationary spatial process modeling of atmospheric pollution data. Journal of Agricultural, Biological, and Environmental Statistics. 2004 under revision.

Berger JO, De Oliveira V, Sansó B. Objective Bayesian analysis of spatially correlated data. Journal of the American Statistical Association. 2001; 96(456):1361–1374.

Carlin, BP.; Louis, TA. Bayes and Empirical Bayes Methods for Data Analysis. Chapman & Hall Ltd; 2000.

Christensen O, Roberts G, Sköld M. Robust MCMC methods for spatial GLMMs. Journal of Computational and Graphical Statistics. 2006 in press.

Cowles MK, Carlin BP. Markov chain Monte Carlo convergence diagnostics: A comparative review. Journal of the American Statistical Association. 1996; 91:883–904.

Crainiceanu, CM.; Ruppert, D.; Carroll, RJ. Spatially Adaptive Bayesian P-Splines with Heteroscedastic Errors. Technical Report 61, Department of Biotatistics; Johns Hopkins University. 2004.

Cressie, N. Statistics for Spatial Data. Revised ed. New York: Wiley-Interscience; 1993.

Damian D, Sampson P, Guttorp P. Bayesian estimation of semi-parametric non-stationary spatial covariance structure. Environmetrics. 2001; 12:161–178.

Denison D, Mallick B, Smith A. Bayesian MARS. Statistics and Computing. 1998; 8:337–346.

Denison, DG.; Holmes, CC.; Mallick, BK.; Smith, AFM. Bayesian Methods for Nonlinear Classification and Regression. New York: Wiley; 2002.

DiMatteo I, Genovese C, Kass R. Bayesian curve-fitting with free-knot splines. Biometrika. 2001; 88:1055–1071.

Friedman J. Multivariate adaptive regression splines. Annals of Statistics. 1991; 19:1–141.

Fuentes M. A high frequency kriging approach for non-stationary environmental processes. EnvironMetrics. 2001; 12:469–483.

Fuentes, M.; Smith, R. A New Class of Nonstationary Spatial Models. Technical report, North Carolina State University; Department of Statistics. 2001.

Gelfand, A.; Sahu, S.; Carlin, B. Efficient parametrizations for generalized linear mixed models. In: Bernardo, J.; Berger, J.; Dawid, A.; Smith, A., editors. Bayesian Statistics 5. 1996. p. 165-180.

Gibbs, M. Bayesian Gaussian Processes for Classification and Regression. unpublished Ph.D. dissertation, Univ. of Cambridge; 1997.

Gradshteyn, I.; Ryzhik, I. Tables of Integrals, Series and Products: Corrected and Enlarged Edition. New York: Academic Press, Inc; 1980.

Green, P.; Silverman, B. Nonparametric Regression and Generalized Linear Models. Boca Raton: Chapman & Hall/CRC; 1994.

Haas TC. Local prediction of a spatio-temporal process with an application to wet sulfate deposition. Journal of the American Statistical Association. 1995; 90:1189–1199.

Hastie, TJ.; Tibshirani, RJ. Generalized Additive Models. London: Chapman & Hall Ltd; 1990.

Higdon D. A process-convolution approach to modeling temperatures in the North Atlantic Ocean. Journal of Environmental and Ecological Statistics. 1998; 5:173–190.

Higdon, D.; Swall, J.; Kern, J. Non-stationary spatial modeling. In: Bernardo, J.; Berger, J.; Dawid, A.; Smith, A., editors. Bayesian Statistics 6. Oxford, U.K.: Oxford University Press; 1999. p. 761-768.

Hodges JS, Sargent DJ. Counting degrees of freedom in hierarchical and other richly-parameterised models. Biometrika. 2001; 88(2):367–379.

Holmes C, Mallick B. Bayesian regression with multivariate linear splines. Journal of the Royal Statistical Society, Series B. 2001; 63:3–17.

Hurvich CM, Simonoff JS, Tsai CL. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. Journal of the Royal Statistical Society, Series B, Methodological. 1998; 60:271–293.

Kammann E, Wand M. Geoadditive models. Applied Statistics. 2003; 52:1–18.

Kim HY, Mallick B, Holmes C. Analyzing nonstationary spatial data using piecewise Gaussian processes. J Am Stat Assoc. 2005; 100:653–668.

Lang S, Brezger A. Bayesian p-splines. Journal of Computational and Graphical Statistics. 2004; 13:183–212.

