Large spatial prediction problems and nonstationary random fields

Douglas Nychka
Geophysical Statistics Project, National Center for Atmospheric Research,
Christopher Wikle,
Department of Statistics, University of Missouri
J. Andrew Royle
U.S. Fish and Wildlife Service Adaptive Management and Assessment Team

Abstract

Many geophysical and environmental problems depend on estimating a spatial process based on a large number of irregularly spaced observations. In addition, such processes often have nonstationary structure. Standard spatial estimates are based on the solution to a large linear system that is of the same size as the number of observations and a straight forward solution is prohibitive for large, spatial datasets. In this work we illustrate how the iterative solution of linear systems makes it possible to find good approximate solutions to large spatial problems. One key to this approach is the use of spatial process models that allow for efficient multiplication of the covariance matrix by an arbitrary vector. This is possible for stationary covariances by making connections with spatial convolution kernels. Because many spatial processes are heterogeneous in space an important extension of these algorithms is to covariances that are nonstationary. Recent developments in multiresolution bases such as wavelets allow one flexibility in capturing nonstationary structure and also permit rapid evaluation of the basis functions. The net result is a flexible models for spatial structure that can be computed efficiently in an interactive data analysis environment. As an illustration of this approach we consider monthly precipitation records at approximately 800 sites in the Rocky Mountain region of the US. Here the goal is to produce spatial predictions on a regular grid and sample the posterior distribution to produce valid realizations of the precipitation spatial process.

Contents

1 Introduction

A basic problem in the geosciences is estimating a spatial surface from irregularly spaced measurements. Geophysical data sets are often characterized by a large number of observations and by a spatial structure that changes depending on location. These two features thwart the application of standard methods of spatial statistics. The matrix computations associated with spatial predictions become difficult beyond several hundred observations. Also, the nonstationary behavior of many geophysical fields invalidate the use of stationary covariance functions that are the norm in geostatistics. In this work we propose two strategies for tackling large geophysical data sets. The first addresses the problems of computing standard spatial estimators for large numbers of observations by discretizing the spatial locations and taking advantage of fast multiplication algorithms for covariance functions evaluated on a grid. The second and more substantial contribution is the use of localized sets of basis functions for modeling nonstationary spatial processes.

A motivating problem for this work is the need to construct monthly precipitation fields for approximately 100 years for the United States. These fields along with other meteorological variables will then be used as the inputs to numerical models for vegetation(e.g. the Century model ??). The number of recording stations in this data base exceeds 5000 sites and has irregular spatial coverage (see figure 1). Precipitation is moderated by many factors including elevation, latitude, local topography and the paths of prevailing storm tracks. Because all of these factors vary over space it is unlikely to expect the spatial covariance for this field at any time to be stationary. Thus, the goal is to derive precipitation fields at regular grid points based on the observational record, accounting for the inherent spatial nonstationarity of this process. Modeling nonstationarity is not only important for more accurate predictions but also to obtain reliable measurements of uncertainty. In particular, to assess the sensitivity of a numerical model to the input climate variables it is useful to study model results for an ensemble of different versions of the input fields. In a statistical context it natural to construct ensembles by sampling from the posterior distribution for the spatial field. The validity of the posterior distribution will in turn hinge on how well the spatial process has been modeled and for this reason accounting for nonstationary structure is important.

A key feature of this work is the reliance on iterative methods to solve large linear systems. This is possible because the covariance matrices have been constructed so that one can evaluate the covariance function on a regular grid and multiple this covariance matrix with a vector rapidly. This set of techniques centered on the conjugate gradient method is a well developed tool in the geosciences but has been given less attention by statisticians. The ability to solve large spatial prediction problems does not come without some concessions. The most important is that the spatial coordinates be discretized and comprise a subset of points from a regular grid. We do not require that the process be observed at all grid points but the grid must be a superset of all observation locations and points for spatial prediction. The discretization poses little practical problem for cases where one does not expect to resolve very fine detail and turns out to be an appropriate assumption for many spatial data sets. Moreover, due to the numerical efficiency of this approach fine grids can be entertained and so the effects of discretization of the locations is minimized.

Building nonstationarity into covariance models directly is difficult and so in this work we take the alternative approach of constructing a model for the process and then finding the implied covariance function. Let $f(\boldsymbol{x})$ be a spatial process where

$$f(\boldsymbol{x}) = \sum_{\nu=1}^{\infty} \alpha_\nu \psi(\boldsymbol{x})$$

and $\{a_\nu\}$ are mean zero random variables with covariance matrix Σ_a. By representing a spatial process in terms of a judiciously chosen set of basis functions and assuming that Σ_a is either diagonal or sparse it becomes possible to compute spatial predictions for large problems and in addition to model nonstationary spatial structure. The basis functions provide a direct model for the process and so it is possible to simulate fields and to add temporal structure to the spatial process. (Wikle et al. 1998) This is in contrast to the more standard approach of modeling the covariance function where incorporating temporal structure is difficult.

A complete treatment of the problems of estimating nonstationary covariance models is beyond the scope of this paper but it is helpful in this introduction to highlight a component of many geophysical processes that simplify the statistical analysis. Many spatial processes for the atmosphere and ocean have a temporal component that provides for replications of the spatial field. For example in the case of precipitation, the monthly precipitation fields are nearly uncorrelated. In this case the covariance between two stations can be found from the sample covariance of observations made at different times. This situation is different

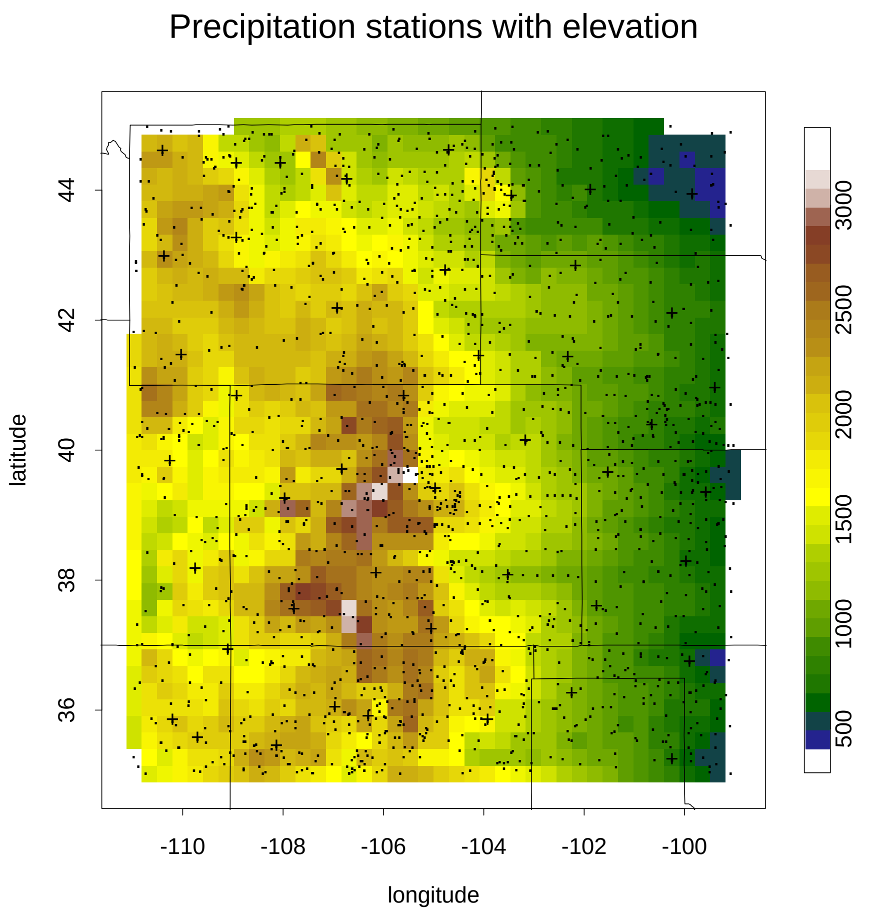

Figure 1: Precipitation sites for the Central Rocky Mountain region of the US for the period 1960 - 1989. Average elevations are overlaid as an image plot.

from the problems in geostatistics where only one realization of the spatial process, e.g. a mineral field, is available. Some care needs to be taken in considering the interaction between properties of the process over space and those over time and an example of a complicated space/time process being built from a hierarchical approach is treated in Wikle et. al. (1998). In any event it should be emphasized that the computational contributions of this paper form a kernel of methods to extend to more complicated modeling situations. For example, the space/time reconstructions of wind fields in Wikle et.al requires the basic spatial prediction computation as a primitive step in a more elaborate Gibbs sampler. Thus, any advances on the simpler problem treated in this paper will have direct improvement on space/time modeling efforts.

The next section presents a standard spatial model and reviews the linear system that defines a spatial estimate. Section 3 discusses the computation of the spatial estimator for stationary covariances and a large number of observations. Section 4 extends this algorithm to include fixed effects and unequal nugget variances. The next section reviews models for nonstationary processes and this is followed by a development for processes using multiresolution bases. This followed by an example using monthly precipitation data and a discussion.

2 Models for Spatial Fields

First a simple spatial model will be discussed that highlights the computational and modeling issues. The main features are reliance on the spatial locations being a subset of a regular grid and the absence of a spatial drift term. Extensions discussed in Section 4 will make this a more practical method by including a fixed part in the model and unequal variances for the measurement error.

Let $g(\boldsymbol{x})$ denote a mean zero, Gaussian spatial process on $\Re^2$. We anticipate that the observation locations, $\{\boldsymbol{x}_k\}$ are irregular but we will assume that the locations for observation and prediction are a subset of an $M \times N$ grid of locations. Let $\boldsymbol{g}$ denote a vector of length MN that is the spatial process evaluated on the entire grid.

