Chapter 5

Spatial Model Results

5.1 Introduction

In the previous chapter, I demonstrated that nonstationary covariance functions could be used in a Gaussian process (GP) prior for regression functions. In the regression setting, a single set of observations is observed, and the nonstationary covariance is inferred based on the similarity between responses at nearby locations. This neighborhood-based inference is the same type of inference done in time series analysis in the standard case in which there is only one time profile. Recall that in Chapter 4 the nonstationary covariance was the prior covariance for the regression function.

In spatial analysis, multiple replicates of a field of data (usually taken over time) are often available, and we may be interested in using the replication to model the covariance structure of the data. Ideally, we would have enough data to estimate the covariance by maximum likelihood as a simple averaged cross-product of residuals. However, for the maximum likelihood (ML) estimate of the covariance to be non-singular, we need at least as many replicates as we have observations (locations) within each replicate. Even when this many replicates are available, a smoothed estimator of the covariance is likely to perform better than the unsmoothed ML estimator.

In such cases, a covariance model is needed, and in many settings a stationary covariance may not be sufficient. This appears to be the case for the storm activity data analyzed in Paciorek, Risbey, Ventura, and Rosen (2002). As we see in Figure 5.1, the residual spatial correlation structure at two different locations appears to have very different correlation scale, which indicates nonstationarity in the correlation structure. In this chapter, I use the nonstationary covariance model described in Chapter 2 and developed further in Chapter 3 as a model for the residual spatial correlation in a hierarchical space-time model.

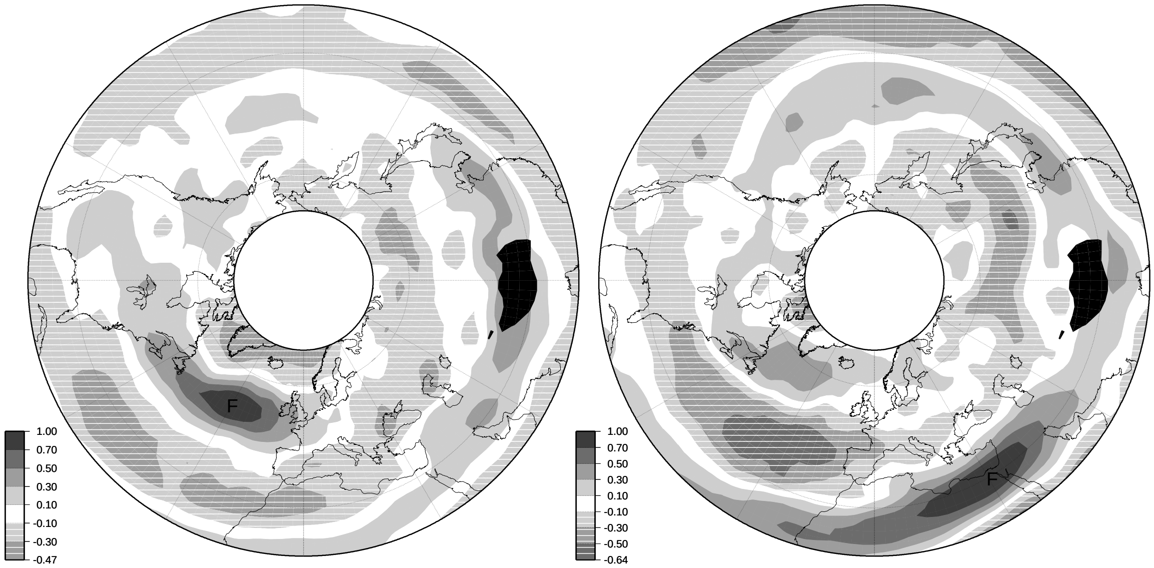

Figure 5.1. Plots of residual correlation of storm activity between all locations and each of two focal locations, each marked with an 'F': (left) $50°$ *N,* $330°$ *E (in the North Atlantic) and (right)* $30°$ *N,* $30°$ *E (in Egypt). The residual correlation is calculated after removing location-specific linear trends. Gray shading indicates the value of the correlation coefficient, with darker colors indicating values large in magnitude. Negative values are indicated by horizontal hatching. The high topography of the Himalayas is blacked out.*

5.2 Scientific Problem

In recent decades, atmospheric science and climatology have become much more prominent, in part because of advances in weather prediction and demand for meteorological information as an important economic product, and in part because of increasing scientific and public concern about climate change induced by anthropogenic changes in atmospheric greenhouse gas concentrations. One aspect of climate change receiving recent attention has been the possibility of changes in extreme events such as hurricanes, tornados, and winter storms. In Paciorek et al. (2002), we analyzed several indices of Northern Hemisphere extratropical winter cyclonic storm activity. Of particular interest was the possibility of long-term changes in storm activity over the 51 years for which we had data (1949-1999). Based on time series analysis, we concluded there was little evidence for temporal autocorrelation, probably because we calculated winter averages, and correlation does not seem to occur from year to year. Since standard linear regression seemed appropriate, we calculated the maximum likelihood linear time trend estimates based on the 51 years and mapped these estimates. To assess the reliability of the estimates in the face of simultaneously testing 3024 locations, we used the False Discovery Rate (FDR) approach introduced by Benjamini and Hochberg (1995). We found that for most of the indices of storm activity, there was significant evidence of trend at some locations but no significant evidence at many locations. Some questions arose as to which version of the FDR methodology to use in the face of correlation between test statistics induced by the spatial correlation of the data. In a separate project, we have been assessing this question in a simulation study of several versions of the methodology (Ventura, Paciorek, and Risbey 2003). The goal of this chapter is to further assess the trends in storm activity by building and fitting a Bayesian model of the time trends that properly accounts for the uncertainty induced by variability in both time and space.

The effect of variability in time on the trend estimates is simple to understand. We fit linear regressions for storm activity as a function of year at each location. Temporal variability causes uncertainty in our estimation of the underlying trends. Analyzing a single location, this uncertainty can be assessed easily in the usual fashion, either classically or from a Bayesian perspective. The difficulty comes in incorporating spatial correlation into a joint estimate of the time trends. The goal is to borrow strength across locations to better estimate the time trend spatial field and its uncertainty in a cohesive fashion. One reason that using spatial information is particularly important is that the data themselves are quite smooth spatially, as are the ML trends (as well as the intercepts and residual variances) at each location. There are two main reasons for this. One is that storm activity is spatially correlated; storms are spatially coherent phenomena that move through the atmosphere, generally along certain preferred paths, called storm tracks. The result is that locations that are close in space tend to have similar amounts of activity because they are affected by the same storms. The second reason for correlation in the data is that the data we use are the output of a meteorological data assimilation model (the NCEP-NCAR reanalysis model (Kalnay and Coauthors 1996)) that takes raw weather observations (temperature, wind, pressure, etc.) combines them with a deterministic meteorological model, and estimates the values of weather variables on a regular latitude-longitude grid ($2.5^\circ \times 2.5^\circ$). This assimilation process induces spatial correlation in the variables and in the storm activity indices calculated from these variables.

A worst case scenario for the trend analysis is that there are no true trends in time, but that temporal variability in the data in time cause apparent trends, and the spatial correlation makes these false trends seem more reliable by causing them to be spatially coherent. The key issue lies in determining the extent to which the spatial correlation causes spatially smooth real trends as opposed to spatially smooth noise. The goal of the spatial model that I construct in the next section is to embed the individual linear time trend models in a model for the spatial correlation structure of the data, so that the estimation of the trends is done in a manner that accounts for the spatial structure of the data. There are three main aims; the first is to get a better point estimate of the trend activity by borrowing strength across locations. The second is to come up with uncertainty estimates for the trends that account for the spatial correlation structure. Hopefully the results of the model will allow us to obtain a more complete picture of the trends and their reliability to compare to the results of the classical testing analysis. Separating trend from spatially smoothed noise is essentially the same problem faced in time series analysis in separating signal from correlated noise. Luckily, in the case of these spatial data, we have repeated observations, the multiple years, with which to make headway. The third aim is to evaluate several covariance models, stationary and nonstationary, to see which is the best model for the data and therefore which models might be preferred for similar climatological data.

Performing spatial smoothing and estimating trends at unobserved locations are not primary goals of this analysis for several reasons. First, as mentioned above, the data are already quite smooth, so further smoothing seems of limited importance, and since the latitude-longitude grid is already fairly dense, smooth maps can be produced with little effort directly from the unsmoothed data with little need for spatial interpolation of either data or estimated trends. Second, by working only with observed locations, I have a simple baseline ‘model’ for the data, namely the predictions from the ML trends and intercepts at the locations. I can compare the results from the various covariance models to this simple baseline model, which is not possible if I estimate trends at held-out locations. Finally, most of the variability in the data is accounted for by residual variation about the trends, not by the trends themselves. Therefore, the main driver behind the accuracy in predicting unobserved locations will be the the model's performance in spatially smoothing the residuals from the temporal model, not the model's performance in smoothing the trends. Hence, assessment of model performance with respect to the main scientific problem of trend estimation is better addressed by cross-validating in time (holding out time points) rather than by cross-validating in space (holding out locations). For other datasets, spatial prediction might be an important goal, and cross-validating in space would be an important means of evaluation. Prediction on unobserved years does not fully address the question of model evaluation, but I believe it does the best possible job under the constraints imposed by these data.

In addition to the empirical evidence for nonstationarity discussed briefly above, basic climatology also suggests that nonstationarity would be present because of the morphology of storm activity. Winter storm activity tends to follow certain preferred paths, called storm tracks. The correlation structure in these storm track areas is likely to be different from the correlation structure in areas far from the storm tracks with their differing climatology. Other important aspects of storm activity are that the correlation scale in the west-east direction is likely to be longer than the correlation scale in the north-south direction because the storm tracks tend to be oriented west-east. Also, storm tracks can shift in time, so there may be strong negative correlations between locations. For example negative correlations may occur between locations at the same longitude but separated by latitude, since if a storm track shifts in the north-south direction, storms that would have hit one location instead hit the location at the other latitude.

While many analyses have modelled spatiotemporal structure, few have focused on including spatial context in assessing long-term time trends at multiple locations. Holland et al. (2000) extensively analyzed atmospheric concentrations of sulfur dioxide, estimating 35 location-specific trends after accounting for covariates. Using kriging methods, they smoothed the raw trend estimates. They used the jackknife to estimate both diagonal and unstructured covariances of the unsmoothed trend estimates. They found that accounting for correlation between the location-specific estimates using the unstructured covariance better fit the data than the diagonal covariance and resulted in small changes in regional trend estimates and increases in the standard errors of these regional trends. Oehlert (1993) discusses but does not fit (because of insufficient data) a model in which linear time trends of wet sulfate deposition are embedded in a spatiotemporal model and the estimated residual spatial correlation is used in making inference about the trends. Wikle, Berliner, and Cressie (1998) propose a spatiotemporal model in which location-specific temporal models are parameterized by spatial processes with Markov random field priors. This chapter represents a Bayesian effort to fully account for the residual spatial structure in assessing time trends.

Methods that accomplish these goals are needed. Spatio-temporal data for which the scientific questions of interest involve analysis of changes over time are common. They are of obvious importance in assessing scientific and public policy questions about the evolving state of natural systems of various kinds, from climate stability to changes in biological populations and environmental phenomena of various sorts.

5.3 Basic Spatial Model

The basic spatial model builds on the location-specific models used in Paciorek et al. (2002). In that work, we fit simple linear regressions of storm activity against time,

$$Y_t \sim \mathrm{N}(\alpha + \beta t, \eta^2).$$

In this analysis, I use the same local model, but tie the residuals, now indexed by location i, $Y_{it} - \alpha(\boldsymbol{x_i}) - \beta(\boldsymbol{x_i})t$, together in space using various covariance models. I model $\alpha(\cdot)$ and $\beta(\cdot)$ a priori as Gaussian processes. The joint likelihood for a single time, $t \in \{-25, \cdots, 25\}$, is taken to be

$$\boldsymbol{Y}_t \sim \mathrm{N}(\boldsymbol{\alpha} + \boldsymbol{\beta} t, C_{\boldsymbol{Y}}), \tag{5.1}$$

and $\boldsymbol{Y}_t|\boldsymbol{\alpha}, \boldsymbol{\beta}$ is taken to be independent of $\boldsymbol{Y}_s|\boldsymbol{\alpha}, \boldsymbol{\beta}$ for $t \neq s$, so that $C_{\boldsymbol{Y}}$ is not a function of time. This is justified based on an exploratory analysis of the correlation structure of the data, which indicated that for winter-long averages of the storm indices, there is little correlation from year to year. The covariance matrix of the observations is taken to be

$$C_{\boldsymbol{Y}} = D(\boldsymbol{\eta}) R_{\boldsymbol{Y}} D(\boldsymbol{\eta}) + \delta I,$$

where $D(\boldsymbol{\eta})$ is a diagonal matrix of standard deviations, $R_{\boldsymbol{Y}}$ is a correlation matrix, and δ is the variance of location-independent noise. In the analysis I compare several models for $R_{\boldsymbol{Y}}$, including a stationary model, kernel-based nonstationary model, and smoothed versions of the empirical correlation structure; description of these models is deferred to Section 5.4. A directed acyclic graph (DAG) of the model, with the nonstationary parameterization for $R_{\boldsymbol{Y}}$, is shown in Figure 5.2.

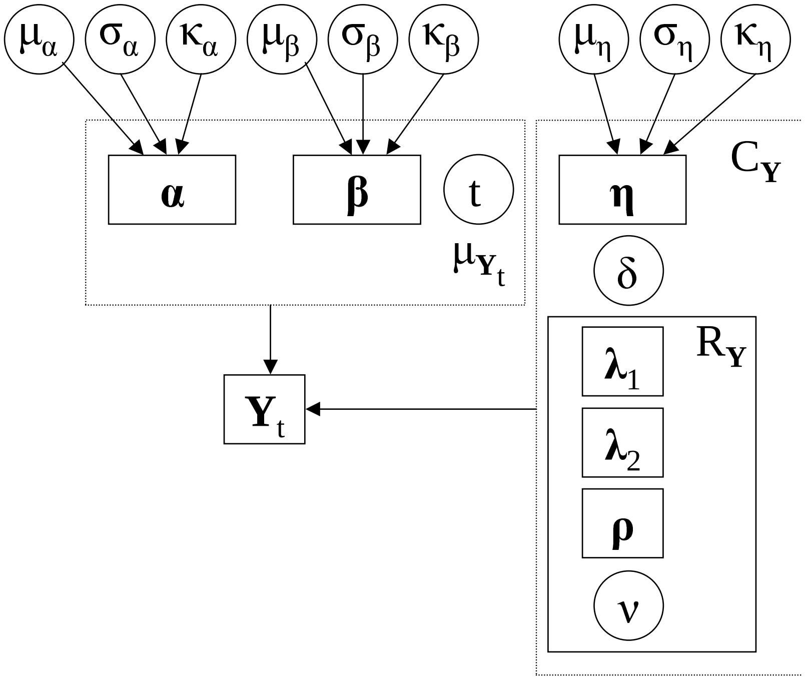

Figure 5.2. Directed acyclic graph of nonstationary spatial model. Bold letters indicate vectors..

The mean function parameters and the residual variance parameters are taken to be spatial fields. For $\phi(\cdot) \in \{\alpha(\cdot), \beta(\cdot), \log \eta(\cdot)^2\}$, we have

$$\phi(\cdot) \sim \text{GP}\left(\mu_\phi, \sigma_\phi^2 R_\phi^S(\cdot;\kappa_\phi,\nu_\phi)\right),$$

where $R_\phi^S(\cdot)$is the stationary Matérn correlation function with scale parameter κ_ϕ and fixed smoothness parameter, $\nu_\phi = 4.0$. This value of ν_ϕ reflects my belief that the parameter processes vary smoothly in space. The atmospheric phenomena that generate storms are based on masses of air and energy that move through the atmosphere. Furthermore, the data, Y_{it}, are winter-long averages; this averaging should increase the smoothness. For $\nu_\phi = 4$ we have $\lceil \nu_\phi - 1 \rceil = 3$ sample path derivatives of the $\phi(\cdot)$ processes. I fix ν_ϕ because I don't believe the data can reasonably inform this parameter based on the model structure and coarseness of the data grid. Use of this value does limit the sample path differentiability (Section 2.5.5) of the underlying spatial residual processes.

To use the Matérn function on the sphere, I transform angular distance to Euclidean distance in $\Re^3$, calculating the chordal distance, $\tau = \sin\left(\frac{\rho}{2}\right)$, where ρ is angular distance. As a function of chordal distance, the usual Matérn function defined by

$$R(\tau) = \frac{1}{\Gamma(\nu)2^{\nu-1}}\left(\frac{2\sqrt{\nu}\tau}{\kappa}\right)^{\nu} K_{\nu}\left(\frac{2\sqrt{\nu}\tau}{\kappa}\right)$$

is a legitimate correlation function. While the chordal distance underestimates the true distance, I believe it is a reasonable distance measure for pairs of locations close enough in proximity to have non-negligible correlation.

One might argue on principle with my use of stationary priors for $\alpha(\cdot), \beta(\cdot)$, and $\log \eta(\cdot)^2$ based on the climatological evidence I discuss for why nonstationary models are more reasonable for the actual observations. However, using nonstationary priors for these fields is not practical because of computation limitations. Adding the necessary parameters high in the model hierarchy will make it even harder for the model to mix adequately. In practice, stationary priors may be sufficient, because the posterior covariance of these fields incorporates the data covariance model and will therefore be nonstationary if the data covariance model is nonstationary. As more data are collected, the posteriors for the fields will be less and less influenced by their stationary priors. For $\boldsymbol{\phi} \in \{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$ we can see this in closed form. Conditional on the other parameters, represented as $\boldsymbol{\theta}$, including the variance vector $\boldsymbol{\eta}$, these vectors are a posteriori normal with moments,

$$\begin{aligned} \mathrm{E}\boldsymbol{\phi}|\boldsymbol{Y}, \boldsymbol{\theta} &= C_{\boldsymbol{\phi}}\left(C_{\hat{\boldsymbol{\phi}}} + C_{\boldsymbol{\phi}}\right)^{-1}\hat{\boldsymbol{\phi}} + C_{\hat{\boldsymbol{\phi}}}\left(C_{\hat{\boldsymbol{\phi}}} + C_{\boldsymbol{\phi}}\right)^{-1}\mu_{\phi} \\ &= \mu_{\phi} + C_{\boldsymbol{\phi}}\left(C_{\hat{\boldsymbol{\phi}}} + C_{\boldsymbol{\phi}}\right)^{-1}(\hat{\boldsymbol{\phi}} - \mu_{\phi}) \qquad (5.2) \\ \mathrm{Cov}(\boldsymbol{\phi}|\boldsymbol{Y}, \boldsymbol{\theta}) &= \left(C_{\hat{\boldsymbol{\phi}}}^{-1} + C_{\boldsymbol{\phi}}^{-1}\right)^{-1}, \qquad (5.3) \end{aligned}$$

where $\hat{\boldsymbol{\phi}}$ is the vector of MLEs for the field, and $C_{\hat{\boldsymbol{\phi}}}$ is the variance matrix of $\hat{\boldsymbol{\phi}}$. For $\boldsymbol{\alpha}$ we have $\hat{\alpha}_i = \frac{\sum_t Y_{it}}{T}$and $C_{\hat{\boldsymbol{\alpha}}} = \frac{C_{\boldsymbol{Y}}}{51}$, while for $\boldsymbol{\beta}$ we have $\hat{\beta}_i = \frac{\sum_t tY_{it}}{\sum_t t^2}$ and $C_{\hat{\boldsymbol{\beta}}} = \frac{C_{\boldsymbol{Y}}}{\sum_t t^2}$ with $C_{\boldsymbol{Y}} = D(\boldsymbol{\eta})R_{\boldsymbol{Y}}D(\boldsymbol{\eta}) + \delta I$ as defined previously. So with a nonstationary correlation model, $R_{\boldsymbol{Y}}$, the posterior specified by (5.2-5.3) is nonstationary for $\boldsymbol{\phi}$. With little data, the prior will dominate and the data will be smoothed in a generally stationary fashion, but with more and more data, the denominators of $C_{\hat{\boldsymbol{\phi}}}$ will increase, driving the posterior variance down and causing the posterior mean to be more and more influenced by the MLEs. The smoothing will then be primarily nonstationary in structure, based on the model-inferred data covariance.