McLeish D. A robust alternative to the normal distribution. The Canadian Journal of Statistics. 1982; 10:89–102.

Neal, R. Probabilistic Inference Using Markov Chain Monte Carlo Methods. Technical Report CRG-TR-93-1, Department of Computer Science; University of Toronto. 1993.

Neal, R. Bayesian Learning for Neural Networks. New York: Springer; 1996.

Ngo L, Wand M. Smoothing with mixed model software. Journal of Statistical Software. 2004:9.

Paciorek, C. Nonstationary Gaussian Processes for Regression and Spatial Modelling unpublished Ph.D. dissertation. Carnegie Mellon University: Department of Statistics; 2003.

Paciorek, C.; Ryan, L. Computational techniques for spatial logistic regression with large datasets. Technical Report 32; Harvard University Biostatistics. 2005.

Paciorek, C.; Schervish, M. Nonstationary covariance functions for Gaussian process regression. In: Thrun, S.; Saul, L.; Schölkopf, B., editors. Advances in Neural Information Processing Systems 16. Cambridge, MA: MIT Press; 2004. p. 273-280.

Paciorek, C.; Schervish, M. Spatial modelling using a new class of nonstationary covariance functions. Technical Report 822; Department of Statistics, Carnegie Mellon University. 2005.

Rasmussen, CE.; Ghahramani, Z. Infinite mixtures of Gaussian process experts. In: Dietterich, TG.; Becker, S.; Ghahramani, Z., editors. Advances in Neural Information Processing Systems 14. Cambridge, MA: MIT Press; 2002.

Ruppert, D.; Wand, M.; Carroll, R. Semiparametric regression. Cambridge, U.K: Cambridge University Press; 2003.

Sampson P, Guttorp P. Nonparametric estimation of nonstationary spatial covariance structure. J Am Stat Assoc. 1992; 87:108–119.

Schmidt A, O'Hagan A. Bayesian Inference for Nonstationary Spatial Covariance Structures via Spatial Deformations. Journal of the Royal Statistical Society, Series B. 2003; 65:743–758.

Schoenberg I. Metric spaces and completely monotone functions. Ann of Math. 1938; 39:811–841.

Seeger M, Williams C. Fast forward selection to speed up sparse Gaussian process regression. Workshop on AI and Statistics. 2003; 9

Smith, R. Environmental Statistics. Technical report, Department of Statistics; University of North Carolina. 2001.

Stein, M. Interpolation of Spatial Data : Some Theory for Kriging. N.Y.: Springer; 1999.

Stein, M. Nonstationary spatial covariance functions. Technical Report 21; University of Chicago, Center for Integrating Statistical and Environmental Science. 2005.

Swall, J. Non-Stationary Spatial Modeling Using a Process Convolution Approach, unpublished Ph.D. dissertation. Duke University: Institute of Statistics and Decision Sciences; 1999.

Tresp, V. Mixtures of Gaussian processes. In: Leen, TK.; Dietterich, TG.; Tresp, V., editors. Advances in Neural Information Processing Systems 13. MIT Press; 2001. p. 654-660.

Wikle, C. Spatial modeling of count data: A case study in modelling breeding bird survey data on large spatial domains. In: Lawson, A.; Denison, D., editors. Spatial Cluster Modelling. Chapman & Hall; 2002. p. 199-209.

Williams, C.; Rasmussen, C.; Schwaighofer, A.; Tresp, V. Observations on the Nyström Method for Gaussian Process Prediction. Technical report, Gatsby Computational Neuroscience Unit; University College London. 2002.

Wood A, Chan G. Simulation of stationary Gaussian processes in $[0, 1]^d$. Journal of Computational and Graphical Statistics. 1994; 3:409–432.

Wood S, Jiang W, Tanner M. Bayesian mixture of splines for spatially adaptive nonparametric regression. Biometrika. 2002; 89:513–528.

Wood SN. Modelling and smoothing parameter estimation with multiple quadratic penalties. Journal of the Royal Statistical Society, Series B. 2000; 62(2):413–428.

Wood SN. Thin plate regression splines. Journal of the Royal Statistical Society, Series B. 2003; 65(1):95–114.

Zhang H. Inconsistent estimation and asymptotically equal interpolation in model-based geostatistics. J Am Stat Assoc. 2004; 99:250–261.