2.1 Observational model

The observational model is

$$\boldsymbol{y} = H\boldsymbol{g} + \boldsymbol{\epsilon}. \tag{1}$$

$\{\epsilon_k\}$ are mean zero uncorrelated random variables with $VAR(\epsilon_k) = \sigma^2$ and $\boldsymbol{y}$ is a vector of n observations. The $n \times MN$ matrix H will be referred to as an incidence matrix and models the relationship between the observed values and the underlying spatial field. For the simplest case, H consists of ones and zeroes with a single 1 in each row indicating the position of each observation with respect to the grid. The discretization of locations in this model is a purely computational device to facilitate working with large covariance matrices. Although in this work we expect n to be large (perhaps several thousand) this sample size will usually be substantially smaller than the total number of grid points ($M \times N$). Computing resources permitting , the grid can be chosen to be very fine and little is lost through this discretization. It should also be noted that the uncorrelated random component also serves to model the departures due to small scale discretization error.

2.2 Spatial Predictions

Let $\Sigma = COV(\boldsymbol{g})$ denote the $MN \times MN$ covariance matrix among all of the MN points in the grid and so

$$COV(\boldsymbol{y}) = H\Sigma H^T + \sigma^2 I$$

and

$$COV(\boldsymbol{g}, \boldsymbol{y}) = \Sigma H^T$$

We first focus on the standard spatial estimate derived from these covariances. Development for the best linear unbiased estimate appears in many places see for example (Cressie 1991, Welch et al. 1992, and Nychka 1998) and in this simple setting it is easy to show that the spatial predictions at the grid points have the form:

$$\hat{\boldsymbol{g}} = \Sigma H^T (H\Sigma H^T + \sigma^2 I)^{-1} \boldsymbol{y}$$

With this notation, the predictions at the observed data are $H\hat{\boldsymbol{g}}$.

Rewriting this expression as the solution to a linear system one first finds $\boldsymbol{\delta}$ so that

$$(H\Sigma H^T + \sigma^2 I)\boldsymbol{\delta} = \boldsymbol{y} \tag{2}$$

and then $\hat{g} = \Sigma H^T \boldsymbol{\delta}$.

If one assumes that the observation errors are multivariate normal then the conditional distribution of $\boldsymbol{g}$ given $\boldsymbol{y}$ is

$$\text{Gaussian} \quad (\Sigma H^T Q^{-1} \boldsymbol{y}, \Sigma - \Sigma H^T Q^{-1} H \Sigma)$$

where $Q = H\Sigma H^T + \sigma^2 I$. The conditional mean is just the spatial estimate given above and the conditional variance can be used to measure the uncertainty in the spatial estimate. A more comprehensive way to draw inferences from the spatial estimate is to examine the full conditional distribution for $\boldsymbol{g}$. This is the approach used in this work because of the need to create realistic input fields for vegetation models. Thus, we focus on generating random samples from this conditional distribution. The procedure for generation has four steps. (1) Generate a mean zero field with covariance Σ and simulate observations from this field, (2) compute the conditional mean based on these simulated quantities, (3) subtract this conditional mean from the actual values of the field and (4) add this simulated difference to the conditional mean found for the actual data. The net result is a multivariate normal vector with the same mean and covariance as the conditional distribution given above.

3 Computing a spatial prediction

The naive approach to solving (2) can result in very large matrices. For example if a spatial data set consists of 5000 points then $H\Sigma H^T$ will be a 5000×5000 matrix requiring about 100Mb of storage in single precision form. It can much get worse. If one works with a grid where $M = N = 256$, a modest size grid for an image, and one actually constructs Σ then it will require about 17Gb of storage for single precision numbers. Clearly a more parsimonious approach is needed. The basic idea to solve find solutions to the linear system without having to construct the full matrix, $H\Sigma H^T$. A standard approach is to use an iterative method such that at any step one only needs to compute the product of this large matrix with a vector. Solving a linear system in an iterative manner is well known in numerical analysis. The conjugate gradient algorithm (CGA) used here, is a mainstay for the data assimilation problems in the geophysical sciences and is simple to implement.

Any iterative method for the solution of (2) will avoid the problems of storing large matrices but still may be very slow simply because because the straight forward multiplication of large vectors is computationally intensive, the number of floating point operations being of order n^2. The second part to make this a practical technique is to formulate models where the relevant matrices for the CGA can be multiplied by arbitrary vectors efficiently, of order $nlog(n)$ or better. Specifically we need to consider $H\Sigma H^T\boldsymbol{u}$ and $\Sigma\boldsymbol{u}$ where $\boldsymbol{u}$ is any vector. By construction, H is a sparse matrix. If the locations are unique then H has only one nonzero element in each row and column. Therefore both $H\boldsymbol{u}$ and $H^T\boldsymbol{u}$ can be rapidly evaluated and the efficient multiplication of $H^T\Sigma H$ depends on the inner multiplication by Σ. In this section a brief description of the conjugate gradient method is given and we then describe how the covariance matrices for stationary covariance functions can be multiplied efficiently.

3.1 Conjugate gradient method

Let A denote a symmetric, positive definite matrix. Then the solution to $A\boldsymbol{x} = \boldsymbol{b}$ is equivalent to minimizing the quadratic form

$$\frac{1}{2}\boldsymbol{x}^T A\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x} \tag{3}$$

The resulting algorithm that exploits this duality is known as the conjugate gradient algorithm (CGA). In this section only a brief description is given of this algorithm and the reader is referred to Golub and Van Loan (1989) for an informative and complete development. The motivation for the conjugate gradient algorithm is the following. Let a current approximation to the solution be $\boldsymbol{x}$, and for a direction, $\boldsymbol{p}$, let $\boldsymbol{x}_{\text{new}} = \boldsymbol{x} + \alpha\boldsymbol{p}$. Now choose α to minimize (3). This operation is just a one dimensional search and has the solution $\alpha = \boldsymbol{p}^T(\boldsymbol{b} - A\boldsymbol{x})/\boldsymbol{p}^T A\boldsymbol{p}$. Thus, the computation of α only involves multiplication of A with a vector and one does not need to construct and store A provided that one perform the multiplications. The algorithm proceeds now by selecting a new direction and updating the solution. Each step results in an x_{new} that decreases (3) and so provides a closer approximation to $A^{-1}b$.

Of course the main difficulty is that it is not clear how the directions should be chosen at each step. An important feature of the CGA is that it is possible to generate a sequence of directions that insure convergence in a finite number of steps. If A is $n \times n$ then in the absence of roundoff error this algorithm will converge in n iterations to the solution. In practice, when n is large one expects to perform far fewer iterations. For this reason, the rate of convergence of CGA is a consideration. This rate will depend on the starting vector. In addition, the algorithm can be accelerated if one can identify a matrix A_0 so that A_0 is in some sense close to A and the system $A_0\boldsymbol{x} = \boldsymbol{b}$ can be solved easily. This matrix is then used to precondition A. Both of these features of the algorithm are important for computing spatial estimates and more will be said about these in Section 8.

3.2 Multiplying stationary covariance matrices

Because stationary covariances are the standard model for spatial methods it is appropriate to first describe how they can be accommodated by this computational approach. More complicated covariance functions are treated in subsequent.

If Σ is derived from a stationary covariance function then the multiplication $\Sigma \boldsymbol{u}$ is equivalent to a two dimensional convolution of the covariance kernel with a rectangular image. In making this identification with convolution one should note that the results at the boundaries are found by simply truncating the convolution kernel if it extends beyond the range of the data. A standard way to find this convolution efficiently it to use the Fast Fourier Transform (FFT) (Press et al. 1992). The order of this calculation is $MNlog(MN)$ if M and N are both powers of 2. Moreover, in this application the convolution can be made more efficient by storing the FFT of the convolution kernel, and so saving an FFT each time the convolution is found.

One drawback in applying this approach to images is that the padding by zeroes that is necessary to get the right boundary convolutions must happen in both dimensions. To reproduce the truncated kernels a circular convolution is effected on a $2M \times 2N$ size image and so the operation count and storage are inflated by a factor of 4 over working with original, unpadded image. This worst case may be avoided if the covariance kernel has limited support and so the buffer areas of zeroes need only be as wide as the support of the kernel.

4 Extension to fixed effects and nonconstant variances

In order to highlight the basic components of the algorithm the presentation did not include a fixed component in the spatial model and unequal variances for the measurement error. Here we address these extensions. The observational model (1) can be generalized to include a fixed linear component

$$\boldsymbol{y} = T\boldsymbol{\beta} + H\boldsymbol{g} + \boldsymbol{e}$$

where T is an $n \times m$ matrix of spatial covariates and β a parameter vector. These covariates can arise from several sources as a spatial drift, as adjustments for instrument biases or as spatial covariates (such as elevation for precipitation fields). We will assume that m is small relative to the size of n and so decompositions and multiplications of T require a computational burden similar to other operations that are order n. In addition, we will assume that the measurement error has covariance matrix $\sigma^2 W$ where W is a known, diagonal matrix. One reason that W is necessary is to adjust for the variance when several observation are made at the same locations and the sample mean is used for the spatial estimate. Although coincident observations of the field may not be common in practice, when the locations are discretized to a grid this is more likely to occur.

The best linear unbiased estimate for g and $\boldsymbol{\beta}$ (or the generalized Kriging estimate) with respect to this model is the solution to the linear system

$$\begin{aligned} T^T\boldsymbol{\delta} &= 0 \\ T\boldsymbol{\beta} + H\Sigma H^T\boldsymbol{\delta} + \lambda W^{-1}\boldsymbol{\delta} &= \boldsymbol{y} \end{aligned}$$

See Nychka (1998) for a derivation of this form.