It is possible to integrate both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ out of the model. Doing so produces the marginal likelihood

$$\boldsymbol{Y}_t|\boldsymbol{\theta} \sim \mathrm{N}(\mu_\alpha + \mu_\beta t, C_{\boldsymbol{Y}} + C_{\boldsymbol{\alpha}} + t^2 C_{\boldsymbol{\beta}}),$$

which for computational efficiency can be expressed in an alternate form (not shown) that requires inverting only $C_{\boldsymbol{Y}}, (C_{\hat{\boldsymbol{\alpha}}} + C_{\boldsymbol{\alpha}})$, and $(C_{\hat{\boldsymbol{\beta}}} + C_{\boldsymbol{\beta}})$. To generate samples of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, the marginal distributions are

$$\begin{aligned}
\boldsymbol{\alpha}|\boldsymbol{Y},\boldsymbol{\theta} &\sim \mathrm{N}\left(\mu_\alpha + C_{\boldsymbol{\alpha}}\left(C_{\boldsymbol{\alpha}} + C_{\hat{\boldsymbol{\alpha}}}\right)^{-1}\left(\hat{\boldsymbol{\alpha}} - \mu_\alpha\right), C_{\boldsymbol{\alpha}}\left(C_{\boldsymbol{\alpha}} + C_{\hat{\boldsymbol{\alpha}}}\right)^{-1} C_{\hat{\boldsymbol{\alpha}}}\right) \\
\boldsymbol{\beta}|\boldsymbol{Y},\boldsymbol{\theta} &\sim \mathrm{N}\left(\mu_\beta + C_{\boldsymbol{\beta}}\left(C_{\boldsymbol{\beta}} + C_{\hat{\boldsymbol{\beta}}}\right)^{-1}\left(\hat{\boldsymbol{\beta}} - \mu_\beta\right), C_{\boldsymbol{\beta}}\left(C_{\boldsymbol{\beta}} + C_{\hat{\boldsymbol{\beta}}}\right)^{-1} C_{\hat{\boldsymbol{\beta}}}\right).
\end{aligned}$$

For the results reported in this thesis, I sampled from the full model, choosing not to integrate $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ out of the model. This is primarily because while I was developing the model, I was interested in devising a sampling scheme that would not require marginalization because I was interested in non-Gaussian likelihoods. Also, sampling $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ conditionally on the chain for the marginalized model still involves a great deal of calculation and in the end, these are the parameters I am most interested in, so I do need to sample them, although integrating them out would allow me to avoid sampling them at every iteration of the chain. While it is conceivable that integrating out the intercepts and trends would make it easier to sample the remaining parameters, the fact that most of the variability in the data is related to the residual variance and not to the mean structure suggests this will not be the case.

The spatial model defined in this section has obvious similarities with the regression models in Chapter 4, except that here I have built a hierarchical model based on the components used in the regression modelling, and the nonstationary covariance is used to model the residual structure. In fact, we can reexpress the model in a way that makes clear that there is an underlying nonstationary GP prior for functions, just as in the regression modelling. The model is

$$\boldsymbol{Y}_t \sim \mathrm{N}(\boldsymbol{f}_t, \delta I)$$

$$\begin{aligned} \boldsymbol{f}_t &\sim \mathrm{N}\left(\boldsymbol{\mu}_t, D(\boldsymbol{\eta})R_{\boldsymbol{Y}}D(\boldsymbol{\eta})\right) \\ \boldsymbol{\mu}_t &= \boldsymbol{\alpha} + \boldsymbol{\beta}t, \end{aligned} \tag{5.4}$$

with each function, $f_t(\cdot)$, sharing common correlation, $R_{\boldsymbol{Y}}(\cdot,\cdot)$ and variance $\eta(\cdot)^2$ functions, but differing in their mean function, $\mu_t(\cdot)$. In contrast to the regression modelling, in which the regression function has constant mean, μ, and variance, σ^2, hyperparameters, here we have the hyperparameters $\mu_t(x) = \alpha(x) + \beta(x)t$ and $\eta(x)^2$ that vary over the covariate space and themselves have GP priors. I avoid having to sample $\boldsymbol{f}_t$ by integrating them out of the model, which gives the model as originally stated (5.1).

5.4 Covariance Modelling

The previous section did not specify a correlation model for the data. There is a limited literature on fitting spatial covariance models to replicated data of the sort that I use here. In this section, I describe some possible correlation models that could be used, including a stationary model, and discuss potential advantages and disadvantages of the models. I choose three models to compare in this thesis and describe in detail how the three are parameterized. Recall that the covariance of the data is modelled as

$$C_{\boldsymbol{Y}} = D(\boldsymbol{\eta})R_{\boldsymbol{Y}}D(\boldsymbol{\eta}) + \delta I.$$

This construction leads me to focus on models for the correlation structure, since I treat the variance structure separately.

5.4.1 Correlation models used in this work

5.4.1.1 Independence model

The simplest model takes $R_{\boldsymbol{Y}} = I$ and estimates trends for each location independently, thereby disregarding the spatial correlation of the data. A second baseline model takes $R_{\boldsymbol{Y}} = I$ but assumes $\boldsymbol{\beta} = \boldsymbol{0}$, the joint null hypothesis at all locations.

Unfortunately, I can't use the MLE of the residual covariance $\hat{C}_{\boldsymbol{Y}} = Y^{*}Y^{*T}$ where Y^{*} is a matrix of the standardized residuals, because Y^{*} has T columns, one for each year of data, and therefore, $\hat{C}_{\boldsymbol{Y}}$, has rank $T-2$, based on estimating $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$. This situation is not unusual with climatological and other geophysical data, in which data are available at more locations than time points. The singularity of $\hat{C}_{\boldsymbol{Y}}$ prevents me from using it in the predictive distribution calculations, which involve the inverse of the covariance matrix, used to compare models, as discussed in Section 5.6.

5.4.1.2 Stationary Matérn correlation

The simplest way to include the covariance structure of the data in a model is to use a stationary covariance model. While examination of the data suggests that stationarity is unlikely, it may be that a stationary model fits the data adequately and that the additional complexity of modelling the nonstationarity is not warranted. One drawback to the stationary model, which seems to be realized in the storm data, is that in regions in which the stationary correlation does not fit well, the residual variance portion of the covariance model, $\boldsymbol{\eta}^2$, is driven up relative to the MLE variances for the locations. It seems undesirable that the modelled variance be much larger than is reflected in the empirical variability about the fitted time trend, merely because of lack of fit in the correlation model. A more sophisticated approach would be to embed the stationary model in the deformation model of Sampson and Guttorp (1992), possibly including the deformation in a fully Bayesian fashion (Schmidt and O'Hagan 2000; Damian et al. 2001), but I have not used the deformation approach here.

I use a stationary Matérn model, with the transformation of angular distance to chordal distance in $\Re^3$ and a uniform prior on $\nu_{\boldsymbol{Y}} \in [0.5, 15]$. Note that for simplicity I had initially fixed $\nu_{\boldsymbol{Y}} = 2$ but this resulted in fitted correlations and variances not consistent with the data and different results than using fixed $\nu_{\boldsymbol{Y}} = 4$. This suggests that it is better to allow the model to choose the value of ν, as I have done here.

5.4.1.3 Nonstationary smoothed empirical correlation

Nychka et al. (2001) introduced a wavelet-based method for smoothing the empirical covariance of data that lie on a regular grid. As an alternative to the kernel-based nonstationary covariance described next, I use this smoothing method on the residuals, $\boldsymbol{U}_t = \boldsymbol{Y}_t - \hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\beta}}t$, of the storm activity data obtained by subtracting the ML trends from the data. To explain the wavelet method, first consider the eigendecomposition of a positive definite matrix. The decomposition expresses the empirical covariance as $\hat{C} = \Gamma\Lambda\Gamma^T$, based on the matrix of eigenvectors, Γ, and a diagonal matrix of eigenvalues, Λ. If we consider the dual space in which each year of data is produced as a linear combination of basis functions, we have $\boldsymbol{U}_t = \Gamma\boldsymbol{\omega}_t$, with $\boldsymbol{\omega}_t \sim \mathrm{N}(0, \Lambda)$. Instead of the eigendecomposition, the wavelet method uses the W wavelets, which are a nonorthogonal wavelet basis whose functions are piecewise quadratic splines. Denoting the basis matrix as Ψ, each year of residuals can be expressed as $\boldsymbol{U}_t = \Psi\boldsymbol{\omega}_t$, so collecting the vectors $\boldsymbol{\omega}_t$ into a matrix Ω we have $\hat{C} = \frac{1}{T}UU^T = \frac{1}{T}\Psi\Omega\Omega^T\Psi^T = \Psi\hat{\Lambda}\Psi^T$ where the empirical covariance matrix of the coefficients, $\hat{\Lambda}$, is no longer diagonal because we are not using the eigendecomposition. To smooth the empirical covariance instead of reproducing it exactly, the method thresholds the off-diagonal elements of a square root matrix, $\hat{H}$, of $\hat{\Lambda}$ and reconstructs a smoothed version of the original empirical covariance, $\tilde{C} = \Psi\tilde{H}\tilde{H}^T\Psi^T = \Psi\tilde{\Lambda}\Psi^T$. The calculations are very fast because $\hat{\Lambda}$ can be calculated using the discrete wavelet transform, as can elements of $\tilde{C}$ once $\tilde{\Lambda}$ is computed. Calculating $\hat{H}$ is slow $O(n^3)$ if one naively takes the square root of the n by n matrix, $\hat{\Lambda}$, where n is the number of locations. Instead, one can calculate $\hat{H}$ from the SVD of Ω, which is $O(n^2T)$, which greatly speeds computation if the number of replicated observational fields, T, is much less than the number of locations. Even greater efficiency can be obtained by only calculating the square root matrix for the leading submatrix of $\hat{\Lambda}$, which is possible if one decides to threshold only the leading submatrix and zero out all of the remaining elements of $\hat{\Lambda}$, save the diagonals.

Nychka et al. (2001) have shown that such a smoothing approach can closely approximate a stationary Matérn covariance in the sense that the elements of the resulting smoothed matrix are similar to those of the original Matérn covariance matrix. The method can give a nonstationary covariance because the original empirical covariance is nonstationary, and nothing in the thresholding enforces stationarity. The intuition behind the method is that real structure is modelled through the retained coefficients, while noise is zeroed out during the thresholding. Unfortunately, there is no principled way to choose the degree of or exact procedure for the thresholding. In their empirical example, Nychka et al. (2001) set the smallest (in magnitude) 90% of the off-diagonal elements in the leading block (which corresponds to the coefficients of the coarsest father wavelets, which look like smooth bumps on a grid) of $\hat{H}$ to zero, retain all the diagonal elements of $\hat{H}$, and zero out the remaining elements.

Since it is not clear how one should do the thresholding, I create two smoothed correlation matrices based on the wavelet decomposition. The first aims to mimic the empirical residual correlation closely, while being positive definite. To create this correlation matrix, I use the empirical residual covariance, $\hat{C}_{\boldsymbol{Y}}$, and implement the thresholding in the following way. In the leading block of $\hat{H}$ I retain all the elements, while in the remainder of $\hat{H}$ I retain the diagonals and the largest (in magnitude) 50% of the off-diagonals. Because I want to model the variances as spatial processes with Gaussian process priors in the same way for all of the correlation models, I then create a smoothed correlation matrix, $\tilde{R}_{\boldsymbol{Y}}$ from $\tilde{C}_{\boldsymbol{Y}}$. In the MCMC, I treat $R_{\boldsymbol{Y}} = \tilde{R}_{\boldsymbol{Y}}$ as fixed, and model $\alpha(\cdot)$, $\beta(\cdot)$, $\eta(\cdot)$, and δ as before. Comparing the matrix $\tilde{R}_{\boldsymbol{Y}}$ to the empirical residual correlation matrix, the differences between elements of the matrices are at most 0.02. The second matrix aims to smooth the empirical correlation matrix, while retaining obvious structure. To do this I follow the procedure above, but include only the largest 0.5% of the off-diagonals that are not in the leading block.

The wavelet decomposition could be done within the context of a full Bayesian model, with priors on the elements of either Λ or H, but this would seem to require sampling matrix elements one by one. This would be very slow in a MCMC sampling scheme because, as each element changes, $C_{\boldsymbol{Y}}$ and its Cholesky factor would need to be recomputed. Furthermore, developing a prior framework for Λ is beyond the scope of this work.

5.4.1.4 Kernel-based Matérn nonstationary correlation

To fit a nonstationary covariance model within a fully Bayesian framework and account for uncertainty in the nonstationary covariance, I use the Matérn form of the kernel-based nonstationary covariance defined in Chapters 2 and 3. I use the basis kernel approach described in Section 3.2.4 to limit the computations. I parameterize the basis kernels, using the Givens angle approach given in Section 3.2.3.1. The pqth element of the kernel matrix at location i is given by

$$(\Sigma_i)_{pq} = \frac{\sum_{k=1}^{K} w_{ik} (\Sigma'_k)_{pq}}{\sum_{k=1}^{K} w_{ik}},$$

where Σ'_k is the kth basis kernel matrix, and w_{ik} are the weights defined below. Since sums of positive definite matrices are positive definite, the resulting kernel matrix, Σ_i, obtained by averaging the basis kernel matrices element by element, is positive definite. I weight the basis kernels based on a squared exponential weighting function

$$w_{ik} = \exp\left(-\left(\frac{\rho_{ik}}{\kappa_{\boldsymbol{Y}}}\right)^2\right), \tag{5.5}$$

where ρ_{ik} is angular distance between location i and the center of basis kernel k and $\kappa_{\boldsymbol{Y}}$ determines how quickly the weight decays with distance. The squared exponential weighting function (5.5) is infinitely differentiable as a function of ρ_{ik}, and we can express $\rho_{ik} = 2\sin^{-1}(\tau_{ik}) = \frac{1}{2}\cos^{-1}(-2\tau_{ik}^2 + 1)$, where τ_{ik} is distance in $\Re^3$. If we consider any element of Σ_i as a function in $\Re^3$ we can see that the function is infinitely differentiable as a function of location, which satisfies the conditions needed in Section 2.5.5 to show sample path differentiability of GPs using this nonstationary covariance model. This means that the differentiability of the residual functions, $f_t(\cdot)$ (5.4), which are integrated out of the model to give (5.1), are based on $\nu_{\boldsymbol{Y}}$ and ν_ϕ. I take $\nu_{\boldsymbol{Y}} \sim U(0.5, 15)$.

In Section 2.4, I discussed how to define kernel matrices on the sphere in a way that generates a nonstationary positive definite covariance on the sphere. To avoid having to do numerical integration, I use a shortcut here that does not seem to cause numerically non-positive definite matrices. To calculate the covariance between two locations, I locate the two points in the Lambert azimuthal equidistant Euclidean projection of the sphere that is centered at the midpoint of the great circle connecting the two locations. This projection accurately preserves the distances and directions from the centerpoint of the projection such that any circle in the Euclidean projection whose origin is the centerpoint of the projection gives a locus of points that are truly equidistant from the centerpoint. The effect of the projection is as if one took the globe with the centerpoint pointed straight up and squashed it directly down onto a plane. I calculate the closed form kernel covariance given in (2.5) based on the kernel matrices defined in this projection. Since I change the midpoint of the projection each time I change the locations under consideration, it is not strictly true that the kernels are solely a function of location, as is required for positive definiteness (2.2), but for the portion of the hemisphere used here, this approximation does not seem to cause problems.

A disadvantage of this nonstationary model is that even though it is more flexible than a stationary model, I still rely on the formulation of convolving kernels, which constrains the types of correlation structure that can be modelled. In particular, the kernel convolution model will not model correlations that do not drop off monotonically from the focal location well, and negative correlations are not possible in the model. This latter inability is a drawback for these data because the empirical data suggest that negative correlations are present, particularly between latitude bands (Figure 5.1). Presumably these negative correlations occur as the storm tracks shift in latitude over time.

5.4.2 Alternate approaches not used in this work

5.4.2.1 Smoothing the MLE fields

An approach that does not require modelling the full covariance of the data is to take the MLE fields for $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ and smooth these spatially. If one wanted to do this in a nonstationary way, it would essentially involve doing the regression modelling of Chapter 4, taking the MLEs as the observations. This approach is similar to that of Holland et al. (2000). The main drawback to this approach is that it does not provide uncertainty estimates of the trends that take account of the data correlation unless this is included, as done by Holland et al. (2000) in an ad hoc way. Because the ML estimates are smooth spatially, the estimated intercept and trend functions will be smooth, and the modelled noise about these functions will have very small variance. Represent model (5.1) as

$$Y_{it} = \alpha(\boldsymbol{x_i}) + \beta(\boldsymbol{x_i})t + r_t(\boldsymbol{x_i}) + \epsilon_{it},$$

where $r_t(\cdot)$ is a spatial residual process with $\text{Cov}(\boldsymbol{r_t}) = D(\boldsymbol{\eta})R_{\boldsymbol{Y}}D(\boldsymbol{\eta})$. Naive smoothing of the ML estimates, $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$, ignores the $r_t(\cdot)$ component of the model even though exploratory analysis suggests that $\alpha(\cdot)$ and $r_t(\cdot)$ account for most of the variability in the data. This is precisely what I want to avoid.

5.4.2.2 Smoothing the empirical covariance toward a stationary covariance

Like Nychka et al. (2001), Loader and Switzer (1992) focus on smoothing an empirical estimate of the covariance structure based on replicated data. They smooth toward a stationary model of the covariance structure, $C^S(\hat{\theta})$, giving

$$\tilde{C}_{\boldsymbol{Y}} = w\hat{C}_{\boldsymbol{Y}} + (1-w)C^S(\hat{\theta}),$$

where $w \in (0,1)$ is a smoothing parameter. The estimate, $\tilde{C}_{\boldsymbol{Y}}$, can be derived as the posterior mean of Bayesian model in which the prior for $C_{\boldsymbol{Y}}$ is inverse Wishart, and Loader and Switzer (1992) use this derivation to optimize w. It would be interesting to assess the success of this simple smoothing approach relative to those used in this work, particularly because the optimization of the degree of smoothing in this approach is straightforward, whereas choosing the degree of thresholding in the Nychka et al. (2001) model is difficult.

5.4.2.3 Fitting an unconstrained data covariance

Another possibility is to define a simple prior for the data correlation matrix, such as a uniform prior over the compact space of correlation matrices (Lockwood 2001), and then sample the matrices. It is possible to sample from unconstrained correlation matrices element by element in a way that ensures positive definiteness (Barnard, McCulloch, and Meng 2000; Lockwood 2001), but element by element sampling would be extremely computationally intensive, since the data covariance and its Cholesky factor would change each time an element changed. With 288 locations, this does not seem feasible; furthermore with only 51 years of data we do not have much data for fitting an unconstrained correlation matrix, and it is not clear how such a prior weights various types of correlation structures.