6 Appendix 1: Convolution

Here we show that the nonstationary covariance function introduced by Higdon et al. (1999) has a simple closed form when Gaussian kernels, $K_{\boldsymbol{x}}(\boldsymbol{u}) = (2\pi)^{-P/2}|\Sigma|^{-1/2}\exp(-(1/2)(\boldsymbol{x}-\boldsymbol{u})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{u})$ are used. We make use of the equivalence of convolutions of densities with sums of independent random variables:

$$\begin{aligned} C^{NS}(\boldsymbol{x}_i, \boldsymbol{x}_j) &= \int_{\Re^P} K_{\boldsymbol{x}_i}(\boldsymbol{u}) K_{\boldsymbol{x}_j}(\boldsymbol{u}) d\boldsymbol{u} \\ &= \int \frac{1}{(2\pi)^{\frac{P}{2}}|\Sigma_i|^{\frac{1}{2}}} \exp\left(-\tfrac{1}{2}(\boldsymbol{x}_i - \boldsymbol{u})^T \Sigma_i^{-1}(\boldsymbol{x}_i - \boldsymbol{u})\right) \\ &\quad \{\frac{1}{(2\pi)^{\frac{P}{2}}|\Sigma_j|^{\frac{1}{2}}} \exp\left(-\tfrac{1}{2}(\boldsymbol{x}_j - \boldsymbol{u})^T \Sigma_j^{-1}(\boldsymbol{x}_j - \boldsymbol{u})\right) d\boldsymbol{u}. \end{aligned}$$

Recognize the expression as the convolution

$$\int h_{\boldsymbol{A}}(\boldsymbol{u} - \boldsymbol{x}_i) h_U(\boldsymbol{u}) d\boldsymbol{u} = \int h_{\boldsymbol{A},U}(\boldsymbol{u} - \boldsymbol{x}_i, \boldsymbol{u}) d\boldsymbol{u},$$

where $h(\cdot)$ is the Gaussian density function, $\boldsymbol{A} \sim \mathrm{N}(\mathbf{0}, \Sigma_i)$, $\boldsymbol{U} \sim \mathrm{N}(\boldsymbol{x}_j, \Sigma_j)$, and $\boldsymbol{A}$ and $\boldsymbol{U}$ are independent. Now consider the transformation $\boldsymbol{W} = \boldsymbol{U} - \boldsymbol{A}$ and $\boldsymbol{V} = \boldsymbol{U}$, which has Jacobian of 1. This gives us the following equalities based on the change of variables:

$$\begin{aligned} \int h_{\boldsymbol{A},U}(\boldsymbol{u} - \boldsymbol{x}_i, \boldsymbol{u}) d\boldsymbol{u} &= \int h_{\boldsymbol{W},\boldsymbol{V}}(\boldsymbol{u} - (\boldsymbol{u} - \boldsymbol{x}_i), \boldsymbol{u}) d\boldsymbol{u} \\ &= \int h_{\boldsymbol{W},\boldsymbol{V}}(\boldsymbol{x}_i, \boldsymbol{u}) d\boldsymbol{u} \\ &= h_{\boldsymbol{W}}(\boldsymbol{x}_i). \end{aligned}$$

Since $\boldsymbol{W} = \boldsymbol{U} - \boldsymbol{A}$, $\boldsymbol{W} \sim \mathrm{N}(\boldsymbol{x}_j, \Sigma_i + \Sigma_j)$ and therefore

$$\begin{aligned} C^{NS}(\boldsymbol{x}_i, \boldsymbol{x}_j) &= h_{\boldsymbol{W}}(\boldsymbol{x}_i) \\ &= \frac{1}{(2\pi)^{\frac{P}{2}}|\Sigma_i + \Sigma_j|^{\frac{1}{2}}} \exp\left(-\tfrac{1}{2}(\boldsymbol{x}_i - \boldsymbol{x}_j)^T (\Sigma_i + \Sigma_j)^{-1}(\boldsymbol{x}_i - \boldsymbol{x}_j)\right). \end{aligned}$$