To simplify these equations for computing, let $T = F_1 R$ denote the QR decomposition of T. $F = [F_1 \mid F_2]$ is an orthogonal matrix, F_1 has columns that span the column space of T, F_2 has columns that span the space orthogonal to T and R is an upper triangular matrix. From the derivation in Nychka (1998) the linear system can be transformed as

$$F_2^T(H\Sigma H^T + \lambda W^{-1})F_2\boldsymbol{\omega}_2 = F_2^T\boldsymbol{y}$$

$$R\boldsymbol{\omega}_1 \quad = \quad F_1^T \left(\boldsymbol{Y} - (H\Sigma H^T + \lambda W^{-1}) F_2 \boldsymbol{\omega}_2 \right)$$

with $\boldsymbol{\delta} = F_2$ and $\boldsymbol{\beta} = \boldsymbol{\omega}_1$. The form for the first equation given above, is similar in to (2) except for the additional multiplications by F_2 and F_2^T. These additional multiplications can be carried out in order mn and so the iterative approach proposed for the simpler spatial estimate is also appropriate to solve this system. In the second equation, $\boldsymbol{\omega}_1$ has length m and so the solution can be found quickly relative to the other parts of the overall computation.

5 Nonstationary covariance models

Although, in this work we will concentrate on random field derived from a basis function expansion a brief overview will be given for other models. One reason for this discussion is that the methods for spatial prediction based on stationary fields can be adapted for computing some of these other nonstationary models and more details on these connections will be given below.

5.1 Windowed Kriging

Perhaps the most standard approach in this field has used a locally windowed estimator to avoid the problems of large numbers of spatial locations and nonstationarity. An example with extension to temporal modeling can be found in Haas (1995). For each location a local, stationary analysis is done with spatial predictions being restricted to the center of the window. One conceptual objection to windowed estimates is that a single coherent spatial model has not been formulated for the process. With this comes the problems of choosing the size of the data window and the computational burden of solving many small spatial prediction problems to produce spatial predictions on a grid. Because the spatial domain is broken into many pieces, it is also difficult to include covariates, such as elevation in the precipitation example, that follow a common model across the whole spatial region. The work in this paper has a direct benefit to the windowed Kriging approach in allowing for larger domains for the windowed estimator. Taking into account the correlation range one can often break up a very large spatial domain into overlapping regions of moderate size and retain the character of a global estimate.

5.2 Geographic Deformation

A nonstationary field can be constructed by a nonlinear geographic transformation of a stationary field (Sampson and Guttorp 1992). Let $\Gamma : \Re^2 \rightarrow \Re^2$ be a nonlinear deformation of the plane and suppose that f is a stationary spatial process. $g(\boldsymbol{x}) = f(\Gamma(\boldsymbol{x}))$ defines a nonstationary process and the model provides a useful representation of spatial heterogeneity based on the deformation of Γ. Sampson and Guttorp proposed estimating Γ using thin plate splines. Although effective in incorporating nonstationary structure, the transformation can be difficult to estimate from spatial data, especially for large numbers of spatial locations. This method also depends on being able to compute a stationary spatial prediction in the deformed (f) space. Thus, the deformation model and the spatial predictions themselves can be computed using the algorithms in this paper.

5.3 Variable kernel

A nonstationary field can be constructed as a variable convolution of a stationary process. This method is illustrated by Higdon (1997) for sea surface temperatures in the Atlantic. For example, let ω be a convolution kernel, f a stationary field, $\theta(\boldsymbol{x})$ a positive function on $\Re^2$ and let

$$g(\boldsymbol{x}) = \frac{1}{\theta(\boldsymbol{x})} \int \omega\left(\frac{(\boldsymbol{x}-\boldsymbol{u})}{\theta(\boldsymbol{x})}\right) f(\boldsymbol{u}) d\boldsymbol{u}$$

Here θ can be recognized as a spatially varying bandwidth. This form is flexible in representing nonstationary structure but the way in which the covariance function depends on θ makes it difficult to multiply the covariance matrix times a vector efficiently.

By discretizing the choice of θ and approximating the convolution by a matrix multiplication, one can adapt this model to the computational form of this paper. Suppose that one discretizes the choice of bandwidth parameter to m values $\{\theta_1, ..., \theta_m\}$ and let $\boldsymbol{f}$ be a field evaluated on the regular grid. Let Ω_j be a matrix so that $\Omega_j \boldsymbol{f}$ approximates the convolution of ω with f with $\theta(\boldsymbol{x})$ held constant at the value θ_j. Let $\boldsymbol{g}_j = \Omega_j \boldsymbol{f}$ and

$$\boldsymbol{g} = A(\boldsymbol{g}_1|\boldsymbol{g}_2|\ldots|\boldsymbol{g}_m)$$

A is a $MN \times m$ matrix that weights $\{g_j\}$ to approximate a spatially varying bandwidth. For example, A could be chosen to interpolate between the discrete values for the bandwidth. Given continuous values for $\theta(\boldsymbol{x})$ one would approximate the convolutions at the grid points by linear (or higher order) interpolations of the smaller set of m fields $\{\boldsymbol{g}_j\}$. The covariance for this process is

$$COV(\boldsymbol{g}) = A(\Omega_1|\Omega_2|\ldots|\Omega_m)\Sigma_f(\Omega_1|\Omega_2|\ldots|\Omega_m)^T A^T$$

Under the assumption that Σ_f is sparse this covariance matrix can be efficiently multiplied with a vector. This follows from the limited number of columns of A and using the fast convolution methods for regular grids. Thus at least an approximate form of the variable kernel field is amenable to spatial prediction.

5.4 Basis Functions

Let ψ_ν for $1 \leq \nu \leq MN$ be a sequence of basis functions and we think of representing the spatial process as

$$g(\boldsymbol{x}) = \sum_{\nu=1}^{NM} a_\nu \psi_\nu(\boldsymbol{x}) \tag{4}$$

We restrict evaluation to locations on the grid and take the number of basis functions equal to the number of grid points. Because of the discrete nature of this setup it helps to rewrite (4) as matrix/vector expressions. Let Ψ be an $MN \times MN$ matrix with

$$\Psi_{\mu,\nu} = \psi_\nu(\boldsymbol{x}_\mu)$$

where $\boldsymbol{x}_\nu$ is the ν^{th} grid point. Thus (4) has the discrete analog $\boldsymbol{g} = \Psi \boldsymbol{a}$.

The final ingredient is to assume a stochastic model for the coefficients. In particular we assume that $E(\boldsymbol{a}) = 0$ and covariance matrix for $\boldsymbol{a}$ is Σ_a. The covariance for $\boldsymbol{g}$ is

$$\Sigma = \Psi \Sigma_a \Psi^T$$

.

With this model, the linear system from (2) can be solved to compute spatial predictions. In this case the efficient multiplication of Σ required by the conjugate gradient algorithm depends on efficient multiplication by the matrices Ψ, Ψ^T and Σ_a. We will assume that Σ_a is either diagonal or sparse and so our attention focuses on the choice of basis functions. A useful class for nonstationary models is developed in the next section.

6 Multiresolution bases

Multiresolution methods in particular, regression on wavelets, have received much recent attention in the statistics literature, especially in their ability to provide estimates of functions that have discontinuities or varying degrees of smoothness over their domain. As noted above a crucial property of any basis is the need to perform the multiplication of Ψ and Ψ^T with a vector quickly. The discrete wavelet transform (DWT) is a fast, recursive algorithm for a basis derived from a wavelet. More traditional treatments of wavelets start by defining the basis from the scale function and then ending with the discrete wavelet transform as a final computational device (e.g. Silverman and Nason 1994). Given the reliance on the DWT in this work and more general nonorthogonal transformations, it is better to start with the transform and then work backwards to deduce the basis.

6.1 The one-dimensional W transform

We first consider the operation of $\Psi^{-1}\boldsymbol{u}$, that is determining the coefficients in the basis expansion from the function values on a grid. For concreteness the presentation is restricted to a one dimensional, W transform basis (Kwong and Tang, 1994) for a vector that has length $n = 2^J$ for some J. These basis functions yield members that are approximately quadratic splines (Figure 2) but unlike other wavelet bases are not orthogonal.

Consider the linear transformation

$$H_n\boldsymbol{u} = \begin{pmatrix} 2 & 3 & -1 & & & & & & & & & \\ & -1 & 3 & 3 & -1 & & & & & & & \\ & & & -1 & 3 & 3 & -1 & & & & & \\ & & & & & & & \ddots & & & & \\ & & & & & & & -1 & 3 & 3 & -1 & \\ & & & & & & & & & -1 & 3 & 2 \\ 2 & -3 & 1 & & & & & & & & & \\ & -1 & 3 & -3 & 1 & & & & & & & \\ & & & -1 & 3 & -3 & 1 & & & & & \\ & & & & & & & \ddots & & & & \\ & & & & & & & -1 & 3 & -3 & 1 & \\ & & & & & & & & & -1 & 3 & -2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_{n/2} \\ d_1 \\ d_2 \\ \vdots \\ d_{n/2} \end{pmatrix}$$

The $\boldsymbol{s}$ components are smoothed values of the original vector and the $\boldsymbol{d}$ are details or the results of a high pass filter applied to the values in $\boldsymbol{u}$. The operation is just a convolution of the filter weights with $\boldsymbol{u}$ except for boundary adjustments but one distinction is that only every other value of the convolution is saved [1]

[1] The definition of the filter matrix differs from Kwong and Tang because these authors present the filters

After this first step the detail vector has length $n/2$ and we declare these to be the coefficients for $n/2$ of the basis functions. In fact because these detail filter weights only operate on the adjacent 4 values of $\boldsymbol{u}$ it is natural to associate these basis functions with the finest level of resolution in the expansion. The next step is to apply the operation $H_{n/2}$ to just the smooth vector. One now extracts an additional $n/4$ detailed coefficients that are identified with basis coefficients. The algorithm proceeds recursively accumulating coefficients from the detail vectors. When one stops the recursion the final smooth vector becomes the final coefficients associated with the basis functions. To notate the complete algorithm let I_m be an $m \times m$ identity matrix.

$$\Psi^{-1}\boldsymbol{u} = \begin{pmatrix} H_8 & 0 \\ 0 & I_{n-8} \end{pmatrix} \begin{pmatrix} H_{16} & 0 \\ 0 & I_{n-16} \end{pmatrix} \cdots \begin{pmatrix} H_{n/2} & 0 \\ 0 & I_{n/2} \end{pmatrix} (H_n)\,\boldsymbol{u} \tag{5}$$

Here we have stopped at 8 because the filters need at least 4 values to be defined.