Using Bayesian methods, Daniels and Kass (1999) consider freely estimating small covariance matrices based on a variety of priors, while Smith and Kohn (2002) estimate sparse Cholesky decompositions of covariance matrices in a computationally efficient way. Some combination of the ideas in these papers may be feasible for the storm activity data, but would entail further methodological development, so I have not pursued these avenues further.

5.4.2.4 Covariances modelled indirectly through basis function regression

An approach that shares some features with the wavelet decomposition is recommended by Minka and Picard (1997). They suggest fitting each replicate of the residuals, $\boldsymbol{U}_t$, using a multilayer perceptron (feed-forward neural network), as $\boldsymbol{U}_t = B\boldsymbol{\omega}_t + \boldsymbol{\epsilon}_t$, and then reconstructing the implicitly modelled covariance as $\widetilde{C}_{\boldsymbol{Y}} = B\Omega\Omega^T B^T$ where Ω is a matrix whose columns are the vectors $\boldsymbol{\omega}_t$. The basis B could in principle be any basis; the critical choices that must be made involve the basis used, the number of basis functions, and the fitting method, since the degree to which $\widetilde{C_{\boldsymbol{Y}}}$ smooths the empirical covariance, $\hat{C}_{\boldsymbol{Y}}$, will be determined by these. The attraction in this approach is that one is doing regression modelling, which does not involve the constraints of positive definiteness, rather than covariance modelling. One can in principle model covariances involving locations not in the training set as well as incorporate uncertainty in the estimated covariance, but this requires one to model Cov$(\boldsymbol{\omega}_t)$ in some fashion, which leads back to the difficulties involved in covariance modelling.

5.5 Fitting the Model

5.5.1 Data and data-dependent model specifications

I fit the model for two of the storm indices in Paciorek et al. (2002). First, I fit the model to the logarithm of the temperature variance index in the Pacific region of the Northern Hemisphere, defined as $20° - 75°$ N, $130° - 245°$ E (Figure 5.3). Second, I fit the model to the Eady growth rate in the Atlantic region of the Northern Hemisphere, defined as $20° - 75°$ N, $250° - 5°$ E (Figure 5.3). Using a $5° \times 5°$ subgrid of the original $2.5° \times 2.5°$ NCEP-NCAR reanalysis grid, this gives 288 locations. While we would ideally like to fit the data from the whole hemisphere, $20° - 70°$ N, that was analyzed in Paciorek et al. (2002) and to the finer grid, the computational limitations of this hierarchical model with GP priors and nonstationary covariance limit the number of locations that can be fit. Even with only 288 locations, a full MCMC takes several weeks on a moderately-fast computer. The wavelet code of Nychka et al. (2001) requires a regular grid based on powers of two (with at most one factor of three allowed), so I include $75°$ N (which was not used in the analysis of Paciorek et al. (2002)) so that with one-third of a hemisphere, I have a grid with 24 longitude values and 12 latitude values. I use this same grid for the non-wavelet-based correlation models to facilitate comparison between models.

For the wavelet representation, at the coarsest level, I use a grid with 12 longitude and 6 latitude

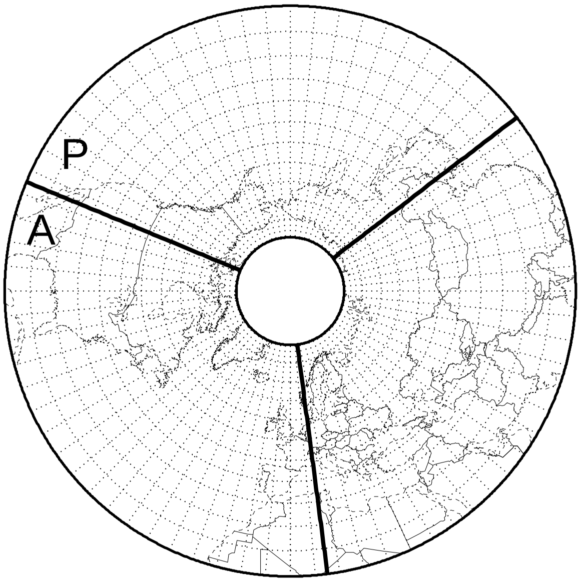

Figure 5.3. Map of the Northern hemisphere, $20° - 75°$ *N, with* $5° \times 5°$ *grid overlaid as dotted lines and Pacific (P) and Atlantic (A) region boundaries indicated by the thick dark lines of longitude.*

values, which means there are 72 ‘smooth’ basis functions (using the terminology of Nychka et al. (2001)). These basis functions are essentially bumps centered on the subgrid; they are able to pick up patterns in the covariance structure of resolution approximately $10°$. The leading submatrix of $\hat{H}$, to which I refer in Section 5.4.1.3, is therefore a 72 by 72 matrix.

While the latitude-longitude grid of the data distorts distances and areas, I rely on the ability of the wavelet-based decomposition to model nonstationarity to account for the changes in distances with latitude. Even if the data truly were stationary, this would require that the decomposition give a longer correlation scale at high latitude than at low latitude, to account for the distances between grid points being shorter at high latitude.

In constructing the kernel convolution nonstationary covariance model, I use nine basis kernels, which I believe are sufficient to represent the basic nonstationarity in the data. I position the nine kernels on a three by three latitude-longitude grid, using three kernels at each of the latitudes $30°$ N, $45°$ N, and $60°$ N, with the kernels 40 degrees apart in longitude. For the Atlantic region, the longitudes are $270°$ E, $310°$ E and $350°$ E, and for the Pacific, $150°$ E, $190°$ E, and $230°$ E.

5.5.2 MCMC sampling procedure

5.5.2.1 Posterior mean centering

I fit the model via MCMC. For the spatial parameters, $\boldsymbol{\phi} \in \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \log(\boldsymbol{\eta}^2)\}$, I use the posterior mean centering (PMC) scheme outlined in Section 3.6.2.2. Because the years are equally-spaced, I can center time about the mean of the years and sample $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ independently. This simplifies the sampling and justifies using independent priors for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. In the PMC sampling for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ I can use the exact conditional posterior mean, since the prior and likelihood are both of Gaussian form for these parameters. Note that I could integrate parameters out of the model, which might speed the calculations and improve mixing. For $\boldsymbol{\phi} = \log(\boldsymbol{\eta}^2)$, I use an approximation to the posterior mean, based on (5.2) using the usual MLE $\hat{\boldsymbol{\phi}}$:

$$\widehat{\phi(\boldsymbol{x_i})} = \log \widehat{\eta(\boldsymbol{x_i})}^2 = \log \left(\frac{\sum_t (Y_{it} - \widehat{\alpha(\boldsymbol{x_i})} - \widehat{\beta(\boldsymbol{x_i})} t)^2}{T} \right),$$

where the hats indicate the usual MLEs calculated from the data independently by location. To approximate $C_{\hat{\boldsymbol{\phi}}}$ in (5.2), I first calculate the covariance of $\hat{\boldsymbol{\eta}}^2$ empirically and then use the delta method to calculate the approximate covariance of $\log(\hat{\boldsymbol{\eta}}^2)$. I derive the empirical covariance of $\hat{\boldsymbol{\eta}}^2$ as follows. Let the residual at location i and time t be denoted $U_{it} = Y_{it} - \alpha(\boldsymbol{x_i}) - \beta(\boldsymbol{x_i})t$ and let $C = C_{\boldsymbol{Y}}$ be the covariance matrix of the residuals with ijth element, C_{ij}. I need the following moments

$$\begin{aligned} \mathrm{E}U_{it}^2 &= C_{ii} \\ \mathrm{E}\left(U_{it}^2 U_{js}^2\right) &= C_{ii}C_{jj} + 2C_{ij}^2, \end{aligned}$$

where the second expression is derived straightforwardly, but tediously, based on (U_{it}, U_{jt}) being bivariate normal with covariance, C_{ij}. Now, consider

$$\begin{aligned} \mathrm{Cov}(\hat{\boldsymbol{\eta}}^2)_{ij} &= \mathrm{E}\left(\frac{\sum_{t=1}^T \sum_{s=1}^T U_{it}^2 U_{js}^2}{T^2}\right) - \mathrm{E}\left(\frac{\sum_{t=1}^T U_{it}^2}{T}\right) \mathrm{E}\left(\frac{\sum_{t=1}^T U_{jt}^2}{T}\right) \\ &= \frac{1}{T^2} \sum_t \sum_s \mathrm{E}\left(U_{it}^2 U_{js}^2\right) - C_{ii}C_{jj} \\ &= \frac{T(T-1)}{T^2} C_{ii}C_{jj} + \frac{T}{T^2}(C_{ii}C_{jj} + 2C_{ij}^2) - C_{ii}C_{jj} \\ &= \frac{2C_{ij}^2}{T}. \end{aligned}$$

Next, I use a second-order Taylor expansion applied to $\log \widehat{\eta(\boldsymbol{x_i})}^2$. Suppressing the dependence on $\boldsymbol{x_i}$, this gives me

$$\hat{\phi} = \log \hat{\eta}^2 \approx \log \eta^2 + (\eta^2 - \hat{\eta}^2)\frac{1}{\eta^2}.$$

Calculating the covariance of the right-hand side and plugging in $\widehat{\eta(\boldsymbol{x_i})}$ for $\eta(\boldsymbol{x_i})$, I can now approximate $C_{\hat{\boldsymbol{\phi}}}$ as

$$\text{Cov}(\hat{\boldsymbol{\phi}}) \approx D(\hat{\boldsymbol{\eta}}^{-2})\text{Cov}(\hat{\boldsymbol{\eta}}^2)D(\hat{\boldsymbol{\eta}}^{-2}),$$

where $D(\hat{\boldsymbol{\eta}}^{-2})$ is a diagonal matrix with the reciprocals of the MLE variances on the diagonal. Using $\hat{\boldsymbol{\eta}}$ as a plug-in estimator (which simplifies the acceptance ratio calculations) again in the expression $C_{\boldsymbol{Y}} = D(\boldsymbol{\eta})R_{\boldsymbol{Y}}D(\boldsymbol{\eta}) + \delta I$, the final result is

$$C_{\hat{\boldsymbol{\phi}}} \approx \frac{2}{T}D(\hat{\boldsymbol{\eta}}^{-2})(D(\hat{\boldsymbol{\eta}})R_{\boldsymbol{Y}}D(\hat{\boldsymbol{\eta}}) + \delta I)^{*2}D(\hat{\boldsymbol{\eta}}^{-2}), \tag{5.6}$$

where the $*2$ notation indicates squaring element by element. Note that for $\delta \approx 0$, which is the case for the spatial model, we have

$$C_{\hat{\boldsymbol{\phi}}} \approx \frac{2}{T}R_{\boldsymbol{Y}}^{*2},$$

which does not involve $\hat{\boldsymbol{\eta}}^2$, which makes sense because the logarithm is variance-stabilizing.

5.5.2.2 Sampling steps

As in Section 4.3, I define the basic posterior mean centering proposal S1. Proposal S1 in the context of the spatial model is as follows.

1. This proposal applies to either μ or to the pair (σ, κ). Propose the hyperparameter(s), using a Metropolis or Metropolis-Hastings proposal. In the notation that follows, I will indicate that all three of the hyperparameters $\mu, \sigma,$ and κ have been proposed, but this is merely for notational convenience.

2. Propose $\boldsymbol{\phi} \in \boldsymbol{\alpha}, \boldsymbol{\beta}, \log \boldsymbol{\eta}^2$ conditionally on $\boldsymbol{\theta}^* = \{\mu^*, \sigma^*, \kappa^*\}$ as

$$\boldsymbol{\phi}^* \sim \text{N}(\widetilde{\boldsymbol{\phi}(\boldsymbol{\theta}^*)} + \sigma^* L(\kappa^*)\boldsymbol{\chi}, v^2 R(\kappa^*)),$$

where $\boldsymbol{\chi} = (\sigma L(\kappa))^{-1}(\boldsymbol{\phi} - \widetilde{\phi(\boldsymbol{\theta})})$. $\widetilde{\boldsymbol{\phi}(\boldsymbol{\theta})}$is the posterior mean of $\boldsymbol{\phi}$ conditional on the current hyperparameters and $\widetilde{\boldsymbol{\phi}(\boldsymbol{\theta}^{*}})$is the posterior mean of $\boldsymbol{\phi}$ conditional on the proposed hyperparameter(s),

$$\begin{aligned}\widetilde{\boldsymbol{\phi}(\boldsymbol{\theta})} &= \sigma^2 R_{\boldsymbol{\phi}}(\kappa)(C_{\hat{\boldsymbol{\phi}}} + \sigma^2 R_{\boldsymbol{\phi}}(\kappa))^{-1}\hat{\boldsymbol{\phi}} + C_{\hat{\boldsymbol{\phi}}}(C_{\hat{\boldsymbol{\phi}}} + \sigma^2 R_{\boldsymbol{\phi}}(\kappa))^{-1}\mu \\ &= \mu + \sigma^2 R_{\boldsymbol{\phi}}(\kappa)(C_{\hat{\boldsymbol{\phi}}} + \sigma^2 R_{\boldsymbol{\phi}}(\kappa))^{-1}(\hat{\boldsymbol{\phi}} - \mu).\end{aligned}$$

$\hat{\boldsymbol{\phi}}$ is the MLE for $\boldsymbol{\phi}$, and $C_{\hat{\boldsymbol{\phi}}}$ is the covariance of the MLE, either the approximation (5.6) for $\log \boldsymbol{\eta}^2$ or the exact covariance in Section 5.3 for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Note again that, as mentioned in Chapter 3, it is allowable to set $v = 0$, so long as one includes the Jacobian of the deterministic mapping $\boldsymbol{\phi} \rightarrow \boldsymbol{\phi}^{*}$, which is the same as the Hastings ratio one uses if $v > 0$.

3. The Hastings ratio (the proposal ratio portion of the acceptance ratio) for the proposal is a ratio of determinants, which cancels the determinant ratio from the prior for $\boldsymbol{\phi}$ as described in Section 3.6.2.2.

Next I describe the sampling steps in a single iteration of the Markov chain for the nonstationary correlation model using the basis kernels. The steps for the stationary correlation and for the fixed wavelet-based correlation models are straightforward simplifications of this scheme.

1. Sample δ using a simple Metropolis step.

2. Sample $\log \kappa_{\boldsymbol{Y}}$, the weighting parameter for the basis kernel averaging, using a simple Metropolis step.

3. For $\boldsymbol{\theta} = (\log \lambda_{1,1}, \ldots, \log \lambda_{1,K})$, the first eigenvalues of the $K = 9$ basis kernels, use a simple multivariate Metropolis step with the same proposal standard deviation for each element. The basis kernels are sampled jointly in this fashion to avoid having to recalculate the Cholesky of the data covariance matrix too frequently. In practice the basis kernel parameters seem to mix much more quickly than the hyperparameters of $\boldsymbol{\phi}$, so this does not seem to be a problem.

4. Repeat step 3 for $\boldsymbol{\theta} = (\log \lambda_{2,1}, \ldots, \log \lambda_{2,K})$, the second eigenvalues of the basis kernels.

5. Repeat step 3 for $\boldsymbol{\theta} = (\rho_1, \ldots, \rho_K)$, the Givens angles of the basis kernels. If I propose ρ_k outside of $(0, \pi)$ I add or subtract π as necessary to bring the proposal back into the parameter space.

6. For $\boldsymbol{\phi} = \boldsymbol{\alpha}$, do the following proposals:

   (a) Sample $(\mu_\phi, \boldsymbol{\phi})$ jointly using a proposal of type S1.

   (b) To account for the high posterior correlation between κ_ϕ and σ_ϕ, sample $(\kappa_\phi, \sigma_\phi, \boldsymbol{\phi})$ jointly using a modified proposal of type S1. The modification of substep (1) of S1 is as follows. First, sample $\log \kappa_\phi^* \sim \mathrm{N}(\log \kappa_\phi, v_1^2)$ using a simple Metropolis step. Next, sample

   $$\log \sigma_\phi^* \mid \kappa_\phi^* \sim \mathrm{N}\left(\log \sigma_\phi \frac{v_1}{r}(\log \kappa_\phi^* - \log \kappa_\phi), v_2^2\right).$$

   I take v_2 to be very small, so that κ_ϕ and σ_ϕ move together closely with the ratio r being a constant scale factor chosen to speed mixing. Finally use substep 2 of S1 to sample $\boldsymbol{\phi}^*$ conditional on $(\kappa_\phi^*, \sigma_\phi^*)$.

   (c) Next, allow $\log \sigma_\phi$ to move independently of $\log \kappa_\phi$. Sample $\log \sigma_\phi$ using a simple Metropolis step based on the centered parameterization without changing $\boldsymbol{\phi}$. In other words, the acceptance ratio will only involve the ratio of the priors for $\log \sigma_\phi^*$ and $\log \sigma_\phi$ and the ratio of the prior for $\boldsymbol{\phi}$ as a function of $\log \sigma_\phi^*$ to that of the prior for $\boldsymbol{\phi}$ as a function of $\log \sigma_\phi$. Note that this does not involve having to invert $R_{\boldsymbol{\phi}}(\kappa_\phi)$. I sample $\log \sigma$ in this way rather than with a PMC step of type S1 because joint sampling mixes very slowly, with the chain spending a long time in parts of the space with large values for the elements of the implicit parameter, $\boldsymbol{\omega} = (\sigma_\phi L(\kappa_\phi))^{-1}(\boldsymbol{\phi} - \mu_\phi)$, and correspondingly low values for the log of the prior density, which is a function of $\boldsymbol{\omega}^T\boldsymbol{\omega}$. The reason for this is still unclear to me, but the phenomenon seems to occur because it is difficult for σ_ϕ and $\boldsymbol{\phi}$ to 'trade-off' when employing joint proposals for $(\sigma_\phi, \boldsymbol{\phi})$. In the joint proposals, when large values of $\sigma_{\boldsymbol{\phi}}$ are proposed using PMC, the result is that $\boldsymbol{\phi}$ is proposed conditionally on σ_ϕ^* such that it has higher variability, and it is difficult to increase the prior density of $\boldsymbol{\phi}$ by increasing σ_ϕ. Using a straightforward centered parameterization style proposal for σ_ϕ allows σ_ϕ and $\boldsymbol{\omega}$ to stabilize with $\boldsymbol{\omega}$ approximately $N(0, I)$ and $\boldsymbol{\omega}^T\boldsymbol{\omega} \approx n = 288$.

(d) Propose $\boldsymbol{\phi}$ using a simple Metropolis step with correlation amongst the elements of $\boldsymbol{\phi}$: $\boldsymbol{\phi}^* \sim \mathrm{N}(\boldsymbol{\phi}, v^2 R(\kappa_{\boldsymbol{\phi}}))$. It is also straightforward to do a Langevin update here, however, I did not use such an update in the model runs reported in this thesis.

7. Repeat step 7 for $\boldsymbol{\phi} = \boldsymbol{\beta}$. Again a Langevin update in substep (d) would be straightforward.

8. Repeat step 7 for $\boldsymbol{\phi} = \log \boldsymbol{\eta}^2$. Note that a Langevin update here in substep (d) is not straightforward because there is no simple form for the gradient of the log posterior with respect to $\log \boldsymbol{\eta}^2$.