Absorbing the necessary constants into the matrices in the quadratic form and dividing by the standard deviation function, $\sigma(\boldsymbol{x_i})=\frac{1}{2^{\frac{p}{2}}\pi^{\frac{p}{4}}|\Sigma_i|^{\frac{1}{4}}}$, we arrive at the nonstationary correlation function, $R^{NS}(\cdot,\cdot)$, defined by

$$R^{NS}(\boldsymbol{x_i},\boldsymbol{x_j})=|\Sigma_i|^{\frac{1}{4}}\left|\Sigma_j\right|^{\frac{1}{4}}\left|\frac{\Sigma_i+\Sigma_j}{2}\right|^{-\frac{1}{2}}\exp\left(-(\boldsymbol{x_i}-\boldsymbol{x_j})^T\left(\frac{\Sigma_i+\Sigma_j}{2}\right)^{-1}(\boldsymbol{x_i}-\boldsymbol{x_j})\right).$$

7 Appendix 2: Proof of Theorem 1

Proof of Theorem 1: The proof is a simple application of Theorem 2 of Schoenberg (1938, p. 817), which states that the class of functions positive definite on Hilbert space is identical to the class of functions of the form,

$$R(\tau)=\int_0^\infty \exp(-\tau^2 s)dH(s), \tag{13}$$

where $H(\cdot)$ is non-decreasing and bounded and $s \geq 0$. The class of functions positive definite on Hilbert space is identical to the class of functions that are positive definite on $\Re^p$ for $p = 1, 2, \ldots$ (Schoenberg 1938). We see that the covariance functions in this class are scale mixtures of the squared exponential correlation function. The underlying stationary correlation function with argument $\sqrt{Q_{ij}}$ can be expressed as

$$\begin{aligned}
& |\Sigma_i|^{\frac{1}{4}}\left|\Sigma_j\right|^{\frac{1}{4}}\left|(\Sigma_i+\Sigma_j)/2\right|^{-\frac{1}{2}}R(\sqrt{Q_{ij}}) \\
= \quad & \int_0^\infty |\Sigma_i|^{\frac{1}{4}}\left|\Sigma_j\right|^{\frac{1}{4}}\left|(\Sigma_i+\Sigma_j)/2\right|^{-\frac{1}{2}}\exp(-Q_{ij}s)dH(s) \\
= \quad & \int_0^\infty |\Sigma_i|^{\frac{1}{4}}\left|\Sigma_j\right|^{\frac{1}{4}}\left|(\Sigma_i+\Sigma_j)/2\right|^{-\frac{1}{2}}\exp\left(-(\boldsymbol{x_i}-\boldsymbol{x_j})^T\left(\frac{\frac{\Sigma_i}{s}+\frac{\Sigma_j}{s}}{2}\right)^{-1}(\boldsymbol{x_i}-\boldsymbol{x_j})\right)dH(s) \\
= \quad & \int_0^\infty \left|\Sigma_i^s\right|^{\frac{1}{4}}\left|\Sigma_j^s\right|^{\frac{1}{4}}(4\pi)^{\frac{p}{2}}\int_{\Re^p} K_{\boldsymbol{x_i}}^s(\boldsymbol{u})K_{\boldsymbol{x_j}}^s(\boldsymbol{u})d\boldsymbol{u}dH(s),
\end{aligned}$$

where $K_{\boldsymbol{x_i}}^s$ is a Gaussian kernel with mean $\boldsymbol{x_i}$ and variance $\Sigma_i^s=\frac{\Sigma_i}{4s}$ and the last step follows by the convolution computation of Appendix 1. Since S is non-negative, it simply scales the kernel matrices, and the last expression can be seen to be positive definite by the same argument as (4):

$$\begin{aligned}
& \sum_{i=1}^n\sum_{j=1}^n a_i a_j |\Sigma_i|^{\frac{1}{4}}\left|\Sigma_j\right|^{\frac{1}{4}}\left|(\Sigma_i+\Sigma_j)/2\right|^{-\frac{1}{2}}R(\sqrt{Q_{ij}}) \\
= \quad & \sum_{i=1}^n\sum_{j=1}^n a_i a_j \int_0^\infty\int_{\Re^p}\left|\Sigma_i^s\right|^{\frac{1}{4}}\left|\Sigma_j^s\right|^{\frac{1}{4}}(4\pi)^{\frac{p}{2}}K_{\boldsymbol{x_i}}^s(\boldsymbol{u})K_{\boldsymbol{x_j}}^s(\boldsymbol{u})d\boldsymbol{u}dH(s) \\
= \quad & (4\pi)^{\frac{p}{2}}\int_0^\infty\int_{\Re^p}\sum_{i=1}^n a_i\left|\Sigma_i^s\right|^{\frac{1}{4}}K_{\boldsymbol{x_i}}^s(\boldsymbol{u})\sum_{j=1}^n a_j\left|\Sigma_j^s\right|^{\frac{1}{4}}K_{\boldsymbol{x_j}}^s(\boldsymbol{u})d\boldsymbol{u}dH(s) \\
= \quad & (4\pi)^{\frac{p}{2}}\int_0^\infty\int_{\Re^p}\left(\sum_{i=1}^n a_i\left|\Sigma_i^s\right|^{\frac{1}{4}}K_{\boldsymbol{x_i}}^s(\boldsymbol{u})\right)^2 d\boldsymbol{u}dH(s) \\
\geq \quad & 0.
\end{aligned}$$

Q.E.D.