Clearly from (5) the coefficients obtained in this manner are a linear function of $\boldsymbol{u}$. Also the inverse transformation can be obtained easily by reversing the order of the matrix products and inverting each of the matrices. One advantage of the filter weights from the W transform and for wavelets in general is that simple inverse filters exist to describe H_n^{-1}. In this case the form is

$$H_n = \frac{1}{16}\left(\begin{array}{cccccccccccc}
4 & & & & & & 4 & & & & & \\
3 & 1 & & & & & -3 & 1 & & & & \\
1 & 3 & 1 & & & & -1 & 3 & 1 & & & \\
 & 3 & 3 & & & & & -3 & 3 & & & \\
 & 1 & 3 & \cdots & & & & -1 & -3 & \cdots & & \\
 & & 1 & & & & & & -1 & & & \\
 & & & & & & & & & & & \\
 & & & & 1 & & & & & & 1 & \\
 & & & & 3 & & & & & & 3 & \\
 & & & & 3 & 1 & & & & & -3 & 1 \\
 & & & & 1 & 3 & & & & & -1 & 3 \\
 & & & & & 4 & & & & & & -4
\end{array}\right)$$

¿From this discussion we have identified recursive linear operations to compute Ψ and Ψ^{-1} and it is straight forward to extend these algorithms to the transposes of these matrices. Because the filter weights are of fixed length, 4 in this case, the multiplication of H_m by a vector is of order m. Thus, the complete multiplication of Ψ and Ψ^{-1} and their transposes is order n.

The basis functions associated with the W transform are defined implicitly. Although the matrix Ψ is never explicitly constructed it is simple to compute each row of Ψ. If e_μ is the μ^{th} unit vector Ψe_μ yields the μ^{th} basis function evaluated on the grid. Figure 2 illustrates the shape of these basis functions when $n = 64$.

Due to the translation of the filters at each level and scaling by 2 at each new level, one can expect that the basis functions at the same scale are translations except for boundary effects. Moving to the next level the support of the basis function doubles in size. The final set of basis functions associated with the smooth coefficients are different in shape but

to yield detail and smoothed coefficients that are interlaced. In this presentation we have already permuted rows so that the smoothed and detail coefficients end up as separate subvectors. This form makes it easier to think about applying the H matrix recursively.

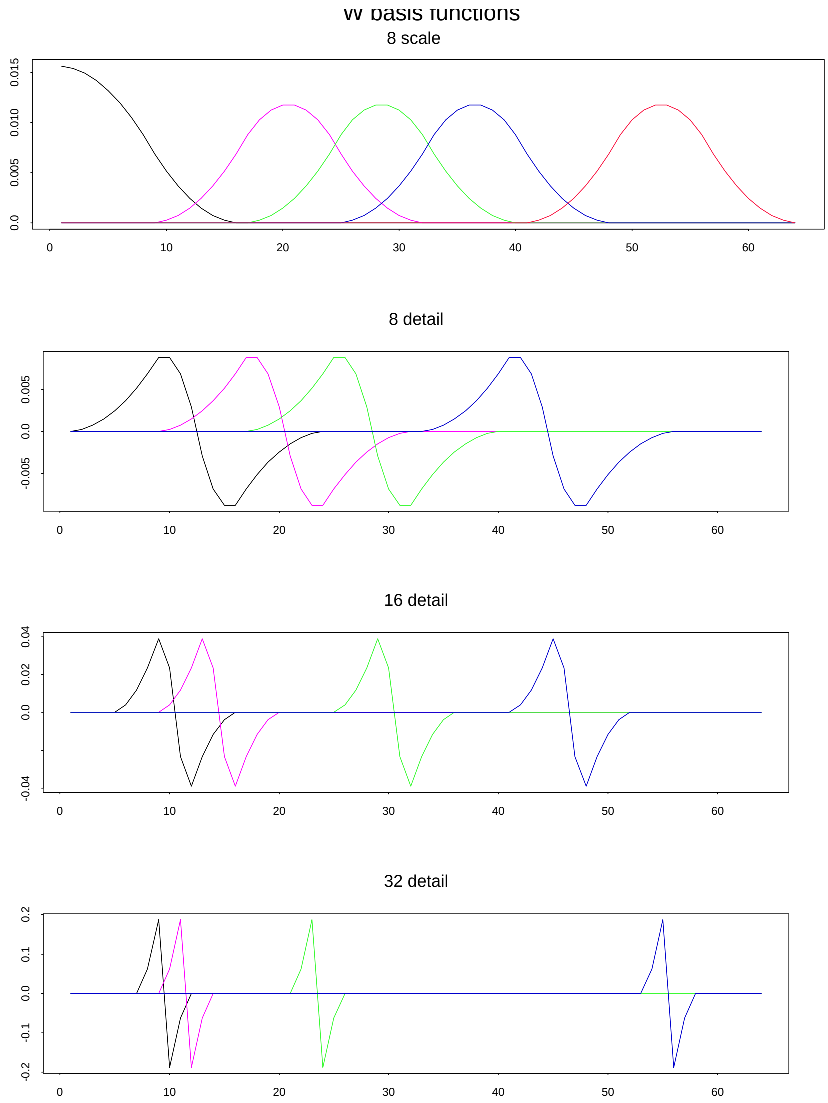

Figure 2: Examples of basis functions associated with the W transform for $n = 64$ equally spaced points.

again are translations except near the boundaries. As the number of levels increases the smooth and detail functions converge to translations and dilations of two fixed functions defined recursively by the filter weights. We refer to these as the smooth function and detail function respectively. In the case of the W transform these are the quadratic splines in figure 3.

By choosing different filter weights and different filter lengths one can alter the properties of the basis. There are several well known families of filter weights that yield orthogonal bases, i.e. $\Psi^{-1} = \Psi^T$, and the basis functions have compact support. Such choices are referred to as wavelets with the most well known example being the Daubechies wavelet (D_4) with smooth and and detail functions in the bottom plot in Figure 3.

6.2 A Two-dimensional Multiresolution Basis

A two-dimensional multiresolution basis is constructed by combining two one-dimensional transforms. Similar to the previous section the derivation will be rooted in the numerical algorithm. Let $\boldsymbol{Y}$ be a two-dimensional image in the form of an $M \times N$ matrix. Here both N and M are a power of 2. The wavelet transform first applies the convolutions of the smooth and detail coefficients to the columns of this matrix and this is followed by applying the same operation to the resulting rows. Schematically, one obtains from this first step.

$$(T_M(T_N\boldsymbol{Y})^T)^T = T_N\boldsymbol{Y}T_M^T = \left(\begin{array}{c|c} SS & DS \\ \hline SD & DD \end{array}\right) .$$

Here D and S refer qualitatively to the composition of the resulting elements as being either smooth or detail and each of the partitioned blocks is $M/2 \times N/2$. For example DS are the obtained by computing the detail coefficients from the one-dimensional transforms of the columns, applying the one-dimensional transform to these rows and then taking the smoothed coefficients. In image analysis these four sets of coefficients represent different parts of the image. SS are the coefficients for a smoothed version of the original with half its resolution. DS is the detail submatrix along vertical lines and SD is the detail along horizontal lines. Finally DD is the detail submatrix for diagonal structure in the image. Figure 4 is an illustration of this decomposition for a simple test image.

This test image has a spatial correlation with approximately three times the range on the left side of the image relative to the right side. Following this structure the three squares of coefficients involving the details tend to have larger magnitudes on their right sides. This is consistent with the way the image was constructed because visually the side with a shorter range of correlation has more detail. Also from this example one can note that the magnitude of the smoothed coefficients reproduce a coarser version of the original image. This is also expected because the smoothed coefficients are local averages of the image. The number of operations to transform all columns is of order NM and to repeat the transform on the rows is again order NM. Thus one step of the transform is linear in the size of the image.

The full algorithm is recursive and is analogous to the one dimensional transform. The three blocks of details are equated with coefficients of the basis functions and the same transformation is applied again to the $M/2 \times N/2$, SS block of the matrix. The algorithm is terminated at a given size of SS block and these last smoothed values are the remaining coefficients of the expansion. The final result is a sequence of vertical, horizontal and diagonal coefficients at different levels of resolution (see Figure 6) and a final block associated from the SS matrix. Taking into account that each of these steps involves a transform on

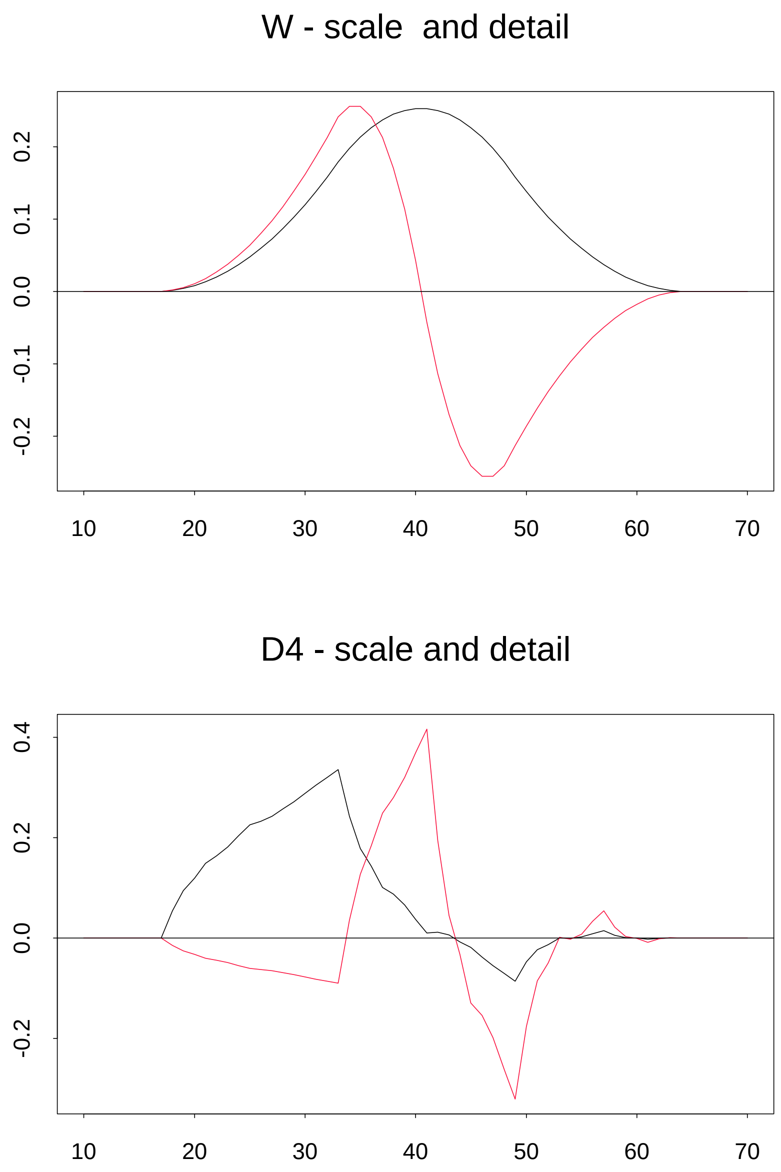

Figure 3: The limiting scale and detail functions for the W transform (top) and Daubechies D_4 basis (bottom).