5.5.2.3 Running the chain

The priors and initial values are listed in detail in the Appendix. Priors are taken to be relatively noninformative. For parameters involved in the correlation structures I use reasonable lower and upper limits based on the induced correlations between the nearest and most distant pairs of locations analyzed. For the initial values, I would like to use the MLEs for the spatial fields and reasonable values for the hyperparameters based on ad hoc analysis of the MLEs. Unfortunately while I am able to do the latter, it is difficult to know what initial prior correlation structure to use so that the MLEs are consistent with the initial prior correlation, i.e., so that we have

$$\hat{\boldsymbol{\phi}} = \mu_\phi + \sigma_\phi L(\kappa_\phi)\boldsymbol{\omega}$$

with the magnitude of $\boldsymbol{\omega}$ remaining reasonable. Instead, I chose to initialize $\boldsymbol{\phi}$ by taking $\boldsymbol{\omega} \sim \mathrm{N}(0, I)$, namely simulating $\boldsymbol{\omega}$ from its prior.

To fit the model, I run a long MCMC chain. Because of the slow mixing and long computation time, I can only run a limited number of runs and cannot do a full convergence assessment using multiple chains. Instead I assess time series and acf plots and compare the beginning of the chain to the end to get some feel for the behavior of the chain. I cannot be sure that the chain has fully converged, but based on many trial runs during which I was adjusting the parameterizations and sampling schemes, I believe that the general results from the runs, in particular the comparison between models, are legitimate, even if some of the parameters cannot be shown to have completely mixed. However, the slow mixing and potential lack of convergence is a cause for real concern with this model and fitting procedure, although perhaps not any more so than for any large Bayesian model of temporal or spatial structure. Furthermore, since the models are compared via cross-validation, the model comparison results still give a valid indication of which of the correlation structures best model these data.

I used at least 50000 iterations for burn-in (longer in some cases) and then sampled 500,000 iterations for inference and model comparison. To economize on storage space, I retained only every 10th iteration, giving a final sample of size 50,000. In adjusting the proposal variances during burn-in, I attempted to achieve the acceptance rates recommended in Roberts and Rosenthal (2001), namely, 0.44 for scalar parameters and 0.23 for vector parameters, and generally come quite close to these rates. The parameters μ_ϕ and δ mix much more quickly than the other parameters, so to reduce computation, I chose to sample those parameters only once for every five updates of the remaining parameters.

5.6 Model Comparison Criteria

As I discussed in Section 5.2, the main scientific goal of the spatial model is to better estimate the slopes and their uncertainty. Therefore I focus on prediction in time rather than prediction in space. To do this I split the data into $T = 44$ training years and $T^* = 7$ test years, and hold out all the data for the 288 locations in the 7 years, which are distributed throughout the 51 year period (1949,1954,1964,1974,1989,1994, and 1999). Based on these 7 years of data, I use two main criteria to compare the success of the models in prediction. The first is the posterior predictive distribution of the test data,

$$\begin{aligned}
\log \prod_{t^*} h(\boldsymbol{Y}_{t^*}|\boldsymbol{Y}_t) &= \log \int \prod_{t^*} h(\boldsymbol{Y}_{t^*}|\boldsymbol{\alpha}, \boldsymbol{\beta}, C_{\boldsymbol{Y}})\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}, C_{\boldsymbol{Y}}|\boldsymbol{Y}_t)d\boldsymbol{\theta} \\
&\approx \log \sum_{k=1}^{K} |C_{\boldsymbol{Y}}^{(k)}|^{-\frac{T^*}{2}} \\
&\quad \times \exp\left(-\frac{1}{2}\sum_{t^*}(\boldsymbol{Y}_{t^*} - \boldsymbol{\alpha}^{(k)} - \boldsymbol{\beta}^{(k)}t^*)^T C_{\boldsymbol{Y}}^{(k)-1}(\boldsymbol{Y}_{t^*} - \boldsymbol{\alpha}^{(k)} - \boldsymbol{\beta}^{(k)}t^*)\right),
\end{aligned}$$

where the integral is approximated by the average over the posterior simulations. Models that perform well should have high predictive density on test data. This criterion is heavily influenced by how well the models predict the covariance structure of the data. As an alternative, I use mean squared error as my second criterion,

$$\text{MSE} = \frac{\sum_{t^*}(\boldsymbol{Y}_{t^*} - \boldsymbol{\alpha} - \boldsymbol{\beta}t)^T(\boldsymbol{Y}_{t^*} - \boldsymbol{\alpha} - \boldsymbol{\beta}t)}{nT^*}.$$

This criterion focuses on the point predictions and ignores the spatial covariance structure of the test data.

5.7 Results

5.7.1 Convergence

All of the models show evidence of slow mixing. In Figure 5.4, I show time series plots of the log posterior densities for the four models for temperature variance, as suggested in Gelman and Rubin (1992) as a way to monitor convergence to the full joint distribution. The wavelet-empirical model has not burned-in, with the log-likelihood still increasing, reaching 53065, which is far greater than for the model with the next highest log-likelihood, the kernel nonstationary model which spends 95% of iterations in the range (28186,28289). Clearly the wavelet-empirical model is closely fitting the covariance structure of the training data, although as we will see in Section 5.7.3, it does a terrible job in predicting held-out data. While the three remaining models appear to be sampling from their posterior distributions, the posterior density plots (Figure 5.4) suggest that mixing is slow. The log-likelihood for the stationary model is approximately (27481,27583) while the wavelet-smooth model is much lower at (23565,23668). The time series plots of the log posterior density are qualitatively similar for the Eady growth rate dataset (not shown). In Figure 5.5 I show time series plots for the hyperparameters $\mu, \sigma,$ and κ of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\log \boldsymbol{\eta}^2$ for the kernel nonstationary model for temperature variance. These indicate that while the μ parameters have mixed, σ and κ have long-term trends (less pronounced for $\boldsymbol{\beta}$) and have not adequately mixed. The mixing properties of the process hyperparameters in the stationary model and the wavelet-smooth model appear to be similar (not shown). For the kernel nonstationary model for Eady growth rate, the hyperparameter mixing is somewhat better, but still not satisfactory, while mixing of σ_η and κ_η are much worse for the stationary model for Eady growth rate (not shown). The lack of hyperparameter mixing appears to impact the process values to some degree, as seen for three locations in the kernel nonstationary model for temperature variance in Figure 5.6. Mixing in the stationary model is similar while mixing in the wavelet-smooth model is even worse (not shown). Interestingly, the values of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}^2$ appear to mix better than those of $\boldsymbol{\alpha}$; in particular, the values of $\boldsymbol{\beta}$, in which we are most interested, do seem to be mixing reasonably well. Mixing of the process values for the Eady growth rate models is qualitatively similar (not shown). Of course these plots only assess the marginal distributions of the process values and not the joint distributions, so the problems with hyperparameter and log posterior density mixing remain a concern. In general, it does appear that the parameters are staying within a range of values, so the sample from the chain may give us some idea of reasonable parameter values even though I do not have confidence that the sample truly reflects the posterior. Note that the PMC scheme greatly improves the mixing of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and their hyperparameters. It also helps in the mixing of the $\boldsymbol{\eta}$ and its hyperparameters. In particular, μ_η mixes well now, and σ_η and κ_η much better than without PMC, albeit still glacially slowly. The impact of the sampling scheme on the mixing of $\boldsymbol{\eta}$ is of most interest, since in a real application, I would integrate $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ out of the model.

I have not carried out a full MCMC convergence diagnosis based on running multiple chains from dispersed starting values and ensuring that the chains all converged to the same posterior. However, for both the stationary and kernel nonstationary models for both datasets, I have run three chains each from dispersed starting values. Although I did not run the chains long enough for the chains to completely converge to the distribution found in the primary runs, the parameters did appear to be converging to the same values. Note that to achieve this, I needed to include a new sampling step in the MCMC, in which I used a step of type S1 (Section 5.5.2.2) for $(\kappa_\phi, \boldsymbol{\phi})$. I believe this is necessary to allow the chain to move quickly from the dispersed starting values in which the pair $(\sigma_\phi, \kappa_\phi)$ are not in the right ratio relative to each other. Once the chain burns-in, this step does not seem to be necessary, and the correlated proposal for σ_ϕ and κ_ϕ seem to be sufficient. For some of the chains, a few of the hyperparameters and some of the process values were somewhat different than the values seen in the primary runs, but I suspect this is an effect of

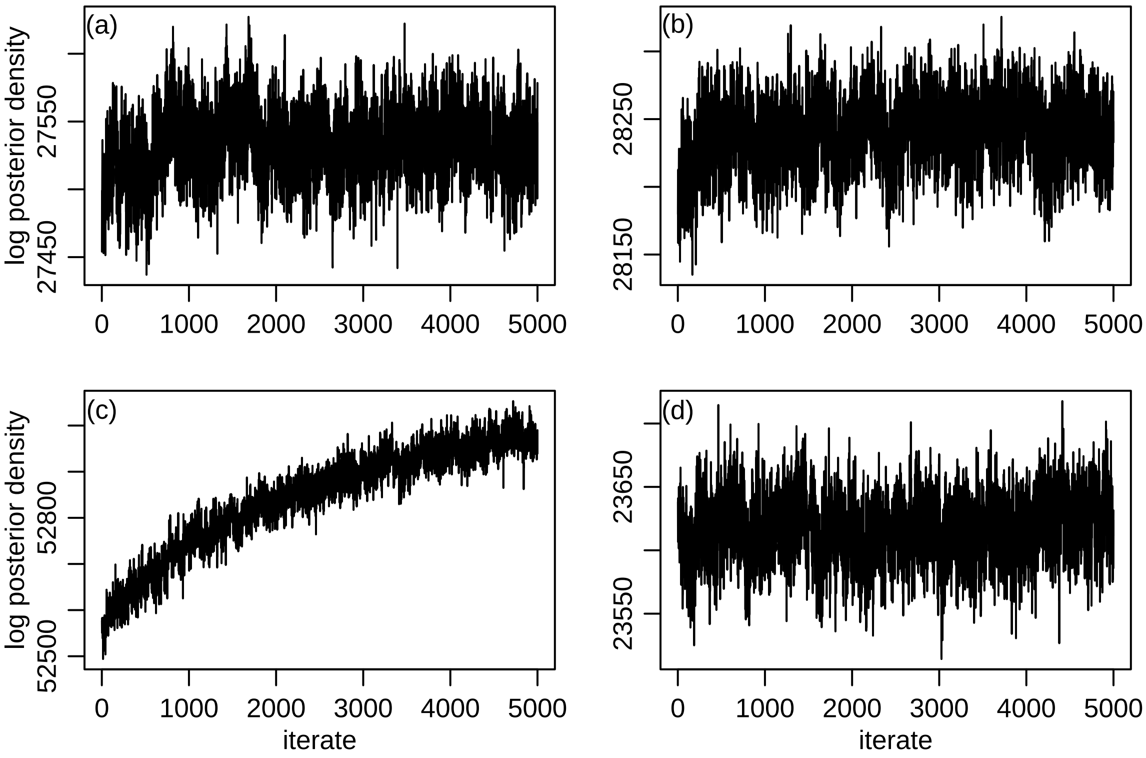

Figure 5.4. Time series plots of the log posterior density for temperature variance for the four Bayesian models: (a) stationary, (b) kernel nonstationary, (c) wavelet-empirical, and (d) wavelet-smooth.

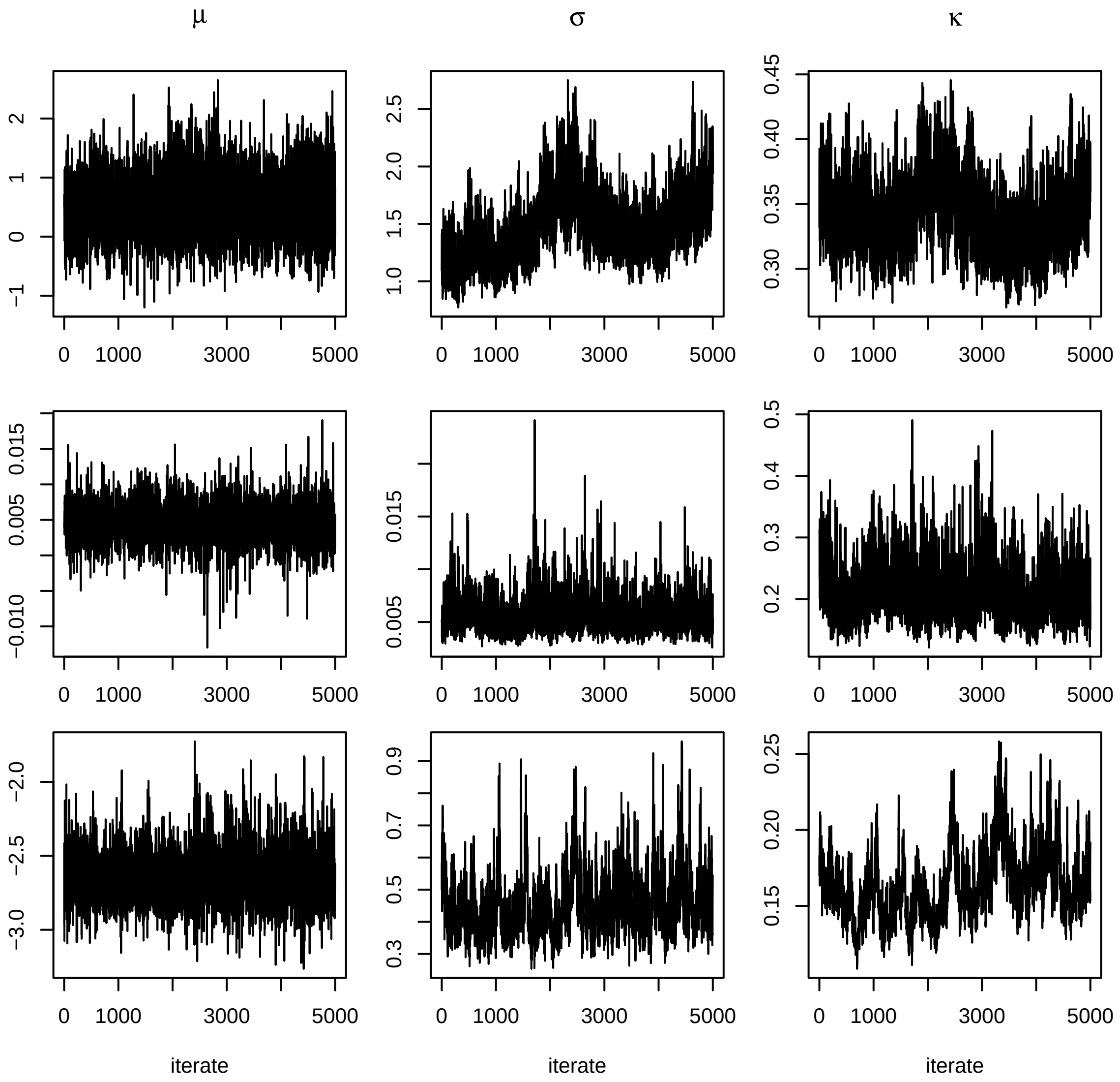

Figure 5.5. Time series plots of the hyperparameters, μ (first column), σ (second column), and κ (third column) for $\boldsymbol{\alpha}$ (first row), $\boldsymbol{\beta}$ (second row), and $\boldsymbol{\eta}^2$ (third row) from the kernel nonstationary model fit to the temperature variance data by MCMC.

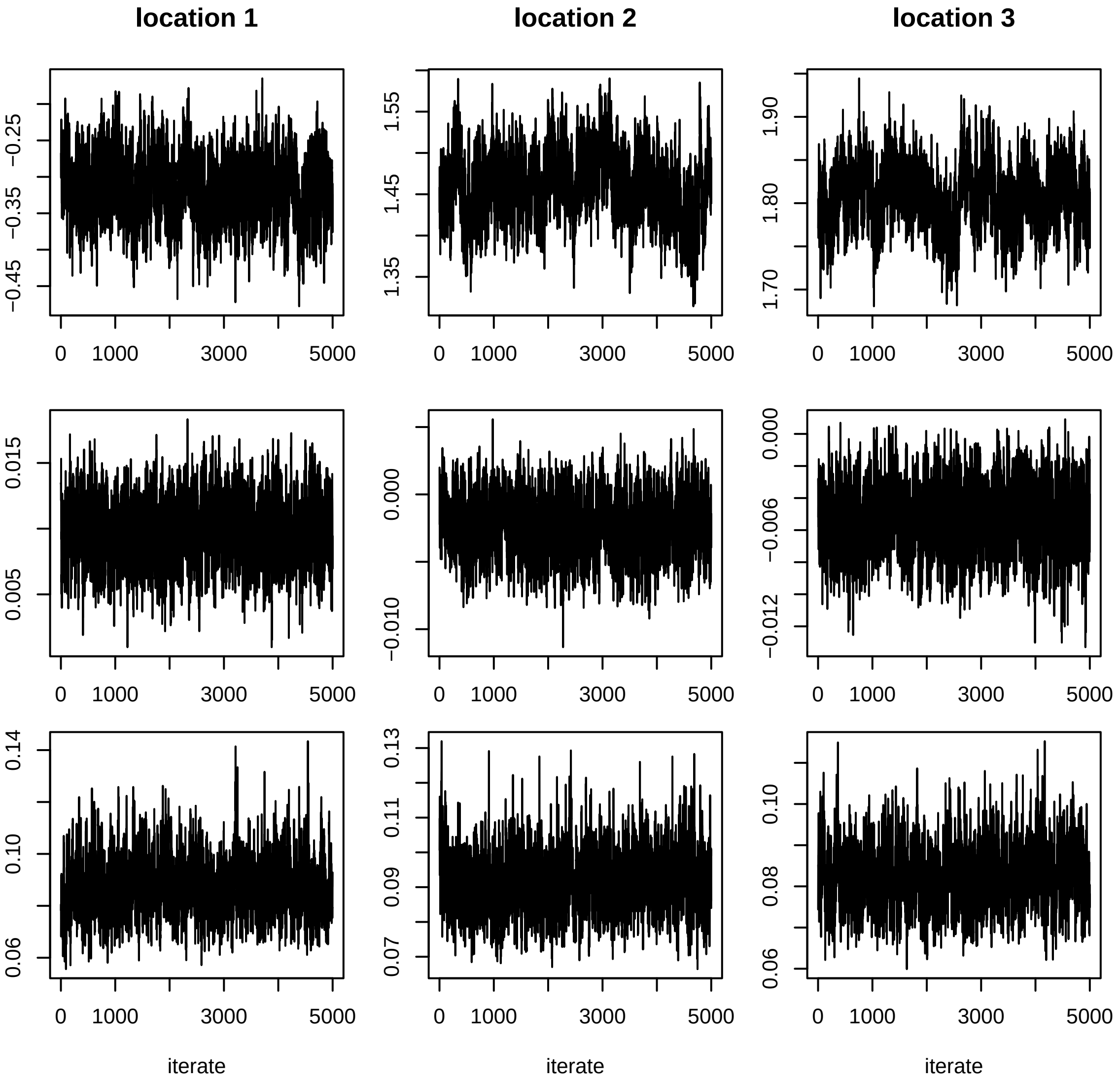

Figure 5.6. Time series plots of three process values of $\boldsymbol{\alpha}$ (first row), $\boldsymbol{\beta}$ (second row) and $\boldsymbol{\eta}^2$ (third row) for the kernel nonstationary model fit to the temperature variance by MCMC.

not having burned-in. In general, this occurs for the σ and κ hyperparameters, and coincides with the chain having log-likelihood values lower than achieved in the primary runs. Also, I have run many runs with these data during the development of the model and therefore have some sense for the range of plausible values of the parameters. While some parameters in the primary runs have not fully mixed, the predictive quantities all seem very stable, suggesting that the results are robust with respect to the comparison of models. In partial defense of the model, mixing is likely to be an issue for most other Bayesian models of similar complexity, and classical fitting methods for large models are prone to finding local minima and not fully exploring the parameter space.