Remark: The new class of nonstationary covariances has, as members, scale mixtures of the original nonstationary covariance of Higdon et al. (1999). Using different distributions, H, for the scale parameter, S, produces different nonstationary correlation functions. Using an integral expression for the Bessel function (Gradshteyn and Ryzhik 1980, p. 340, eq. 9; McLeish 1982), one can easily show that for the Matérn form (8), S is distributed inverse-gamma (ν, 1/4). Another example is the rational quadratic covariance, whose stationary form is $R(\tau)=\left(1+\left(\frac{\tau}{\rho}\right)^2\right)^{-\nu}$, which produces GPs with infinitely differentiable realizations (Paciorek 2003, chap. 2). A nonstationary version of the rational quadratic correlation function is

$$R(\boldsymbol{x}_i, \boldsymbol{x}_j)=|\Sigma_i|^{\frac{1}{4}}\left|\Sigma_j\right|^{\frac{1}{4}}\left|\frac{\Sigma_i+\Sigma_j}{2}\right|^{-\frac{1}{2}}\left(\frac{1}{1+Q_{ij}}\right)^{\nu},$$

which can be seen to be of the scale mixture form by taking $S \sim \Gamma(\nu, 1)$,

$$\int \exp(-Q_{ij}s)dH(s)=E(\exp(-Q_{ij}s))=M_S(-Q_{ij};\nu, 1)=\left(\frac{1}{1+Q_{ij}}\right)^{\nu},$$

where M_S is the moment generating function of S. This makes sense because the rational quadratic correlation function has the form of a t density, which is a mixture of Gaussians with an inverse gamma distribution for the variance, proportional to $\frac{1}{S}$, of the Gaussian.

Paciorek (2003, chap. 2) shows that the existence of moments of S is directly related to the existence of sample path derivatives of GPs parameterized by the nonstationary covariance (this is also true for stationary covariance functions). The number of moments of the inverse gamma distribution depends on its first parameter, which for the scale mixture for the nonstationary Matérn is ν. In the rational quadratic form, the gamma distribution has infinitely many moments, which corresponds to infinitely many sample path derivatives for GPs parameterized by either the stationary or nonstationary versions of the correlation function.

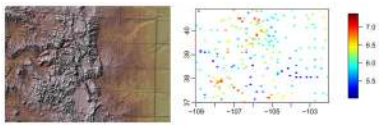

Figure 1.
(a) Topography of Colorado, with thicker line indicating state boundary, and (b) image plot of log-transformed annual precipitation observations in 1981, with the color of the box indicating the magnitude of the observation.

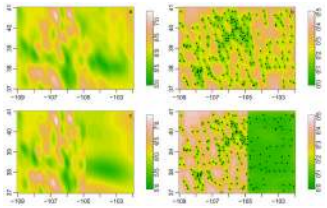

Figure 2.
Surface estimates from stationary (a) and nonstationary (c) kriging with corresponding standard deviations (data locations overlaid as points), (b) and (d).