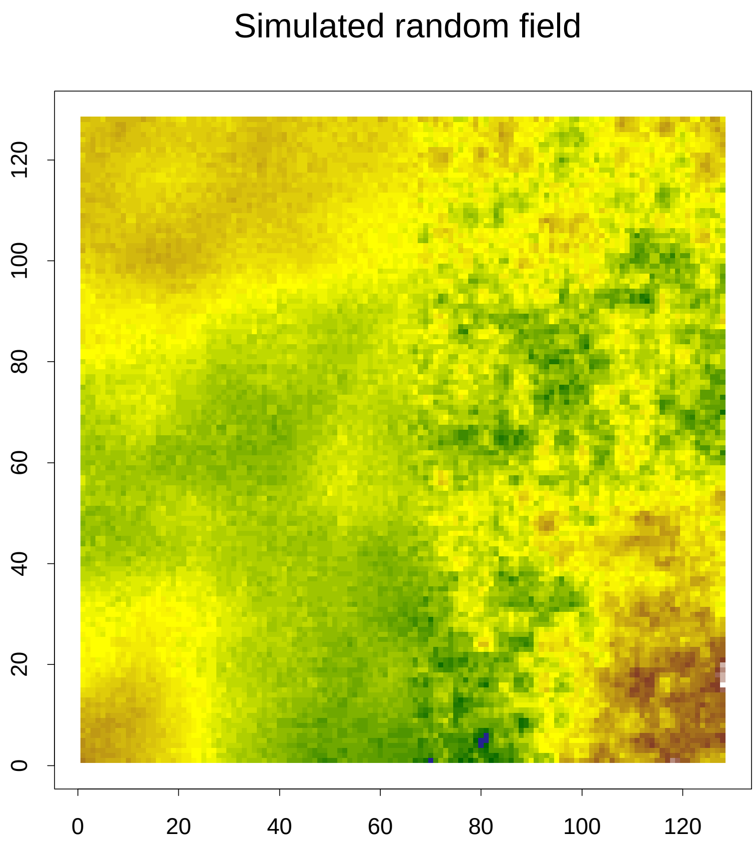

Figure 4: A simulated random field with resolution 128×128 pixels.

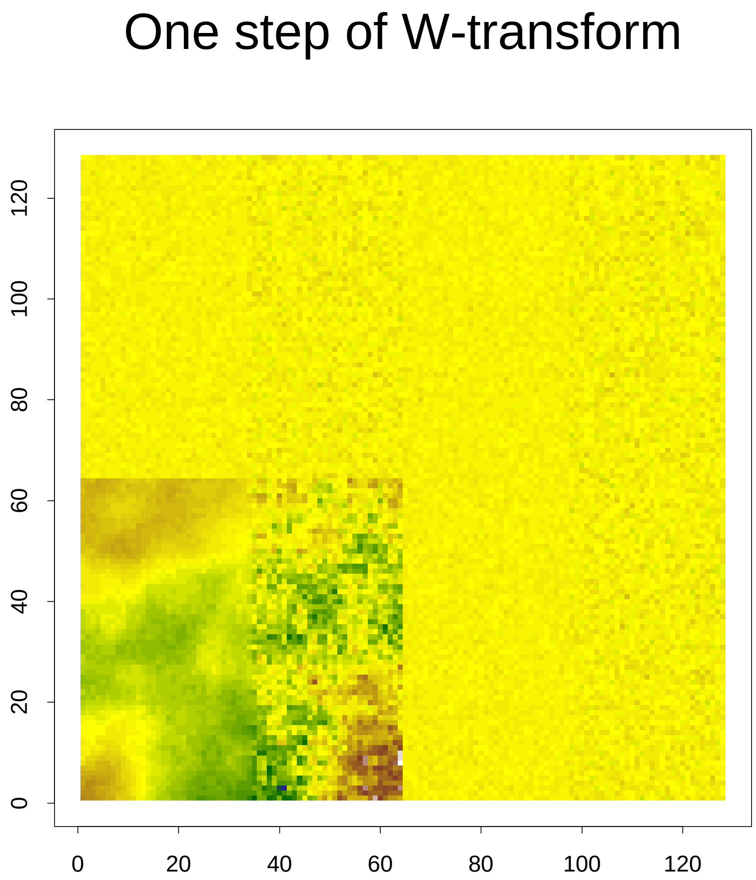

Figure 5: Coefficients from one step of the two dimensional W transform applied to the simulated image in Figure 4.

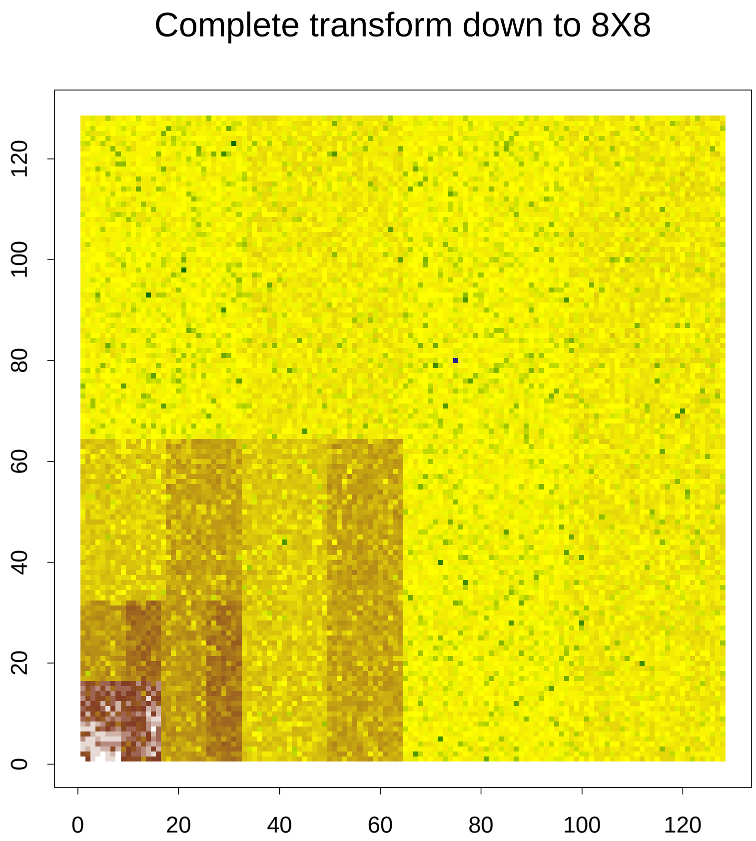

Figure 6: The full W transform of the image from the previous figure stopping at and 8×8 block of smoothed coefficients. Scaling for coefficients is the log of the absolute value.

Level	Smooth 8×8	detail 8	detail 16	detail 32	detail 64
"Gaussian"	1	0.05	0.0001	1e-08	1e-10
"Exponential"	1	0.015	0.0012	1e-06	1e-08

Table 1: Variances for W-transform basis coefficients used to define the covariance model from Figure 8

an image matrix 1/4 the size of the previous level it is clear that the full transform is also of order MN operations. The final block of SS coefficients can be identified as a very coarse representation of the image. For example note the highest and lowest coefficients are located at patches of high and low values in the image.

Repeating the arguments from the one-dimensional case this multiresolution transform for an image is also a linear operator and so can be represented by a matrix, Ψ^{-1}. It is also possible to identify the inverse and transpose of this matrix piggybacking from what we know about the one dimensional transforms. For example, if $U = H_N Y H_M^T$ then $Y = H_M^{T-1} U H_N^{-1}$. Therefore, the full expressions for Ψ and Ψ^T can be derived by expanding the transformation as products of block diagonal matrices in an analogous manner to (5).

Figure 7 illustrates the asymptotic form for the 4 types of basis functions associated with the 2 dimensional W transform.

Not surprisingly they are tensor products of the one dimensional smooth and detail functions. The full basis can be visualized as families of the detail functions translated and scaled with the last few basis functions being translations of the scale function.

6.3 Covariance functions from the W transform

One issue in selecting a basis is whether the covariance function implied by diagonal Σ_a can approximate standard parametric covariance families. One advantage of the W transform is the potential to model spatial processes that are commonly encountered in applications. This is the reason that this particular basis was chosen over more common wavelet basis that have compact support and are orthogonal. Of course one promise of this approach is that the added flexibility from a multiresolution basis will represent nonstationary processes. But as a start, it is useful to verify that this basis can approximate stationary processes.