5.7.2 Model estimates

To evaluate the effect of the correlation model used, I first compare the posterior mean estimates of the slopes and residual variances from the four models for temperature variance to the MLEs for the slopes and variances, all based on the training data. In mapping the slopes (5.7), we see that the stationary and kernel nonstationary models smooth the MLE field but retain much of its structure, while the wavelet methods drastically smooth out the peaks and troughs in the MLE field. In the residual variance maps (Figure 5.8), the stationary estimates bear no particular resemblance to the MLEs, while the kernel nonstationary estimates appear to be a smoothed version of the MLEs. The wavelet-smooth estimates are an order of magnitude larger than the MLEs, but the spatial pattern largely mimics that of the MLEs. The wavelet-empirical estimates closely match the pattern of the MLEs but are smaller in value. Note that in mapping these and subsequent quantities, I have used the contour function in the R statistical package, which performs some additional interpolation on top of the interpolation done by the Bayesian model.

In Figure 5.9 I focus further on the smoothing being done by the models of temperature variance using scatterplots of the posterior mean values for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\eta}^2$ for the four full models against the MLE values. Consider first the stationary and kernel nonstationary models. For the intercepts, the posterior means coincide closely with the MLEs. The two models appear to smooth the MLE slopes in similar fashion. For the residual variance, the nonstationary model seems to be doing some smoothing of the MLEs, while the estimates from the stationary model are not clearly related to the MLEs and some of the residual variances are rather high. These high variance estimates

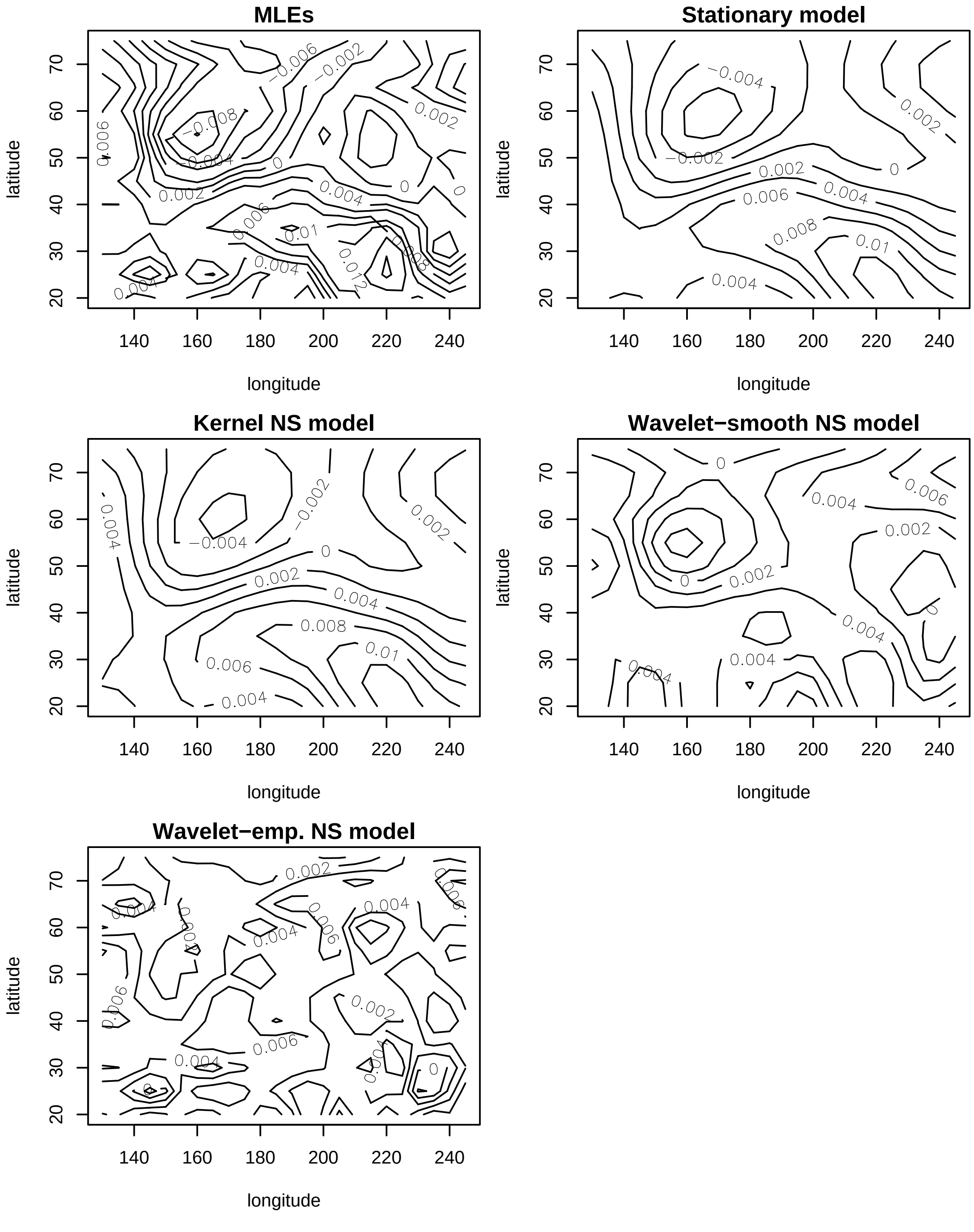

Figure 5.7. Maps of estimated **β** *values for temperature variance for (a) MLE model, and posterior means from (b) stationary model, (c) kernel nonstationary model, (d) wavelet-smoothed covariance model, and (e) wavelet-empirical covariance model.*

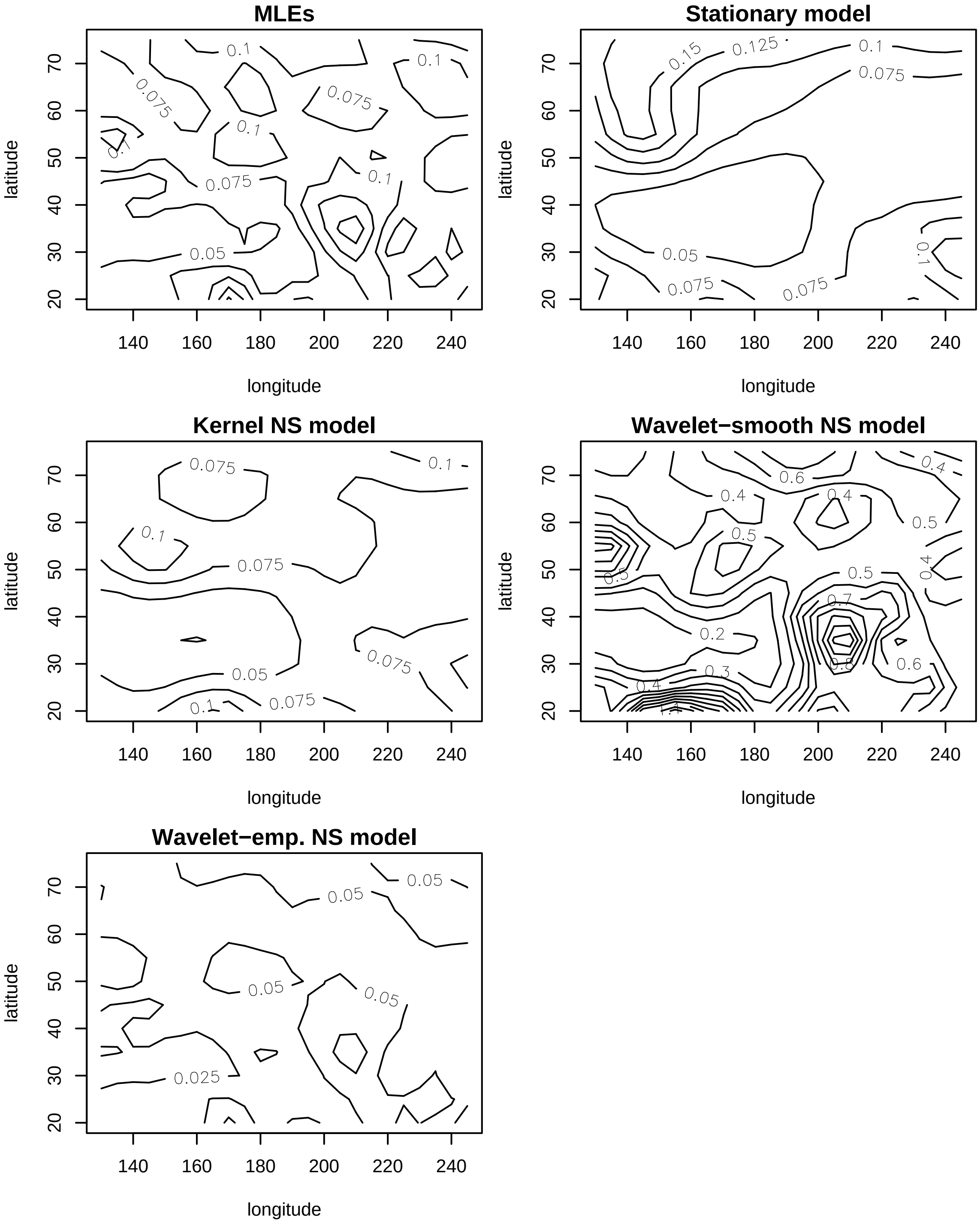

Figure 5.8. Maps of estimated η^2 values for temperature variance for (a) MLE model, and posterior means from (b) stationary model, (c) kernel-based nonstationary model, (d) wavelet-smoothed covariance model, and (e) wavelet-empirical covariance model.

are occurring at locations where the correlation structure is not well modelled by the stationary model; the variance increases at these locations to compensate. We can see this by calculating the standardized residuals, $\boldsymbol{Y}_t^* = L_{\boldsymbol{Y}}(\hat{\boldsymbol{\eta}}^2)^{-1}(\boldsymbol{Y}_t - \tilde{\boldsymbol{\alpha}} - \tilde{\boldsymbol{\beta}}t)$ where $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{\beta}}$ are the posterior mean estimates and where the Cholesky factor $L_{\boldsymbol{Y}}(\hat{\boldsymbol{\eta}}^2)$ of the residual covariance matrix is calculated based on the posterior means for $\kappa_{\boldsymbol{Y}}, \nu$, and δ, but using the MLEs, $\hat{\boldsymbol{\eta}}^2$ rather than the posterior mean estimates. If the covariance structure fits the data, then $\boldsymbol{Y}_t$ should be uncorrelated white noise with standard deviation of one. I calculate the average squared values of these residuals, averaging over the training years, and plot this by location (Figure 5.10). The areas of lack of fit, where there are large, correlated squared residual values, coincide with the areas in which the ratios of the posterior mean variances to the MLE variances,

$$\frac{\widetilde{\eta_i}^2}{\widehat{\eta_i}^2},$$

are larger than one, generally falling in the corners of the plot, particularly the upper left corner.

The wavelet models do not match the MLE intercepts as closely as the stationary and kernel nonstationary models and smooth the slopes to a much higher degree (Figure 5.9). For the wavelet model with the smoothed correlation matrix, the residual variances are very large relative to the MLEs, suggesting that the correlation model is fitting the correlation structure of the data very poorly, and the variances must increase to compensate. I am not surprised that these variance estimates are larger than the MLEs since the correlation matrix is fit without reference to the likelihood, so the variance estimates and/or the value of δ must compensate for any lack of fit of the correlation matrix to the data. However, I am surprised that the variance estimates are so high, given that for the stationary model, the posterior mean residual variances are not more than about twice as large as the MLEs. The estimate of δ for the wavelet smooth model is three orders of magnitude larger than for the other models, again suggesting compensation for lack of fit of some sort in the model. For the wavelet model that mimics the empirical correlation matrix, the variances are a linear function of the MLEs and are lower than the MLEs. It appears that this correlation model fits the training data so well (as shown by the large log-likelihood) that the residual variance estimates are lower than the MLEs even though the intercept estimates appear quite different than their MLEs, which one would expect to detract from the model fit and drive the variances up. In any event, it is clear that the wavelet models are fitting the data in rather strange ways. Given that there

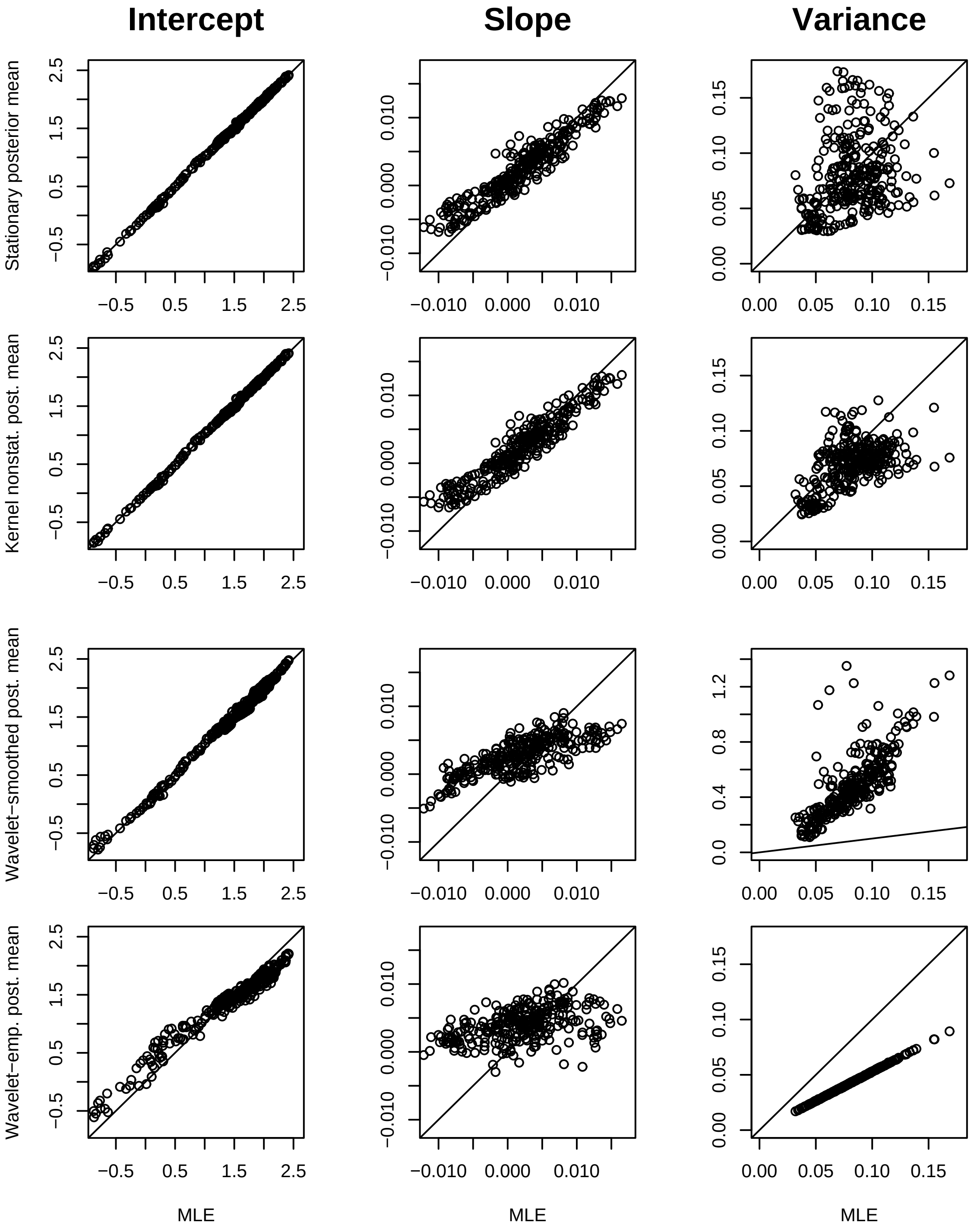

Figure 5.9. Scatterplots of model estimates (posterior means) of intercept (column 1), slope (column 2), and residual variance (column 3) fields compared to the MLE values for the four models: stationary (row 1), kernel nonstationary (row 2), wavelet-smoothed (row 3) and wavelet-empirical (row 4) for temperature variance.

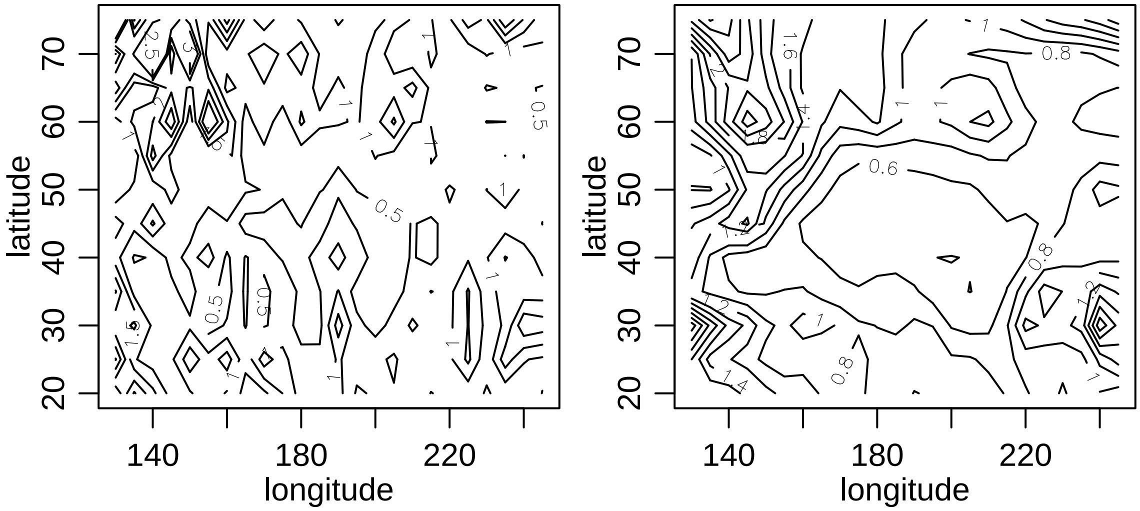

Figure 5.10. (a) Plot of standardized residuals (defined in text) as a function of location for temperature variance; these residuals are calculated based on the posterior mean parameters, but using the MLEs for the residual variances. (b) Plot of the ratio of the posterior mean residual variance estimates to the MLE variance estimates.

is no clear theory behind the thresholding, that the predictive performance of the wavelet models is poor (Section 5.7.3), and that the smoothing being done is not intuitive, I would be reluctant to use the wavelet models in the fashion employed in this work without further development.

In Figure 5.11, we see that for Eady growth rate, the slopes are smoothed much more toward zero in both the stationary and nonstationary models than was the case for temperature variance (Figure 5.9). Given that the predictive calculations suggest that the trends are not strong in the Eady growth rate dataset (Section 5.7.3), this level of smoothing is not surprising, and suggests that the model is correctly adjusting to the information in the data about the certainty in the trend estimates. For the variance estimates, the picture is similar to that for temperature variance; the nonstationary model estimates are somewhat similar to the MLEs, while the stationary model has some estimates that are much larger than the MLEs, suggesting that once again lack of fit in the correlation model is forcing the residual variance to rise to compensate.