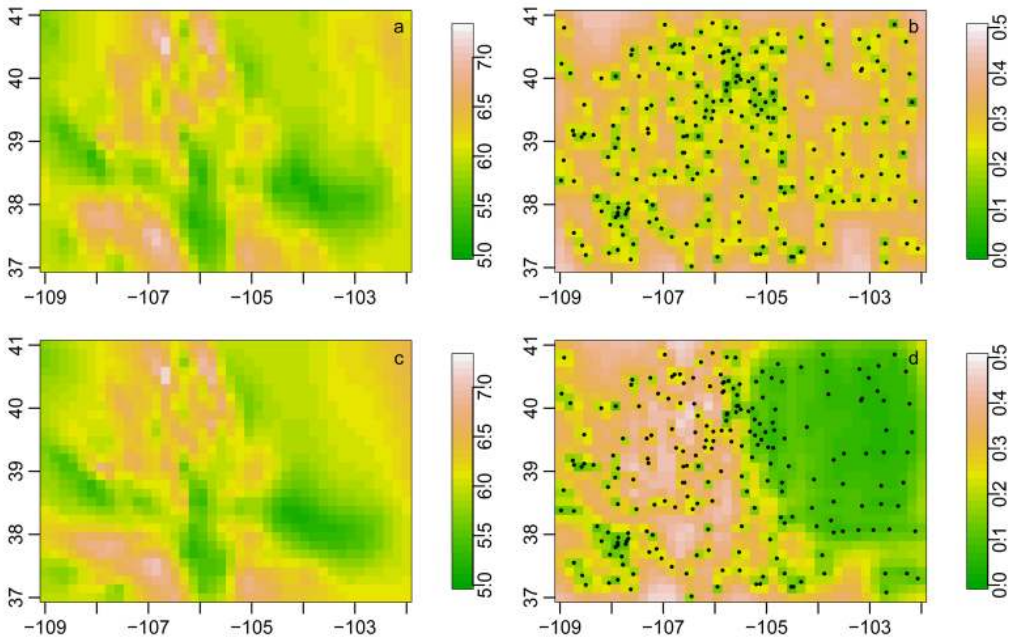

Figure 3.
Posterior mean surface estimates from stationary (a) and nonstationary (c) GP models with corresponding posterior standard deviations (data locations overlaid as points), (b) and (d).

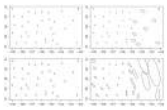

Figure 4.
Kernels (ellipses of constant probability density of Gaussian densities) representing the estimated correlation structure for (a) stationary kriging, (b) nonstationary kriging based on two regions, (c) the fully Bayesian stationary GP model, and (d) the nonstationary GP model. For the Bayesian models, the ellipse-like figures are the posterior means of constant probability density ellipse values at a sequency of angles, $0,\ldots,2\pi$.

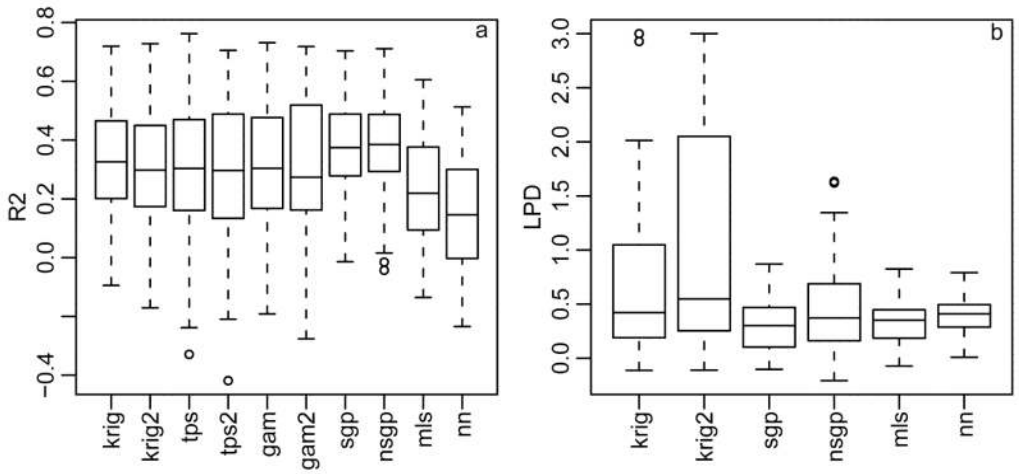

Figure 5.

(a) Test R^2 and (b) log predictive density (LPD) for the methods (labels indicated in Section 4.3) on 47 years of Colorado precipitation for test locations within the convex hull of the training data. Values of LPD larger than 3 are not plotted. Methods with a '2' in their label (krig2, tps2, gam2) were fit separately to eastern and western Colorado.

Table 1

Parameter estimates for the stationary and nonstationary models for Colorado precipitation in 1981.

	$\hat{\mu}$	$\hat{\eta}$	$\hat{\sigma}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\psi}$
stationary	6.05	0.17	0.39	0.089	0.037	105°
nonstationary, West	6.12	0.13	0.44	0.076	0.028	100°
nonstationary, East	6.11	0.14	0.30	0.30	0.16	165°

Environmetrics. Author manuscript; available in PMC 2007 December 27.