As an example, the W transform covariance function was evaluated on a 128×128 grid according to specific choices for the diagonal elements of Σ_a. The transform was terminated at a block of 8×8 smoothed values and so there are 64 smooth basis functions. The elements of Σ_a are one for the smoothed basis functions and constant for each level of detail basis functions. Note that for a 128×128 image with 64 smooth basis functions there are 4 levels of detail and for this first example the sequence of variances is given in Table 1. (The rapidly decreasing sequence is partly the result of the W transform of using integer valued filter weights and so not being normalized between levels.) Figure 8 plots the implied covariance surfaces at particular locations. The first plot is the covariance function for a location in the middle of the domain and we see that covariance function has nearly circular contours and a shape in cross-section that is similar to a Gaussian. The similarity of this surface with ones when the location is translated suggests that the covariance is approximately stationary. The next pair of plots is for a location nearer to the edge of the domain and here we see that the stationarity is breaking down due to the edge effects of the transform. Contours are no longer circular with some distortion near the boundary. The last covariance surface is based on the variances in the second row of Table 1 and illustrates the potential to represent

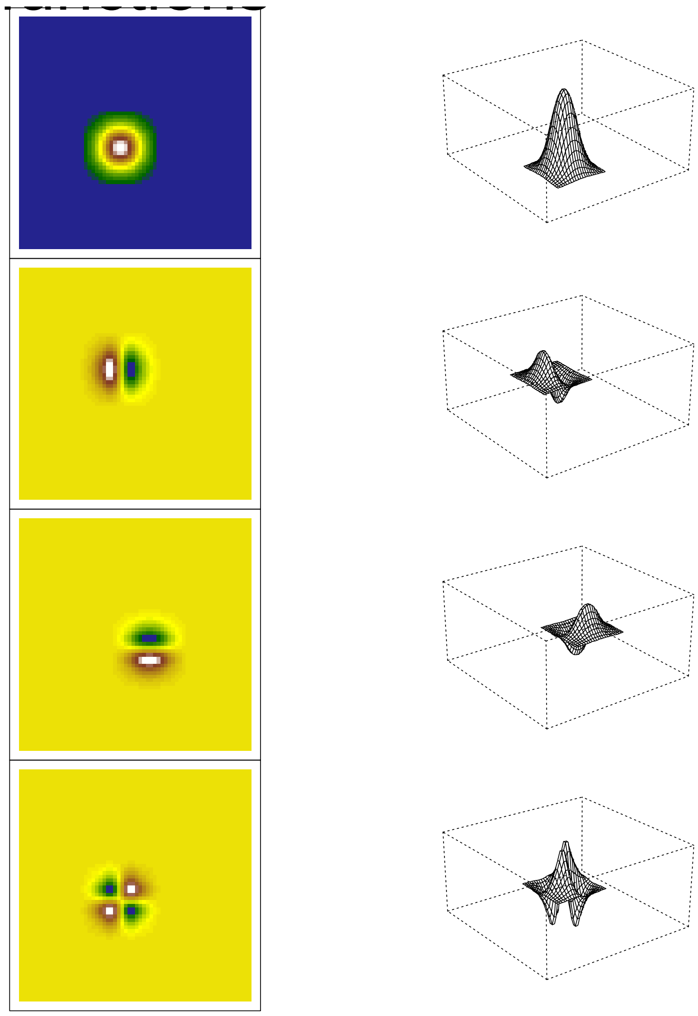

Figure 7: Examples of smooth, smooth-detail, detail-smooth and detail-detail basis functions associated with the two dimensional W transform.

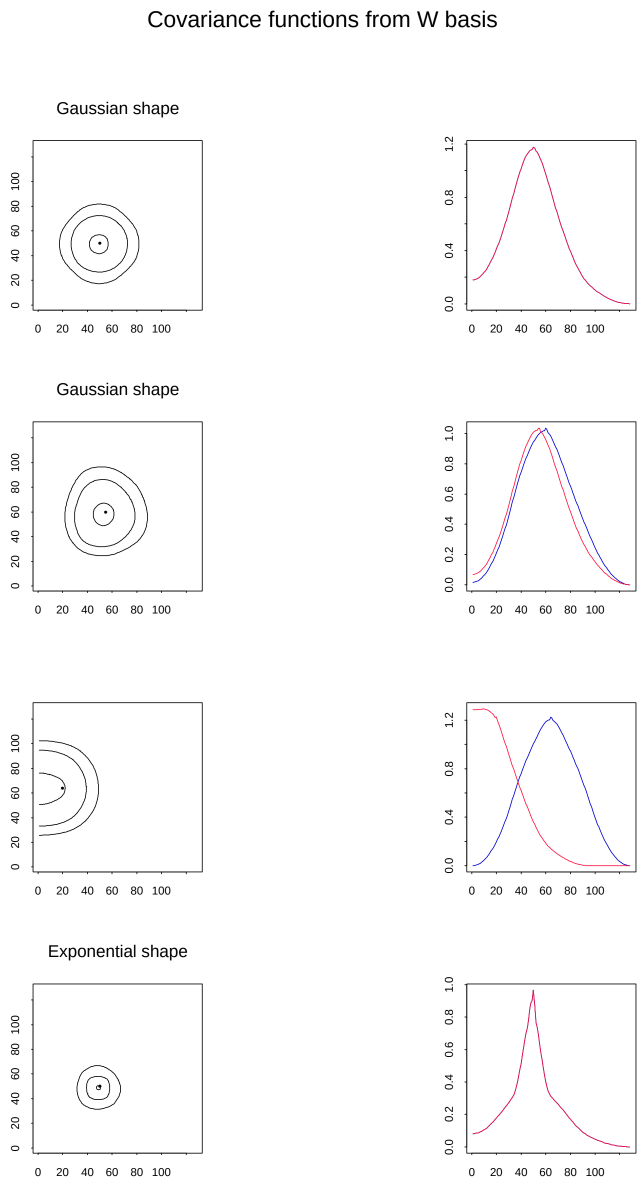

Figure 8: Covariances implied by W-transform basis functions using the variance weights from Table 1. The first column are the contours for the covariance function with respect to the location marked by a dot. The second column are vertical and horizontal slices of each of the surfaces through the marked location. Top three rows are for a model similar in shape to a Gaussian and the bottom row is closer to an exponential.

a covariance more similar in shape to an exponential.

Although a rigorous treatment of the covariance functions spanned by this basis is beyond the scope of this paper, such examples suggest that one can obtain approximate stationary, isotropic covariances with different shapes. A more ambitious test is to see if one can approximate the covariance function for a nonstationary field.

As an example of the potential to represent non-stationary fields, the two forms of covariances cited above are combined to produce an heterogeneous field. This is done by simply filling the right half of each block of coefficients with the variances for the Gaussian-like covariance and filling the left hand side of each block with the variances for the exponential-like covariance. Figure 4 is in fact a realization from this model and Figure 9 illustrates the nonstationarity of this covariance. The first plot in Figure 9 shows that the marginal variance is not constant. This probably due to the discrete nature of the translations of the basis functions. The second plot in this figure contains contours of equal correlation for a grid of different locations. The solid contour in the upper right hand corner of this plot represents the set of points that are correlated at .9 with location at the dot located at the corner. The next line out gives the points that are at the .5 level of correlation. The other contours in this plot are correlations for a grid of points indicated by the dots. This collection of contours may be confusing at first because it plots the covariance surface for more than one location on the same plot but provides a visual means for assessing nonstationarity. If the covariance function were stationary and isotropic then all contours would have the same shape and just be translates of each other. Moreover, if the process were isotropic the contours would be circles. In this example, the contours close to the middle of the region are distorted and reflect the blending between the two types of covariance models.

7 Reconstructed Precipitation Fields

As an example of the algorithms and covariance models presented in this paper precipitation fields are estimated over a section of North America including part of the Midwest and Rocky Mountains. The data record consists of approximately 1600 stations within the longitude/latitude rectangle in Figure 1 although the number of reporting stations in this period range from 701 to 823. Monthly values for the square root of precipitation for the period 1960 to 1989 are calculated for each location. In addition, for each location the transformed series is centered by subtracting the yearly cycle obtained by taking means for each month. Such values are known as anomalies as they represent deviation from the long term (30 year) mean for each month. The seasonal pattern of precipitation contains much of the spatial structure and so the anomalies have less spatial correlation than the raw values. Monthly precipitation is not strongly correlated in time and so this analysis will focus on just spatial aspects rather than a full space and time model. For the purposes of an example just the month of August is analyzed although it is straightforward to extend this work to the remaining 11 months. To apply the efficient algorithms presented above, the locations were discretized to a 128×128 grid. At this level of discretization approximately 10 (less than 2 %) of the grid boxes contained more than one observation. For these grid points the mean was used and the variance matrix, W was adjusted by the square root of the number of members in each box.