The nonstationary model does appear to estimate nonstationary correlation structure, based on the posterior mean basis kernels shown in Figure 5.12 for temperature variance in the Pacific and in Eady growth rate in the Atlantic. We see that the basis kernels vary in size and orientation. However, the degree of nonstationarity based on the correlation structure induced by the basis kernel model is much less striking. In the following figures, I estimate the correlations between each of nine focal locations (the nine locations at which the basis kernels are centered) and all 288 other locations. For the two models in which the correlations are fit during the MCMC, I do this by calculating the induced correlations between locations at each MCMC iterate and averaging over the values. I compare these posterior mean correlations to the empirical correlations in the data by comparing correlation maps. In Figure 5.13, which shows the correlation structure for the kernel nonstationary model for temperature variance, we see that the correlation structure is somewhat nonstationary, primarily at 30° N, but that much of the variability in the basis kernels (Figure 5.12) has been smoothed out by taking the kernels to be spatial averages of the basis kernels. The correlation structure for the kernel nonstationary model for Eady growth rate is somewhat more nonstationary than for temperature variance (Figure 5.14), although the variability in the basis kernels is still greatly smoothed. Comparing the nonstationary correlation structure and the stationary correlation structure (Figures 5.15 and 5.16, for temperature variance and Eady growth,

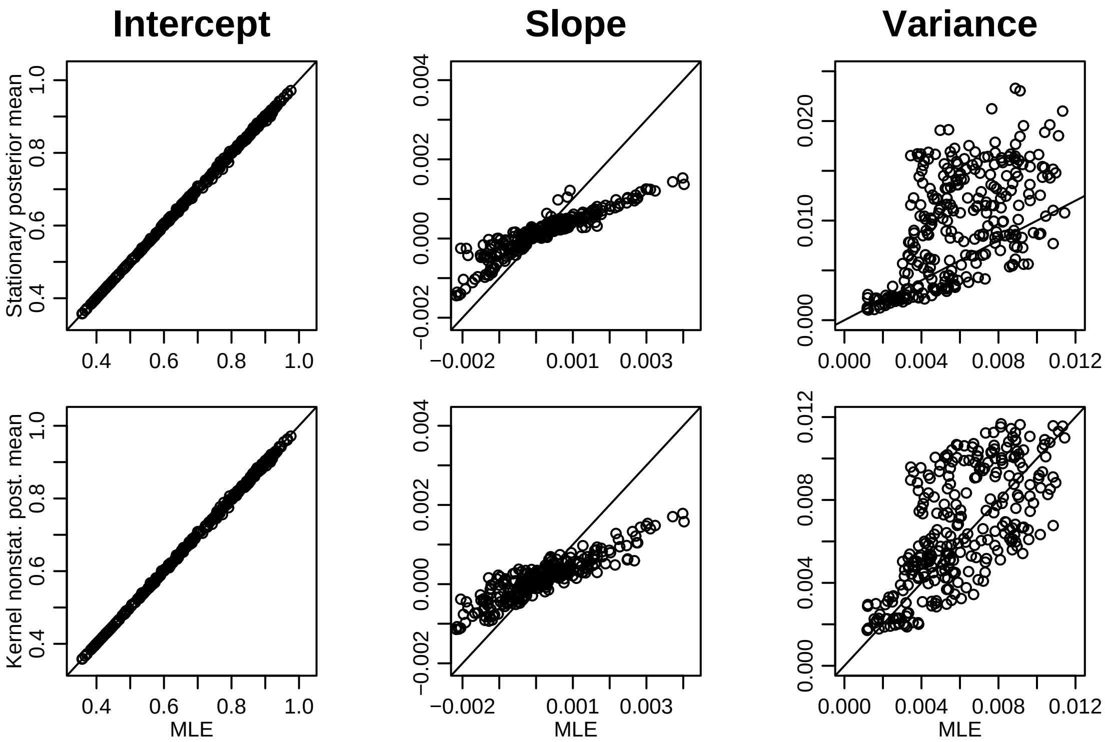

Figure 5.11. Scatterplots of model estimates (posterior means) of intercept (column 1), slope (column 2), and residual variance (column 3) fields compared to the MLE values for the models for Eady growth rate: stationary (row 1) and kernel nonstationary (row 2).

respectively) to the empirical correlations (Figures 5.17 and 5.18) we see that the nonstationary model is reflecting some of the gross patterns of the empirical correlation structure, but not all of the gross structure that one might expect the model to capture. This may be because the empirical correlations are noisy estimates that the model is correctly smoothing out, or because the basis kernel approach is not flexible enough to capture the true structure. Use of more kernels or a different approach to the spatial averaging that does less smoothing of the basis kernels might result in more nonstationarity being modelled. However, note that the model is free to choose the parameter that controls the degree of spatial smoothness. In contrast to the stationary and kernel nonstationary models, the wavelet correlations are fixed by the thresholding chosen. In Figure 5.19 we see that the wavelet-empirical model most closely matches the empirical correlation structure. Yet, this model drastically overfits (Section 5.7.3), suggesting that the structure is not matching the correlation structure of the test data. In Figure 5.20 we see that the wavelet-smooth method is is in fact smoothing out much of the empirical structure, while retaining much of the gross structure that is not retained in the nonstationary model. The poor predictive performance of the wavelet-smooth model could be occurring because it overfits the gross correlation structure, retaining structure that should be smoothed out. Alternately, the model may not correctly capture the intricate high-dimensional structure of the test data, perhaps because the test data do not satisfy certain constraints specified by the wavelet-smooth correlation matrix. One indication that this may be the case is that the correlation scale of the wavelet-smooth structure tends be longer than that of the stationary and kernel nonstationary models. Longer correlation scales tend to produce linear combination constraints amongst the locations. By using shorter scales the stationary and nonstationary correlation structures may avoid imposing such constraints on the test data (assuming that in the fitting process similar constraints are discouraged by the training data). This suggests that it is difficult to closely model the correlation structure and expect it to match the structure of test data, and also that modelling the structure may tend to give lower correlations than truly exist, if by doing so the model can avoid imposing constraints that are not satisfied by the data. The general lesson is that modelling the correlation structure of high dimensional data in a joint fashion is very difficult.

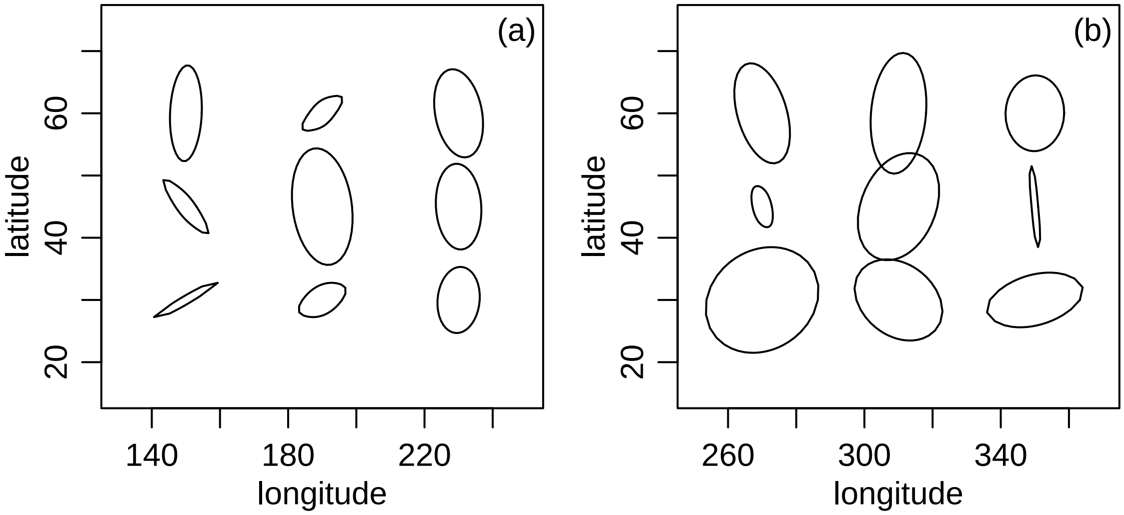

Figure 5.12. Mean basis kernels for (a) temperature variance in the Pacific region and (b) Eady growth rate in the Atlantic. For each posterior sample, I represent the kernel as a constant density ellipse from a normal density with covariance matrix equal to the basis kernel matrix. Mean ellipses are then plotted using the average distances from the ellipse origin to the ellipse itself at 44 different angles, averaged over the posterior samples. Note that since the kernels are plotted on a latitude-longitude grid, distances toward the top of the plot are exaggerated and the true basis kernels there are smaller in size than represented here.

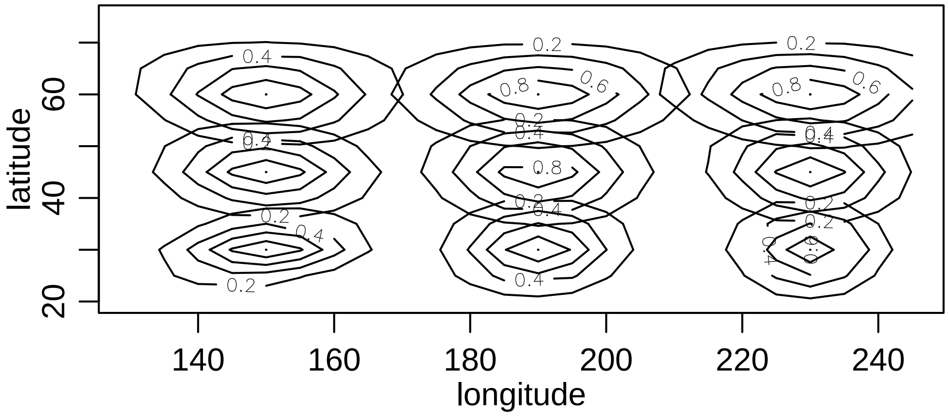

Figure 5.13. Plot of posterior mean correlation structure from the kernel nonstationary model for temperature variance between each of nine focal locations and all 288 locations. Correlation structures at the nine focal locations are overlaid on the same plot because correlations are less than 0.20 except in the bullseye areas. The nine focal locations are at the centers of the bullseyes and are the same locations as the centers of the basis kernels, as listed in the text.

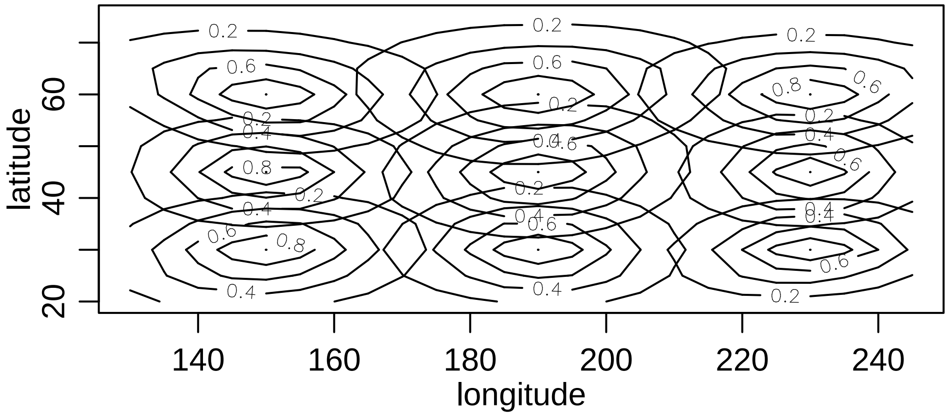

Figure 5.14. Plot of posterior mean correlation structure from the kernel nonstationary model for Eady growth between each of nine focal locations and all 288 locations. Plots are overlaid because correlations are less than 0.20 except in the bullseye areas. The nine focal locations are at the centers of the bullseyes and are the same locations as the centers of the basis kernels, as listed in the text.

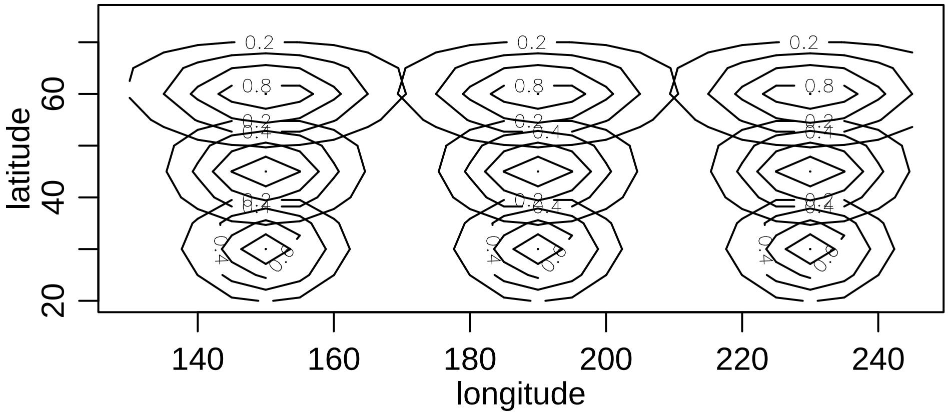

Figure 5.15. Plot of posterior mean correlation structure from the stationary model for temperature variance between each of nine focal locations and all 288 locations. Correlation structure appears different at different latitudes because of the distortion induced by the latitude-longitude grid. Other details are as in Figure 5.13.

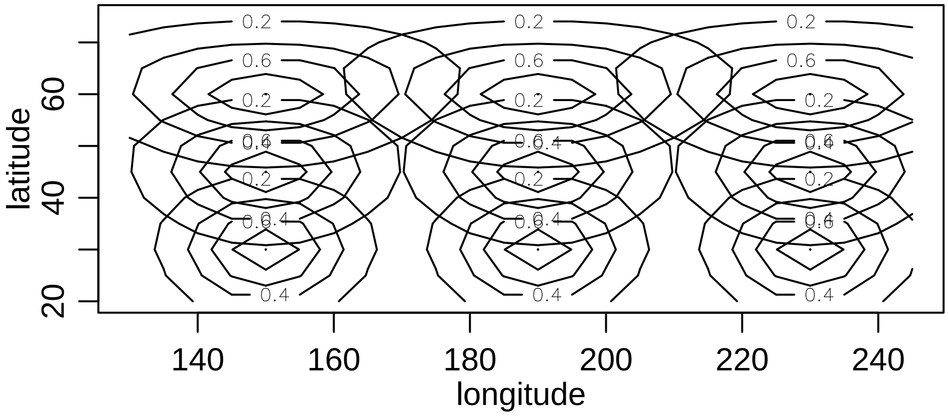

Figure 5.16. Plot of posterior mean correlation structure from the stationary model for Eady growth between each of nine focal locations and all 288 locations. Correlation structure appears different at different latitudes because of the distortion induced by the latitude-longitude grid. Other details are as in Figure 5.14.

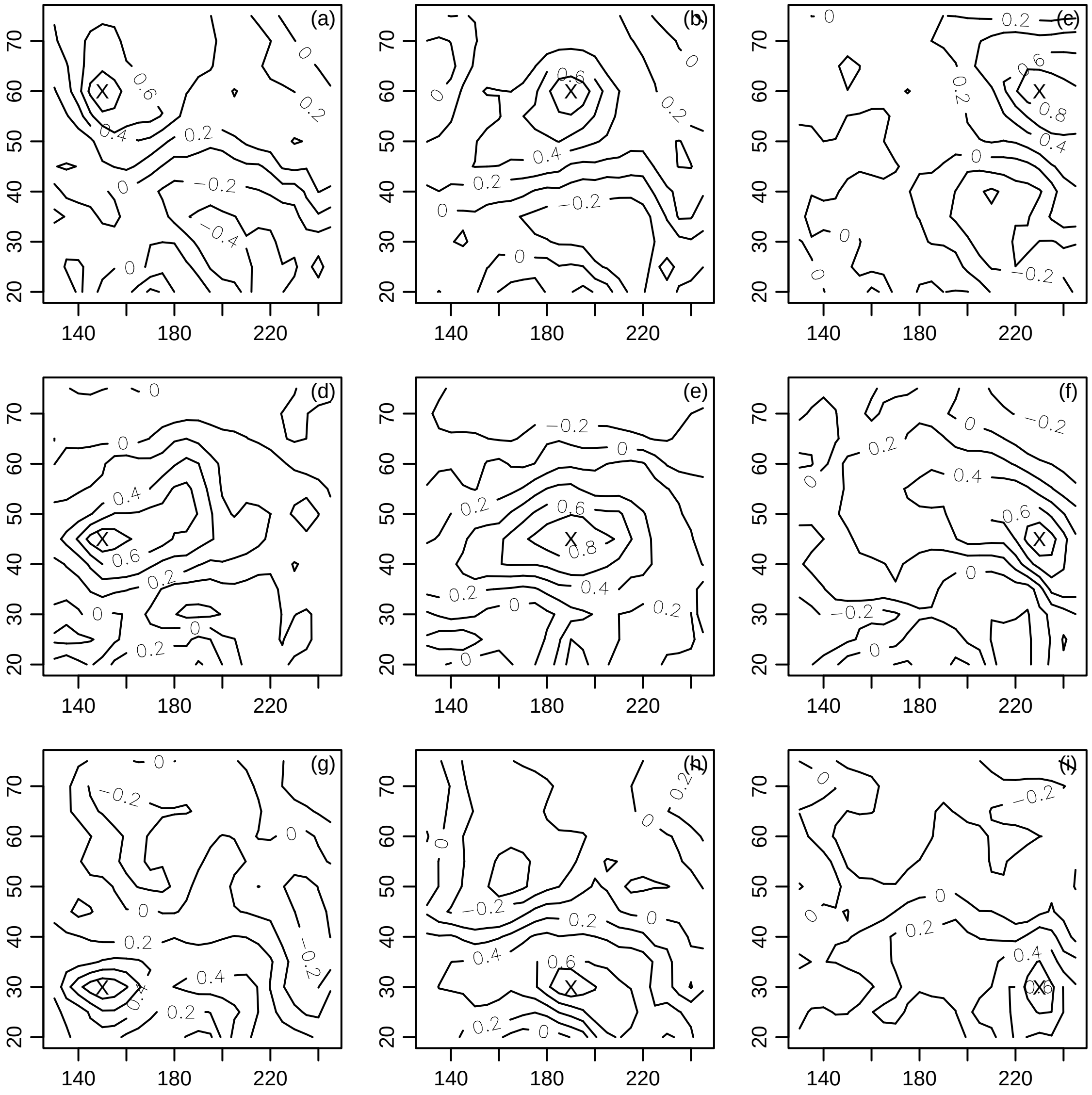

Figure 5.17. Plots of empirical correlations for temperature variance between each of the nine focal locations and all 288 locations. Each subplot displays the correlation structure for one focal location (marked by 'X') with latitude and longitude increasing from bottom to top and left to right respectively: (a) 150° *E,* 60° *N, (b)* 190° *E,* 60° *N, (c)* 230° *E,* 60° *N, (d)* 150° *E,* 45° *N, (e)* 190° *E,* 45° *N, (f)* 230° *E,* 45° *N, (g)* 150° *E,* 30° *N, (h)* 190° *E,* 30° *N, (i)* 230° *E,* 30° *N.*