The analysis is based on two covariance models, a stationary, isotropic exponential covariance and one based on the multiresolution W transform basis. In each case, there is a linear spatial drift in the fixed part of the spatial model. The range for the exponential

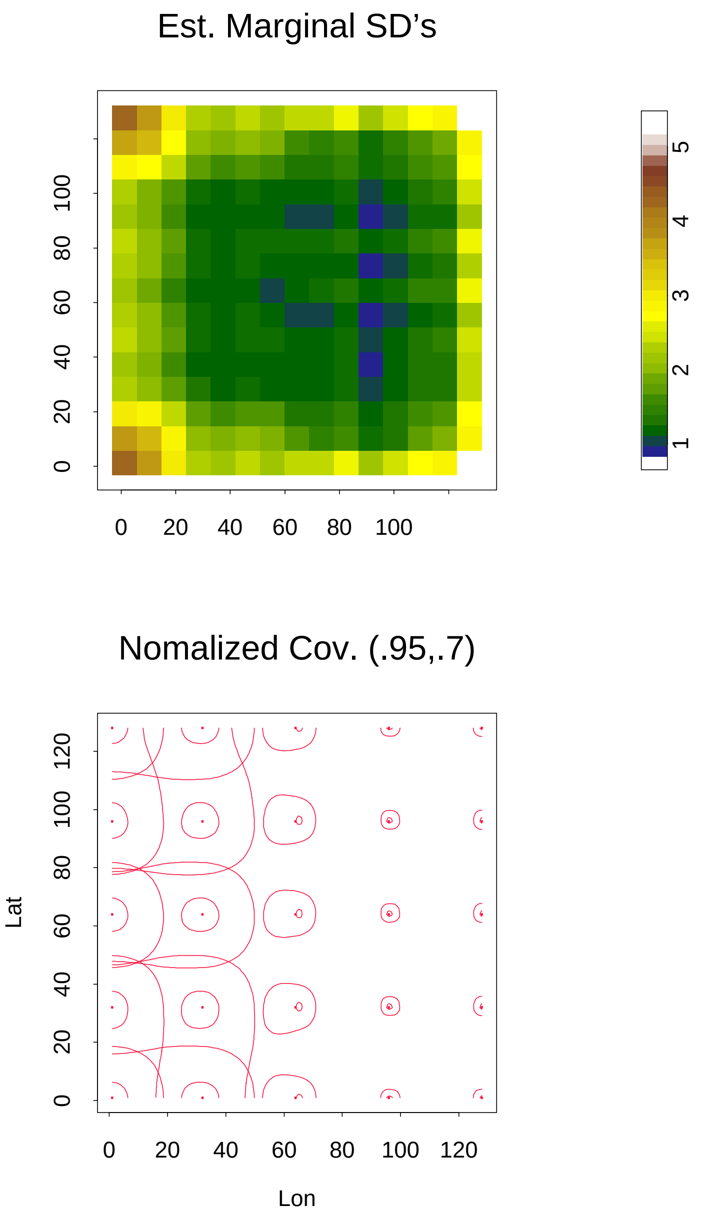

Figure 9: The margin variances and correlations associated with a nonstationary process. The nonstationarity is constructed by weighting the W transform basis functions differently on the right side of the image compared to the left.

covariance was identified from variograms of the fields and was fixed at 1 degree. The sill and the nugget variance (σ) were estimated at each of the 30 times by minimizing the prediction error from out of sample cross-validation. The 45 locations used for cross-validation are indicated on Figure 1 (as crosses) and were generated based on a space-filling design criterion (Nychka and Saltzman 1998). Estimates for the multiresolution covariance model rely strongly on the initial stationary covariance analysis. The estimated field for each time point is evaluated at the 128×128 grid using the stationary model. The W transform is computed for each field having first subtracted the spatial drift. Thus one obtains 30 sets of W-transform coefficients. Σ_a is assumed to be diagonal where the diagonal elements are estimated by the mean sums of squares for each coefficient over the realizations. Figure 10 gives an image plot of these variances and Figure 11 is a surface for the estimated marginal variance from this model and illustrates the correlation function. Here we see that the marginal variance has some spatial structure varying from 2 in the NW corner to over 3 in the eastern edge. The correlation function in contrast appears to be reasonably constant across the region.

As noted in section 3.2 the convergence of the CGA can be improved by preconditioning and choosing good starting vectors. Because the spatial correlations have limited scale the covariance matrices in this application are well conditioned and so using a diagonal matrix as the preconditioning matrix is adequate. The spatial mean only depends on the ratio of the sill and nugget variance parameters, the smoothing parameter and we will denote this ratio by λ. The value for λ is found by minimizing out of sample mean squared error at the locations indicated in Figure 1. As λ ranges over a grid from ∞ , fitting a just spatial drift, to 0, interpolation, the solution vector, ω_2 changes continuously. We start with the spatial drift estimate and then decrease λ. For each new value of the smoothing parameter the spatial mean is reestimated using the previous solution as the starting vector. This choice tends to accelerate convergence because smoothing parameters close to each other will have similar solutions.

The estimates for the stationary and nonstationary covariance models have similar out of sample accuracy. Figure 12 illustrates the variability associated with the spatial predictions and conditional mean field for a particular year and month. Here are plotted the conditional mean field for August, 1963 along with six realizations from the conditional distribution. The realizations are rougher than the conditional mean field and are more appropriate as forcing input for a vegetation model. Although the conditional mean fields for the stationary and nonstationary models are similar in this study. The standard errors associated with the predictions are different. Because of the size of this problem these conditional variances are found by simulation and the standard deviations in this figure are based on 100 replications. In these calculations simulating a mean zero field associated with the multiresolution covariance is exact and easy because both Σ_a^{-1} and Ψ^{-1} can be readily multiplied with a vector. To simulate the stationary field the region is doubled in size in both directions and a periodic spatial process is generated using the FFT. Because the range of correlation is much shorter than the size of the original domain this approximation is accurate and the implicit periodic boundary conditions do not effect the results. Figure 13 reports the prediction standard errors (or conditional standard deviations) for August, 1963 for the stationary and multiresolution model. For these computations we see that the prediction errors are smaller when a more realistic, i.e. nonstationary covariance, is used.

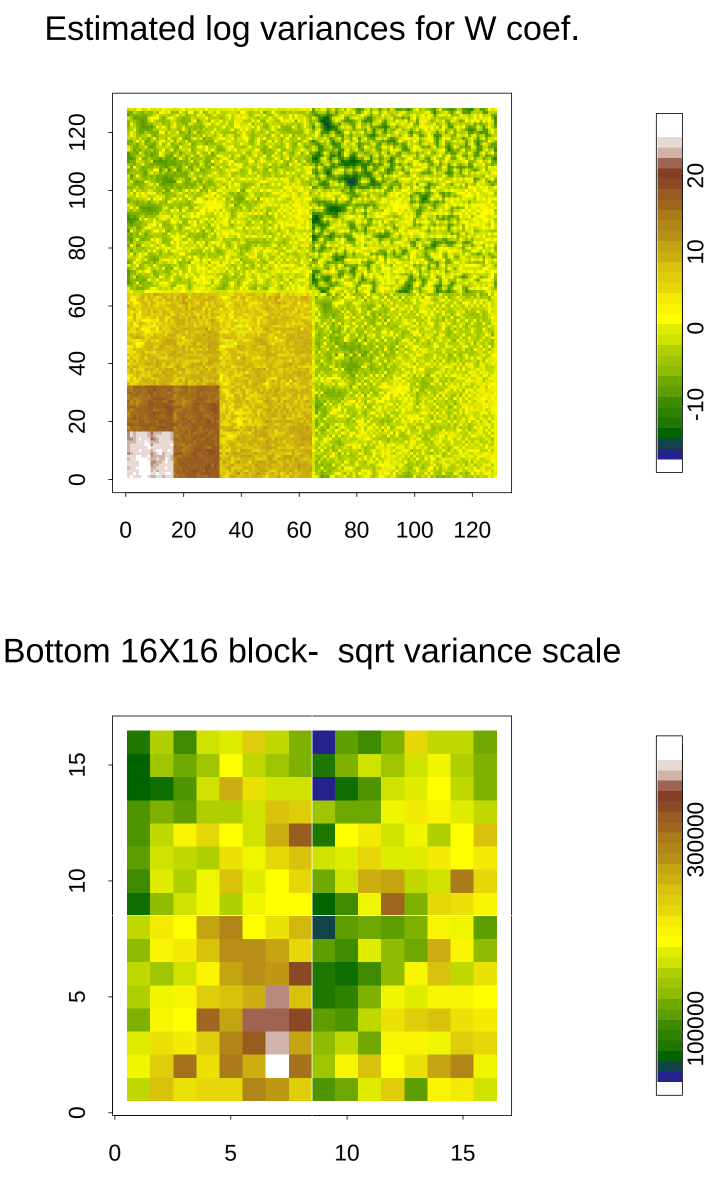

Figure 10: The variances of W transform coefficients estimated from stationary fields.

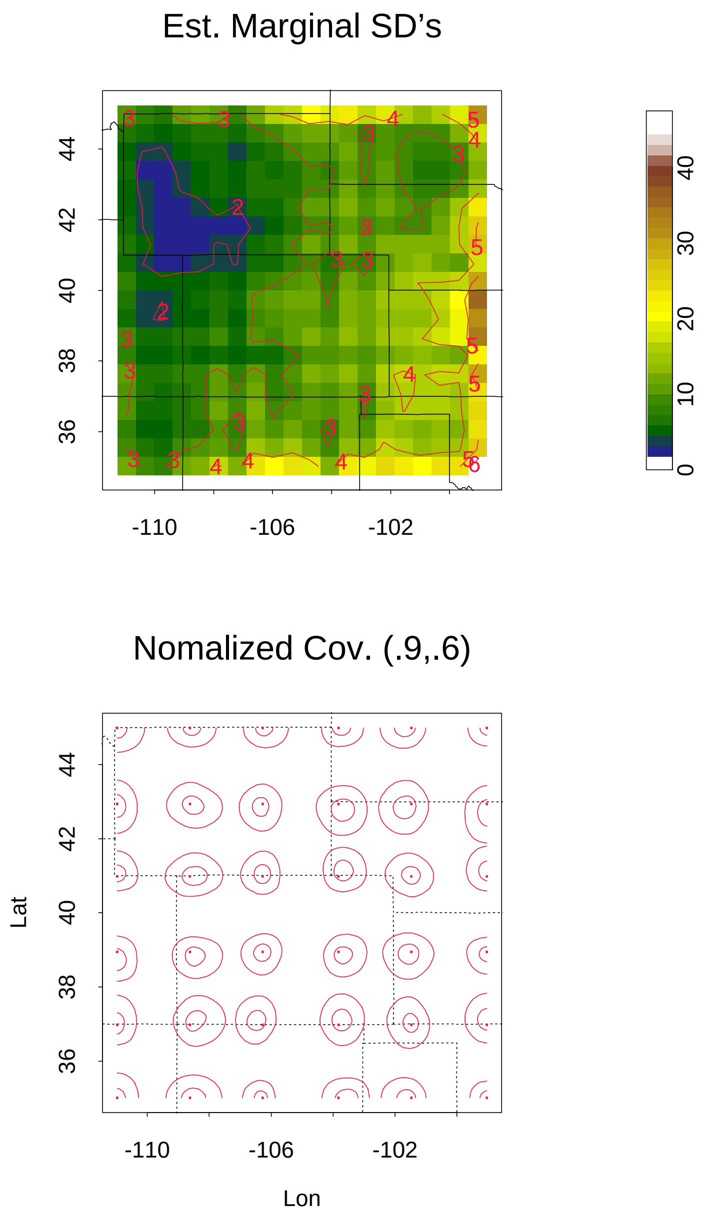

Figure 11: Top plot is the implied marginal standard deviations implied by the variances from Figure 10. Bottom plot are the implied correlation functions for the covariance model at a grid of points. Contours are drawn at .9 and .5.