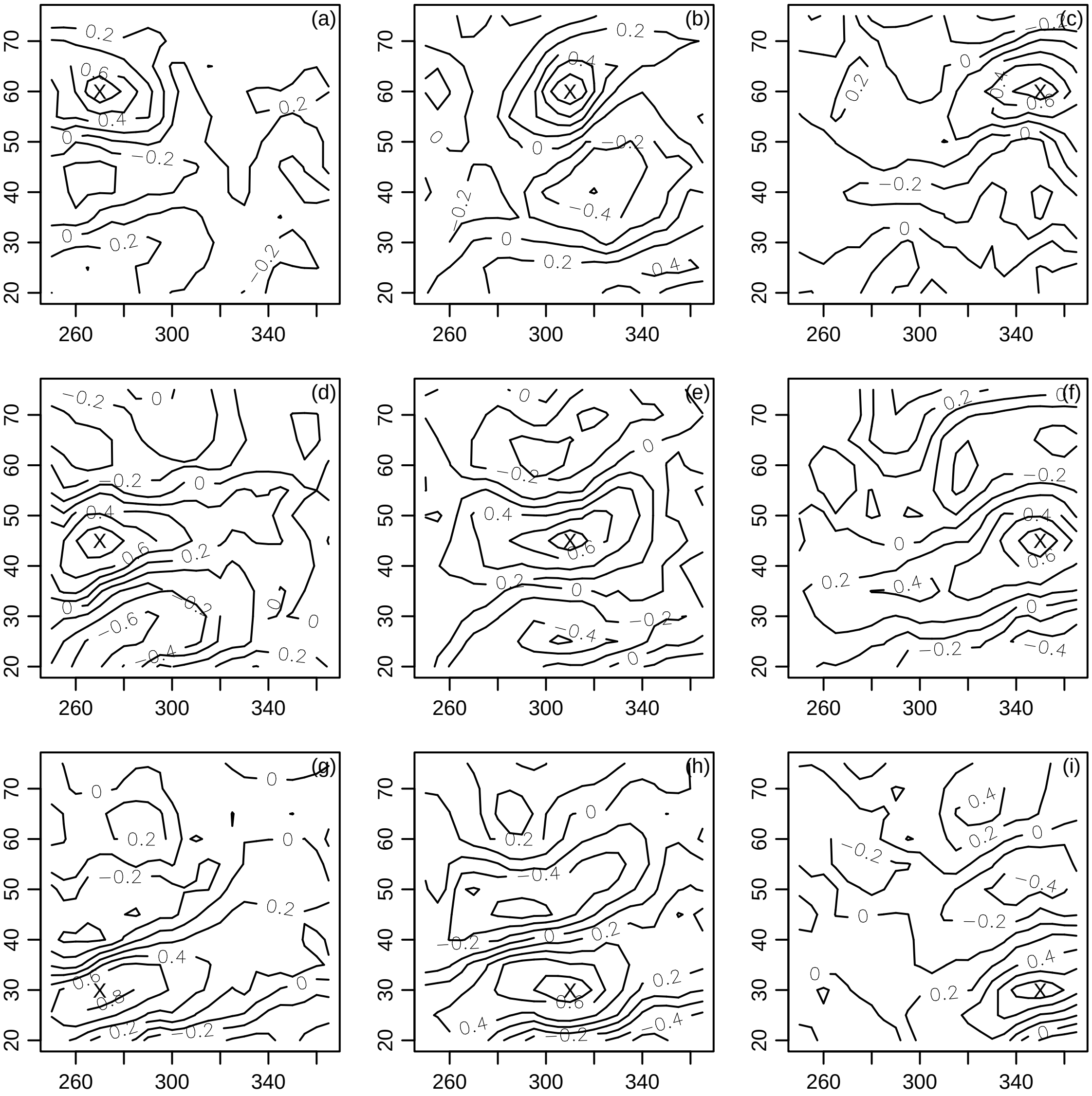

Figure 5.18. Plots of empirical correlations for Eady growth between each of the nine focal locations and all 288 locations. Each subplot displays the correlation structure for one focal location (marked by 'X') with latitude and longitude increasing from bottom to top and left to right respectively: (a) 150° *E,* 60° *N, (b)* 190° *E,* 60° *N, (c)* 230° *E,* 60° *N, (d)* 150° *E,* 45° *N, (e)* 190° *E,* 45° *N, (f)* 230° *E,* 45° *N, (g)* 150° *E,* 30° *N, (h)* 190° *E,* 30° *N, (i)* 230° *E,* 30° *N.*

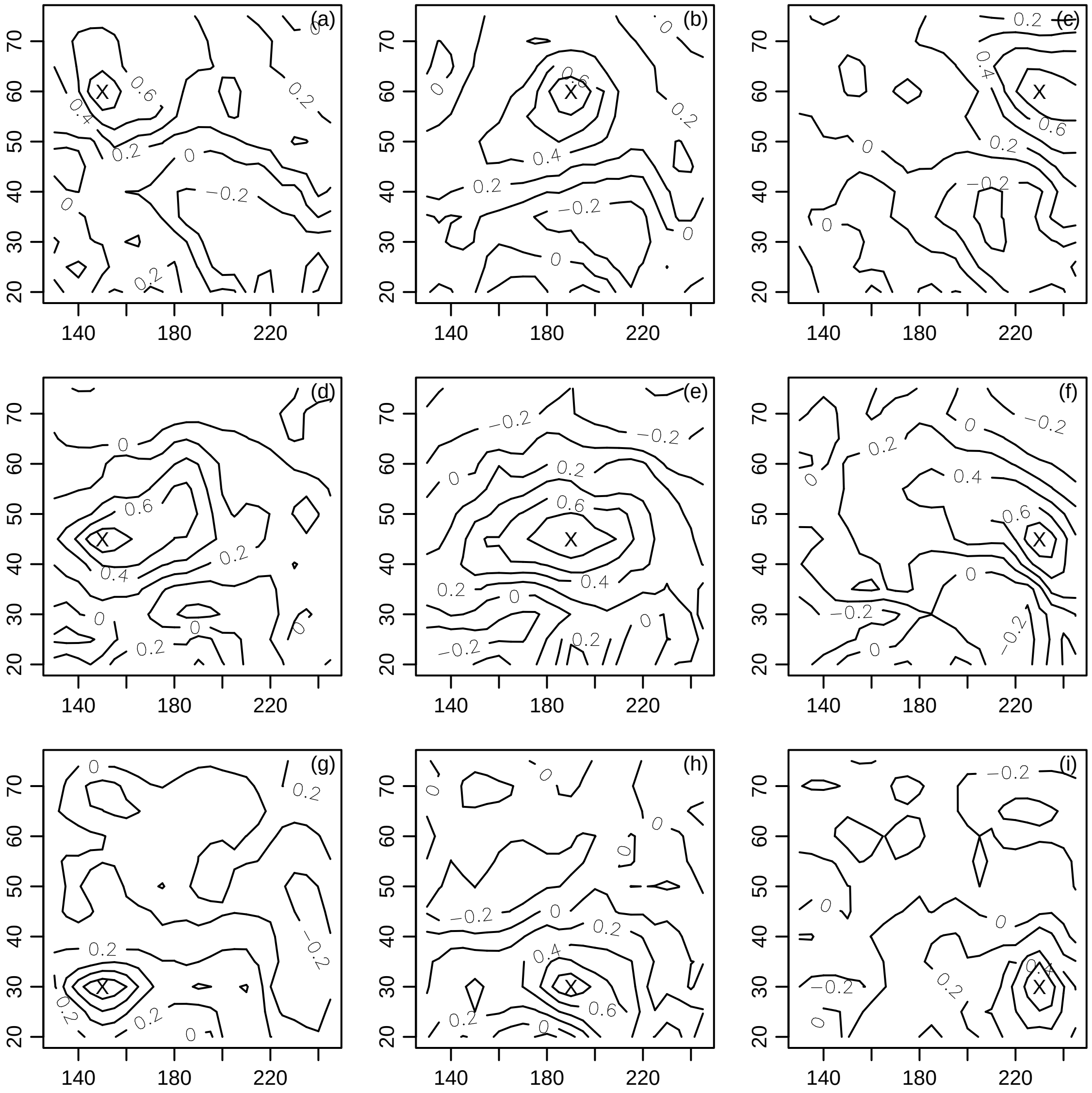

Figure 5.19. Plots of wavelet-empirical model correlations between each of nine focal locations and all 288 locations for temperature variance. Details are as in Figure 5.17.

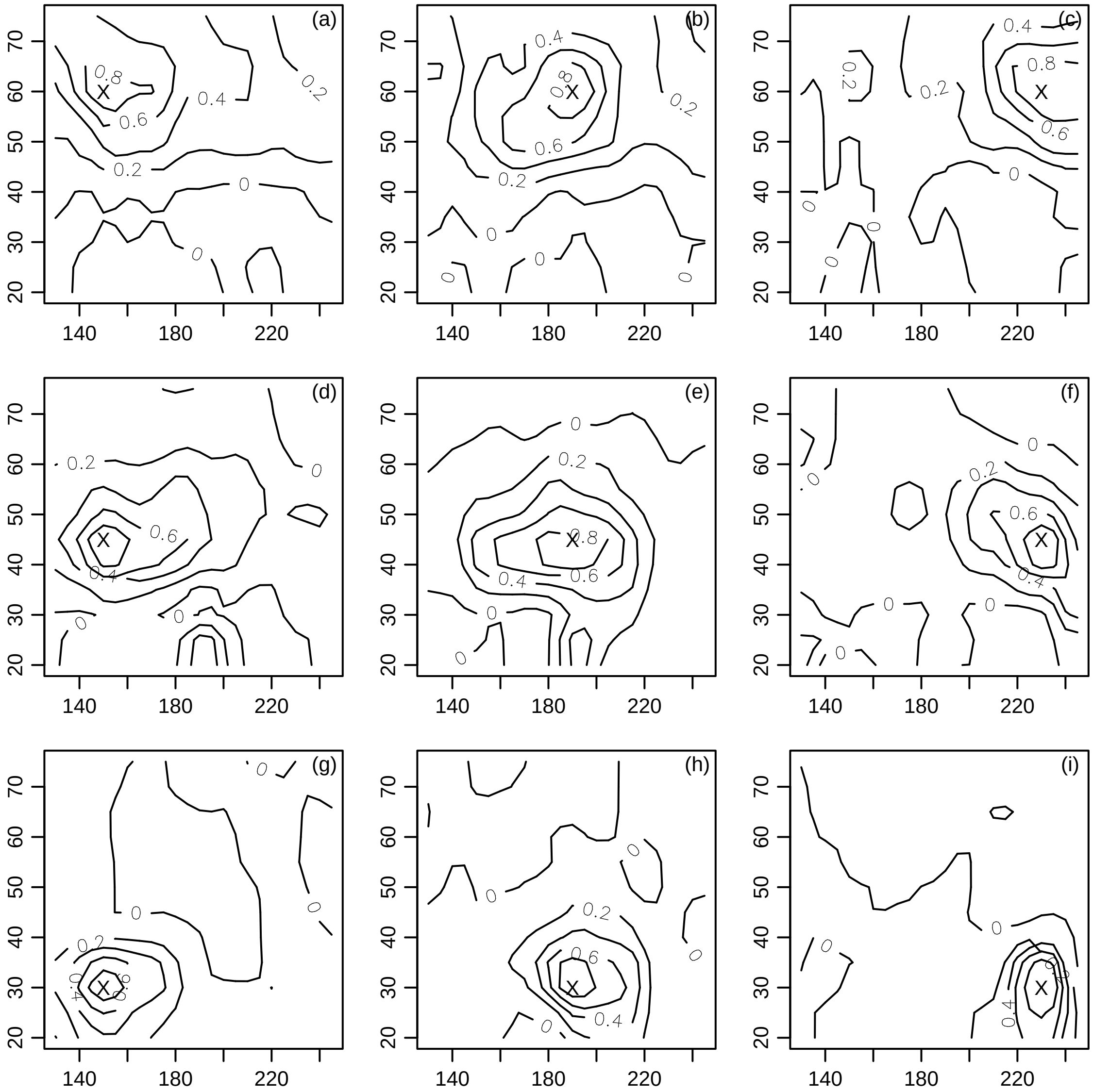

Figure 5.20. Plots of wavelet-smooth model correlations between each of nine focal locations and all 288 locations for temperature variance. Details are as in Figure 5.17.

5.7.3 Model comparison

I first assess the log posterior predictive density. In Table 5.1 I show the posterior predictive density values for test data for both datasets for the the six models. The kernel nonstationary model has the highest density. For the Bayesian models, I assess the uncertainty in these estimates by calculating the log predictive density from 10 subsets of 5000 contiguous MCMC samples and assessing the variability in the blocked estimates. This approach makes it clear that the stationary and kernel nonstationary methods give by far the highest log predictive densities and that the kernel nonstationary model is clearly better than the stationary model. Comparing the predictive density between the stationary and kernel nonstationary models by year, the kernel nonstationary is better in each of the seven test years (not shown), although the relative difference varies somewhat by year. Clearly when we assess performance in a way that takes account of the covariance structure of the data, the kernel nonstationary model is the best model. However, is this improvement substantial? If the observations were independent, we could consider the improvement in the predictive density on a per observation basis, by considering the difference in log predictive densities from the two models divided by the number of observations. This works for the time variable, since I consider the seven years to be independent, and the improvement in the log predictive density is 17.6 and 20.6 per year for temperature variance and Eady growth rate, respectively. If one is interested in predicting a whole year of observations, then one may be interested in the likelihood ratio for all locations which would be $\exp(17.6)$ or $\exp(20.6)$, suggesting great improvement based on the nonstationary model relative to the stationary model. However, if one is thinking about individual locations or wants to consider the improvement in predictive density per observation, the picture is clouded. It is not clear how many effective observations we have each year because of the correlation between locations. If all the locations were independent, the improvement would be only 0.061 or 0.071 per observation, corresponding to a likelihood ratio of 1.06 or 1.07. However, if there is substantial correlation between observations, the likelihood ratio is more compelling. Given the empirical correlations in the data, having the number of effective observations be one-fifth or even one-tenth of the number of nominal observations may be reasonable. Further investigation of this issue would involve estimating the effective number of locations in an appropriate manner, but I have not pursued this issue.

Table 5.1. Log predictive density comparisons of the six models on test data for the two datasets. Values in parentheses are ranges based on blocked values from the posterior simulations.

Model	Eady growth rate - Atlantic	Temperature variance - Pacific
MLE	3832	1267
MLE,$\beta \equiv 0$	3886	1393
stationary	6393(6375-6395)	4439 (4427-4441)
wavelet-empirical	not modelled	~ -100000
wavelet-smoothed	not modelled	2527 (2473-2529)
kernel-nonstationary	6537 (6524-6539)	4562 (4550-4565)

The MLE models perform poorly because I assume independence between locations, which is a poor model for the joint structure of the data. The wavelet-empirical model has a terrible predictive density, presumably because of overfitting. The wavelet-smooth model is not as unreasonable, but is much worse than the stationary or kernel nonstationary models. Note that based on the strange behavior of the wavelet models seen in the previous section, and the poor predictive performance on the temperature variance index seen here, I did not fit the wavelet models to the Eady growth rate data.

The results change somewhat when we look only at point predictions, as shown by the MSE values in Table 5.2. In the temperature variance case, the wavelet models are clearly not performing as well, with the wavelet-empirical model appearing to overfit drastically and the smoothing imposed by the wavelet-smoothed model decreasing predictive performance as well. The MLE models, stationary model, and kernel nonstationary model all give similar MSE values, with relative differences of only a few percent. For temperature variance, the differences do appear to be somewhat robust with respect to simulation error, as seen from the range of MSE values in blocks of the MCMC iterations, as well as when assessing MSE by year. Comparing the MSE values by year for the temperature variance case, the stationary model is worse than the kernel nonstationary model in all years while the kernel-based nonstationary model is better than the MLE model in one year, 1954. For Eady growth, the differences do not appear to be robust with respect to simulation error, as seen from the range of MSE values in blocks of the MCMC iterations, although the nonstationary model does slightly outperform the stationary model in all seven years. The nonstationary model outperforms the MLE model with slopes fixed at zero in one of seven years, while it outperforms the unrestricted model in only two of seven years, but in those two years (1949 and 1999, the extremal years), it sufficiently outperforms the MLE model so that averaged across all seven years, it has slightly lower MSE.

Given the relatively small differences in MSE, it is unwise to overinterpret the results. However, it does seem that the nonstationary model is performing slightly better than the stationary model. Also, in the case of temperature variance, the MLEs seem to perform better than the nonstationary model, suggesting that given the smoothness of the original data, there is little to be gained by borrowing strength across locations. In the Eady growth rate case, the slopes appear not to help with prediction, suggesting that for most locations, there is little real trend over time, although it is still possible that for a subset of locations, there are real trends that are masked by assessing only the effect of forcing all the slopes to be zero simultaneously. Unfortunately the full models do not appear to improve prediction by compromising between the MLEs and the joint null hypothesis. Note that the model comparison results might change dramatically with less smooth data. Also, fitting the full model does allow us to assess joint uncertainty in our estimates, as discussed next.

Table 5.2. MSE comparisons of the six models on test data for the two datasets. Values in parentheses are ranges based on blocked values from the posterior simulation.

Model	Eady growth rate - Atlantic	Temperature variance - Pacific
MLE	0.00796	0.0879
MLE,$\boldsymbol{\beta} \equiv \mathbf{0}$	0.00765	0.0904
stationary	0.00798 (0.00793-0.00805)	0.0901 (0.0891-0.0910)
wavelet-empirical	not modelled	0.155 (0.140-0.177)
wavelet-smoothed	not modelled	0.101 (0.0964-0.105)
kernel-nonstationary	0.00793 (0.00788-0.00798)	0.0887 (0.0882-0.0894)

5.7.4 Trend significance

Here I compare the trend estimates from the Bayesian nonstationary model with the MLEs. First, let's consider the temperature variance model. For the Bayesian model I use the posterior means and take the posterior standard deviation as a standard error, while for the MLEs, I estimate the standard error as $\frac{\widehat{\eta_i}^2}{\sum t^2}$. In Figure 5.21 I plot the standard errors as function of the point estimates for both models. The Bayes point estimates have been shrunk somewhat toward zero, but the most obvious difference is that the standard error estimates are on average about two-thirds as much for the Bayesian model as for the MLEs. It appears that accounting for the spatial correlation has made us more certain about the trend estimates. Using a conventional significance threshold of plus or minus two standard errors, the MLEs suggest that 103 locations have significant trends while the Bayesian model suggests 146 locations. Such an analysis does not account for simultaneously conducting 288 tests. In Paciorek et al. (2002) we used the False Discovery Rate approach to assess joint significance. Using the standard FDR method for independent data, for which Ventura et al. (2003) report successful results with spatially correlated storm activity data, 49 of the 288 locations are significant. To see the effect of the Bayesian shrinkage, I calculated the p-values and FDR result that would apply if one used the point estimates and standard errors from the Bayesian model, and found that I rejected 117 locations. Of course there is no theoretical justification for this, but it gives a sense for the impact of the changes in the point estimates and their standard errors in a multiple testing context. The Bayesian shrinkage has substantially increased our certainty in the estimates of the slopes. However, given that the Bayesian model produced MSE values that are slightly worse than those based on the MLEs (Table 5.2), it is not clear that such certainty is warranted.

Next let's consider the Eady growth rate data; the shrinkage results are very different here. In Figure 5.22 I plot the standard errors as function of the point estimates for both models. Once again, the estimates have been shrunk toward zero and the standard error estimates are much smaller, suggesting that accounting for the spatial correlation has made us more certain about the trend estimates. The linear trends in the Eady growth rate data are much less pronounced than in the temperature variance case. Using a conventional significance threshold of plus or minus two standard errors, the MLEs suggest that 49 locations have significant trends while the Bayesian model

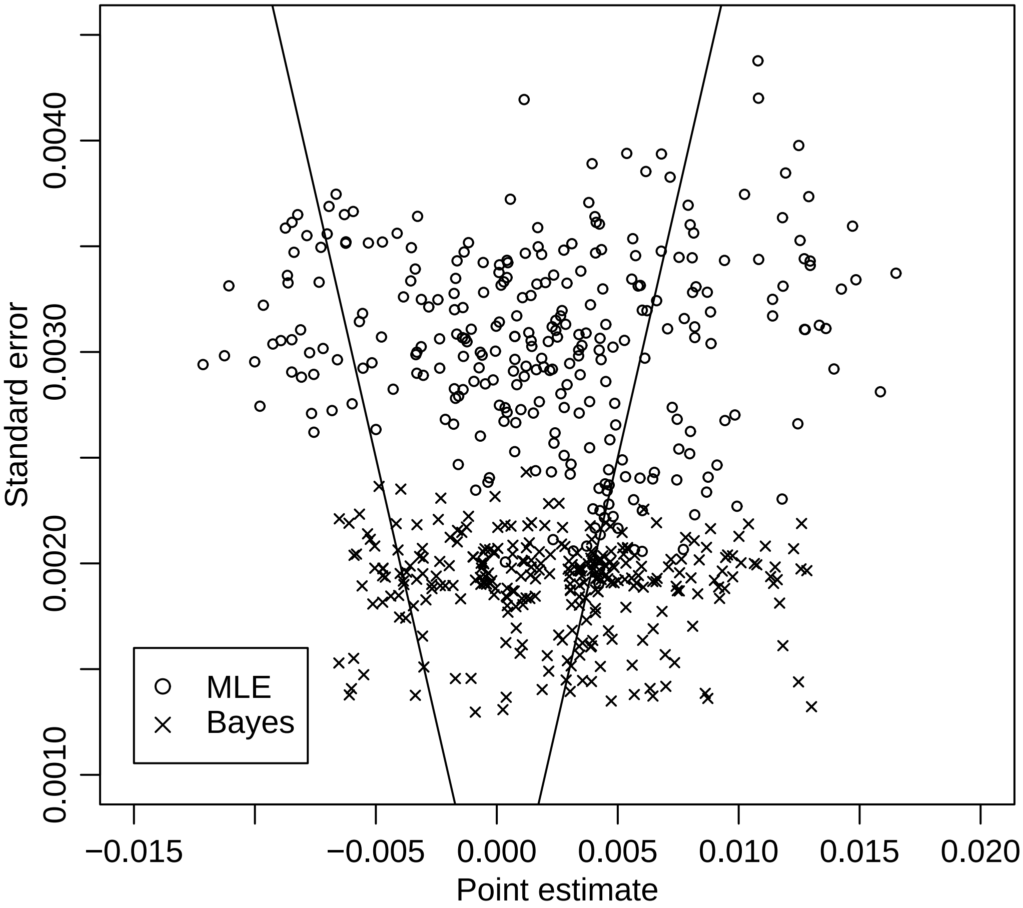

Figure 5.21. Scatterplot of standard error estimates as a function of the point estimates for the linear trends in both the MLE and Bayesian nonstationary models for the 288 locations of temperature variance in the Pacific. Points in the areas toward the outer sides of the plot relative to the nearly vertical lines are individually significant based on the point estimates being at least two standard errors away from zero.

suggests 12 locations. In contrast to the temperature variance data, the Bayesian model suggests many fewer locations are significant than based on the MLEs. Using FDR, 19 of the 288 locations are simultaneously significant. Once again, to see the effect of the Bayesian shrinkage, I calculated the p-values and FDR result that would apply if one used the point estimates and standard errors from the Bayesian model, and found that I rejected no locations.

One might also take a Bayesian approach to multiple testing based on the posterior sample. The first difficulty lies in defining H_0. One might define it in an ad hoc way as $\mathrm{H}_{0,i} : \beta_i \tilde{\beta}_i < 0$, namely that the true slope is of the opposite sign from the posterior mean. Of course this hypothesis depends on the data, which is anathema to a frequentist, but poses no real problems to the Bayesian. Another concern is that this is essentially a one-sided test in which the side is determined based on the data, so this approach will result in more significant locations than a classical approach because the level of the test is essentially twice that in the classical test. However, if one proceeds in this fashion, one might choose to reject the null for all locations at which $P(\mathrm{H}_0|\boldsymbol{Y}) < 0.05$. Such an approach satisfies a Bayesian version of the FDR criterion because

$$\mathrm{E}(\text{FDP}|\boldsymbol{Y}) \leq 0.05, \tag{5.7}$$

where FDP is the proportion of false rejections (rejections for which the null is actually true). This is just the classical FDR, except for the conditioning on $\boldsymbol{Y}$. In fact, so long as the average value of $P(\mathrm{H}_{0,i}|\boldsymbol{Y})$ taken over the rejected locations is less than 0.05, the expectation property (5.7) will hold:

$$\begin{aligned} \mathrm{E}(\text{FDP}|\boldsymbol{Y}) &= \mathrm{E}\frac{n_{fr}}{n_r} \\ &= \mathrm{E}\frac{\sum_i I(\mathrm{H}_{0,i}|\boldsymbol{Y})}{n_r} \\ &= \frac{\sum_i \mathrm{E}I(\mathrm{H}_{0,i}|\boldsymbol{Y})}{n_r} \\ &= \frac{\sum_i P(\mathrm{H}_{0,i}|\boldsymbol{Y})}{n_r}, \end{aligned}$$

where n_r is the number of hypotheses rejected and n_{fr} is the number rejected falsely, which is a random variable. We estimate $P(\mathrm{H}_{0,i}|\boldsymbol{Y})$ using the posterior samples, performing the calculation location by location; the effect of the other locations is already incorporated through the fitting

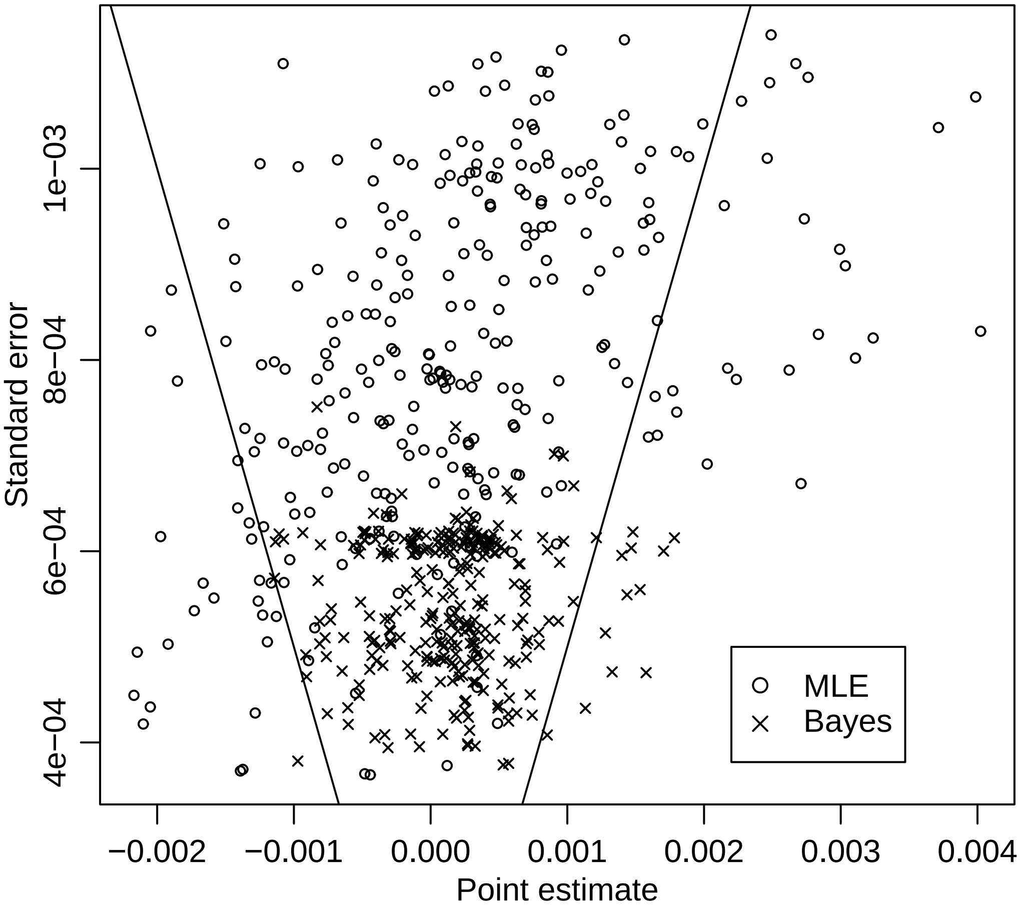

Figure 5.22. Scatterplot of standard error estimates as a function of the point estimates for the linear trends in both the MLE and Bayesian nonstationary models for the 288 locations of Eady growth rate in the Atlantic. Points in the areas toward the outer sides of the plot relative to the nearly vertical lines are individually significant based on the point estimates being at least two standard errors away from zero.

process. The result using this Bayesian version of FDR is that many more locations are rejected than in the frequentist FDR approach. For temperature variance, if we reject only those locations with posterior probability of the null less than 0.05, 170 are rejected (146 if we use 0.025 to account for the use of a ’one-sided’ test). If one is willing to reject locations with posterior probability of the null exceeding 0.05, so that the Bayesian FDR is very close to 0.05, one rejects 241 of the 288 locations (based on rejecting the locations with the smallest posterior probabilities). Note that these are more than are rejected using the Bayesian estimates to get a ‘p-value’ and then using the classical FDR, which suggests that not only are the smaller standard errors having an effect, but also that the use of the Bayesian approach to FDR plays a role. For Eady growth rate, the Bayesian approach accounts for the apparent lack of robust trends; 17 locations are rejected with posterior probability of the null less than 0.05, while 12 are rejected if we use 0.025, and 33 if we reject such that the average probability of the null is 0.05. Thus the Bayesian approach suggests that for temperature variance most of the locations have slopes that can be distinguished from zero, but that the point estimates of those slopes are closer to zero than the MLEs. For Eady growth rate, the Bayesian approach suggests that few locations have real trends. The optimism on the part of the Bayesian approach for the temperature variance data is rather troubling given that the cross-validation results for MSE in Section 5.7.3 give little indication that the trends are so certain. However, it may also be the case that the FDR approach is overly conservative. It is illustrative to consider an extreme example of positive dependence amongst the test statistics. Suppose the test statistics are perfectly correlated. In that case, we are really only doing one test and would want to reject if $p < 0.05$. However, the standard FDR approach will only reject if $p < \frac{.05}{n}$, a much harder threshold to reach. Also, the Bayesian approach does not appear to be overly optimistic in the case of Eady growth rate.

We might also take the approach of trying to determine locations at which the slopes can be determined with high certainty to not be near zero. This is motivated by the fact that there might be small trends that with enough data we could be certain were not zero, but are not substantively of interest. For example, in conjunction with the subject matter experts, we might decide that trends less than some value ϵ are not substantively of interest. One would then determine $P(|\beta_i| > \epsilon)$ and apply the Bayesian FDR approach outlined above to determine a set of locations for which the FDR for this new type of null hypothesis is less than, say, 0.05. Of course, this requires choosing a value for ϵ. A variation on the theme is to plot the number of rejected hypotheses as a function of different values of ϵ. I have not pursued these possibilities here as in-depth data analysis is somewhat outside the scope of this work. Also the fact that the temperature variance index modelling is done on the log scale makes it somewhat difficult to interpret the substantive importance of different values of ϵ.

5.8 Discussion

5.8.1 Covariance modelling assessment

Accounting for spatial covariance in a rigorous manner when fitting trends at multiple locations is a difficult task. Even after reducing the number of locations from 3024 to 288, fitting the full spatial model was very computationally intensive, and the MCMC exhibited very slow mixing. This was the case for the stationary covariance model and the models in which the wavelet-based correlations were fixed, as well as the full kernel-based nonstationary model.

Based on cross-validation, the kernel-based nonstationary model seems to offer some improvement over the stationary model. The nonstationary model reflects some of the local correlation structure, although it is unable to model irregular, long-distance, and negative correlations. In both datasets, the MLEs outperform the stationary model estimates in terms of point estimates for test data, which suggests that unless one is willing to use a nonstationary model, one may not want to model the correlation structure. Of course, if one wants to predict off-grid, some sort of model is necessary, although for point predictions, one might just consider smoothing the MLE fields. For one dataset the MLEs outperform the nonstationary model in terms of MSE, while in the other the nonstationary model performs slightly better. For less smooth data, there may be much greater benefit to fitting the full model rather than relying on the MLEs. With respect to the posterior predictive density, which assesses how well the models jointly predict test data assuming the normal likelihood is reasonable, the nonstationary model is clearly better than the other models. However, on a per-location basis, the gain relative to the stationary model is small.

A less smooth dataset may allow a better evaluation of the various covariance models and better let us determine how critical is the choice of covariance structure in smoothing noisy data. One potential dataset is the storm activity index based on extreme wind speeds (Paciorek et al. 2002), which is much less smooth than the storm activity indices used here. Another possibility is the storm count index in Paciorek et al. (2002). However, these latter index values are non-normal count data and modelling them would require additional methodological development (discussed in Section 6.3). Climatological fields can range from the very smooth, such as pressure fields, to the very noisy, such as precipitation and wind fields. It seems plausible that the advantages of modelling the fields and thereby smoothing the observations will be largest for the noisy fields.

There are several important limitations built into the nonstationary model constructed here. The model cannot account for long-distance and negative correlations, and the use of simple Gaussian kernels limits the correlation structures that can be modelled. These limitations may be important in many datasets, including possibly the storm activity data, as suggested by the long-distance and negative correlations in Figure 5.1. I do not have any particular suggestions for modelling long-distance or highly irregular correlation structures, apart from the wavelet decomposition approach explored in this work, but some relatively simple approaches may model some of the negative correlation structure. One idea is to use kernels that are allowed to take negative values. However, even if such kernels could be parameterized sufficiently simply, the advantage of the closed form expression based on Gaussian kernels (2.5) would be lost. Another potential drawback is that the resulting correlation function is somewhat non-intuitive, because the product of two negative values is positive. An alternative is to use a product of the Matérn nonstationary correlation function used in this work and a stationary correlation function that can take negative values, such as those defined on the sphere (Das 2000) and in $\Re^3$ (Abrahamsen 1997), although this approach has the same drawback as described at the end of the next paragraph.

I have used stationary models that allow only non-negative correlations, but I could easily use the stationary correlation functions mentioned above that take negative values. It would be interesting to see if such correlation functions better fit the data than the Matérn stationary correlation function fit here. However, I suspect that the improvements would be minimal, because the negative correlations seen in Figure 5.1 occur in latitude bands. It does not seem likely that stationary correlation functions that require negative correlations to occur in equidistant rings around the focal location would fit such data well.

In the form that I have used here, the wavelet approach does not perform well. This work points out several areas in which the approach needs further development. First, it is unclear how much thresholding to do or exactly how to carry out the thresholding. I tried two levels of thresholding and neither performed well when incorporated into the full spatial model. One might argue that an intermediate level of thresholding would work much better, but how does one find that level? In particular, this highlights the need for a criterion by which to choose the level of smoothing. Nychka et al. (2001) report that with the thresholding they chose, they could mimic a stationary Matérn correlation with little error in terms of element by element differences between the true correlation and the approximation. However, this was done by smoothing the true correlation matrix, not by smoothing an empirical correlation matrix. In my experience applying the approach to empirical covariance matrices, it is very difficult to decide on the thresholding. Furthermore, even if one is able to choose a level of thresholding that appears satisfactory on an ad hoc basis, there are several reasons for concern. First, covariance matrices are very tricky to work with because of the constraints involved in being positive definite, and they involve very high-level structure when modelling many locations. Indeed, once the level of correlations is high enough, the values at some locations are for all practical purposes linear combinations of the values at other locations. It is entirely feasible that one could choose a smoothed covariance that by eye seems reasonable but in reality does not fit the data well at all when used in a likelihood-based context. This would occur if the data violate a linear combination constraint imposed in the covariance matrix. This lack of fit seems to be the case for the wavelet-thresholded covariance that smoothed the empirical covariance. This covariance gives a very low log-likelihood to the training data when embedded in the Bayesian model, presumably because the thresholding was done without reference to the likelihood, unlike the MCMC fitting of the stationary and kernel nonstationary models. Second, based on my experience here with the wavelet covariance that closely matched the empirical covariance, overfitting is a real concern. The log-likelihood of training data with the wavelet-empirical covariance was extremely high, but the spatial model with this covariance did a very poor job of predicting test data.

If one is interested merely in smoothing the empirical covariance for the purpose of display, ad hoc thresholding may be sufficient, but for model building and extrapolation to other locations, the exact covariance structure and the inherent high-level structure is critical. One potential avenue for development of the wavelet approach is to consider more rigorous criteria for optimizing the thresholding. Choosing the thresholding with reference to a likelihood or similar criterion is one possibility, possibly in conjunction with cross-validation, although this might limit the computational advantages of the approach. Other loss functions for estimating covariance matrices (Yang and Berger 1994) might be useful.

5.8.2 Other approaches

Given the difficulties encountered in the covariance modelling done in this work, other approaches may be more practical, albeit possibly more ad hoc. Holland et al. (2000) smooth estimated trends based on a kriging style approach, but with the ‘data’ covariance based on jackknifing the trend estimates. Such resampling procedures may be the most practical and computationally feasible approaches to data such as these. An alternative that makes use of the model development in this work would be to employ more empirical Bayes methodology to fix as many parameters as possible without unduly influencing the final inference. Hopefully this would get around some of the mixing issue encountered here as well as speeding the computations. Wikle et al. (1998) build hierarchical space-time models in which the analogues to my $\alpha(\cdot)$ and $\beta(\cdot)$ processes are given Markov random field priors, and spatial structure within the residual fields is modelled using nearest neighbor autoregressive models with spatial dependence on the field at the previous time point. Their approach is to build more complicated temporal models than are used here, while attempting to model the spatial structure in a relatively simple nearest neighbor fashion that avoids specifying a joint spatial covariance structure explicitly. This may be more computationally feasible than my approach, but it’s not clear if the spatial structure is sufficiently flexible to capture the important structure in the data.

In problems with many locations, such as with the full storm activity dataset of 3024 locations, computationally-efficient methods are needed. While computationally feasible, smoothing the empirical covariance using the wavelet approach suffers from lacking a defined objective function for optimizing the degree and structure of the smoothing. Further work on this approach is needed.

A final area that I have not addressed in this work involves the linearity of the temporal model. Relaxing the assumption of linearity in the temporal model is an obvious choice for improving the model. In Paciorek et al. (2002) we present evidence of nonlinearity in all the indices for at least some locations. Two main issues arise in moving to nonlinear models. The first is how to parameterize and fit such models. The goal is to extract useful high-level information about long-term trends; deciding how to and the extent to which to partition between signal and noise is a difficult task. One possibility would be to represent the time series at each location by regressing on an orthogonal basis, such as low-order orthogonal polynomials, and then independently smooth the estimated coefficients. Because of the orthogonality, one could then estimate the smoothed trend using the smoothed coefficients at each location. Of course one still needs to account for the residual spatial correlation in estimating the coefficients, which becomes an even larger challenge if one moves away from simple linear models. The second issue is that even if one is able to estimate nonlinear trends, displaying the results is an important challenge. Linear trends can be portrayed as a unidimensional quantity on a map. Nonlinear trends cannot be so easily summarized. One possibility is to display the nonlinear trends as time series plots for a moderate number of representative locations.