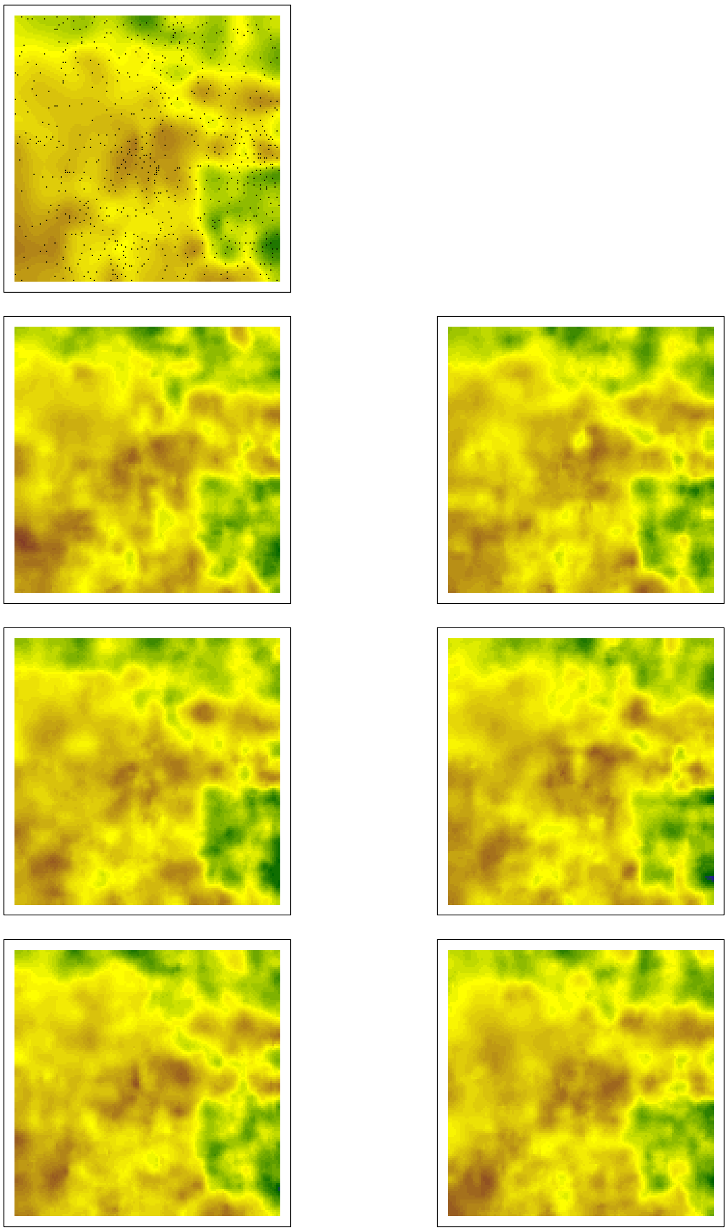

Figure 12: For August, 1963 using the nonstationary model, the conditional mean field and six realizations from the conditional distribution.

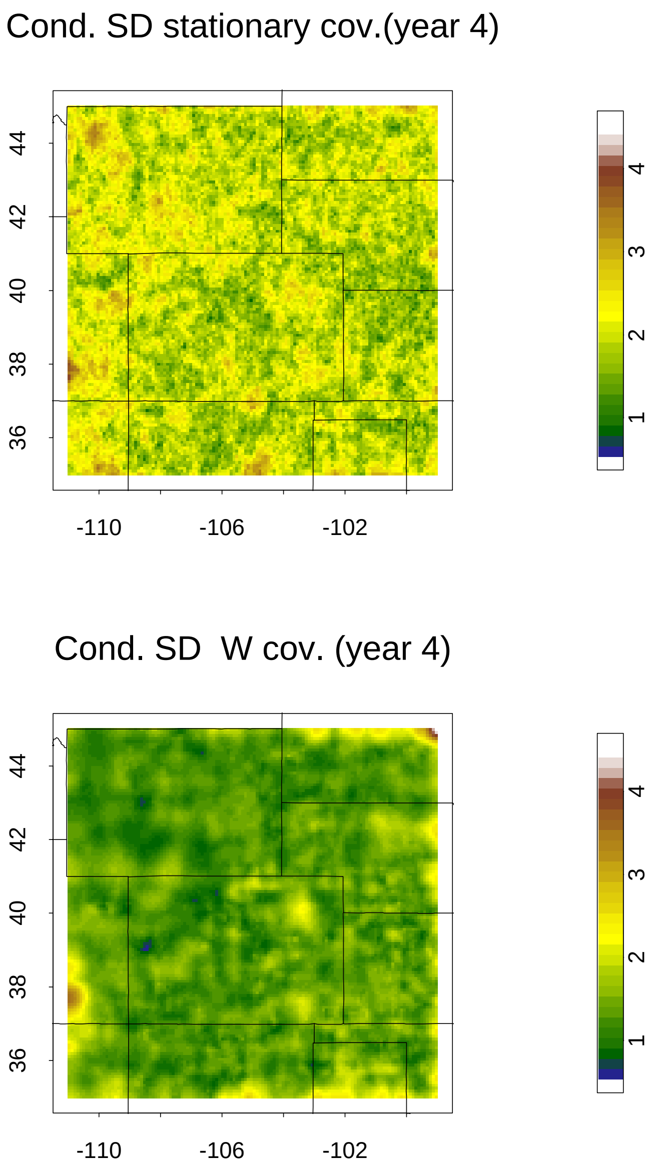

Figure 13: Prediction standard errors (Conditional standard deviations) for monthly precipitation, August, 1963. The stationary covariance model is given in the top and the W transform covariance is given in the bottom.

8 Discussion

This work combines multiresolution basis functions, the iterative solution of large linear systems and a discretization of the spatial locations to produce a practical algorithm for spatial prediction.

Although this computational approach is successful in handling large problems such as reconstructing precipitation fields it is still an open question to determine the classes of covariance functions spanned by different multiresolution bases. Also there are better methods for estimating and modeling the covariance matrix of the coefficients. In the precipitation example the covariance matrix has been assumed to be diagonal but there is some evidence for moderate correlation among the coefficients. This structure might be handled by a formulating a sparse version of Σ_a instead of assuming it is diagonal. The multiresolution basis functions are organized both by levels and by the locations of their support. This suggests using a Markov random field model to build covariances among coefficients at a given level and type. Another improvement of the covariance model is to constrain Σ_a so that it uses the largest eigenfunctions associated with the sample covariance matrix of the data. In this way the covariance can adapt to large scale nonstationary modes that are indicated empirically.

A more challenging problem is modeling precipitation fields to account for the non Gaussian distribution of the measurements. Any deviation from a Gaussian assumption will involve likelihoods that do not yield linear estimators. However, this work is still important in that most nonlinear estimators can be found using a iteratively reweighted sequence of linear calculations. Thus the algorithms developed will serve as the core computation within a larger iterative scheme.

Acknowledgments

This work was supported by National Science Foundation grant DMS-93122686.

References

Cressie, N.A.C. (1991). *Statistics for Spatial Data.* Wiley, New York

Golub, G. and Van Loan, C. (1989). *Matrix Computations.* John Hopkins University Press. Baltimore.

Haas, T. (1995)"Local Prediction of a spatio-temporal process with an application to wet sulfate deposition. " *Journal of the American Statistical Association*, 90 , 1189-1199.

Higdon, D. (1997) " A Process-Convolution Approach to Modeling Temperatures in the North Atlantic Ocean " *INSTITUTE OF STATISTICS AND DECISION SCIENCES Discussion Paper Series* Duke University, 97-16.

Kwong, M.K. and Tang, P.T.P. (1994) "W-Matrices, Nonorthogonal Multiresolution Analysis, and Finite Signals of Arbitrary Length", Technical Report Argonne National Laboratory, MCS-P449-0794

Nason, G.P. and Silverman, B.W. (1994). The discrete wavelet transform. *Journal of Computational and Graphical Statistics*, **3**, 163-191.

Nychka, D. (1998). Spatial Process Estimates as Smoothers. *Smoothing and Regression. Approaches, Computation and Application* ed. M. G. Schimek, Wiley, New York. to appear.

Press, W. , Teukolsky, S. Vettering, W. Flannery, B. (1992) *Numerical Recipes in C.* Cambridge University Press, Cambridge, UK.

Nychka, D. and Saltzman, N. (1998). Design of Air Quality Monitoring Networks. in *Case Studies in Environmental Statistics* Nychka, D., Cox, L., Piegorsch, W. ed., Lecture Notes in Statistics, Springer Verlag

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

Tibshirani, R. (1996). Regression selection and shrinkage via the *lasso. Journal of the Royal Statistical Society, series B* **58**, 267–288.

Welch, W.J., Buck, R.J., Sacks, J., Wynn, H.P., Mitchell, T.J., and Morris, M.D. (1992). Screening, predicting and computer experiments. *Technometrics.* **34**, 15-25.

Wikle, C.K., R.F. Milliff, D.W. Nychka, and L.M. Berliner, (1998) "Spatio-temporal hierarchical Bayesian blending of tropical ocean surface wind data". *Journal of the American Statistical Association* (submitted).