Chapter 6

Conclusions and Future Work

6.1 Summary and Contributions

This thesis makes several original contributions. The first contribution is my extension of the Higdon et al. (1999) (HSK hereafter) kernel convolution nonstationary covariance function. The HSK covariance is a straightforward way to construct a flexible nonstationary covariance function, albeit with some restrictions, such as non-negativity. The HSK covariance function is a generalization of the well-known squared exponential stationary correlation function. When used in Gaussian process (GP) distributions, both produce sample functions that are infinitely differentiable in both a mean square and sample path sense. This high degree of smoothness is generally undesirable. In Chapter 2, I generalize the HSK covariance function to produce a class of nonstationary correlation functions, one of which is a nonstationary version of the Matérn correlation, which has a parameter that controls the degree of differentiability. I prove that the smoothness properties of the new nonstationary correlation functions follow from those of the stationary correlation functions on which they are based, under certain smoothness conditions on the kernels used to construct the nonstationarity. In Chapters 4 and 5, I show that the Matérn nonstationary correlation can be used within a Gaussian process prior in nonparametric regression and spatial smoothing models.

Given the class of nonstationary covariance functions, we need a way to parameterize kernels that vary smoothly in space. In Chapter 3, I describe one parameterization based on the eigendecomposition of the kernel matrices and an overparameterized model for the eigenvectors. This

parameterization seems to be feasible for two- and three-dimensional covariate spaces, but the number of parameters, and presumably the difficulty in fitting the model, increases with the dimension of the covariate space. In higher dimensions, simpler alternatives may be required and even abandoning nonstationarity may be a reasonable approach. When GPs are embedded in a hierarchical model, or used in any nonconjugate fashion, the process cannot be integrated out of the model. This nullifies the general approach taken to fitting GP models via Markov chain Monte Carlo (MCMC) and causes major problems in MCMC convergence. I introduce a sampling scheme, which I call posterior mean centering, that can be used in some cases in which the process cannot be integrated out of the model. The scheme greatly improves mixing of the GP hyperparameters, although mixing is still quite slow. Hyperparameter mixing is important because the covariance hyperparameters control the flexibility of the function; they perform the same role that the number of knots does in a spline model. The sampling scheme can be applied successfully to generalized nonparametric regression models with non-Gaussian response using an analogue to the familiar iteratively-reweighted least squares algorithm, as I describe in Chapter 3 and demonstrate in Chapter 4. Unfortunately, when Gaussian processes are used to construct the kernel matrices of the nonstationary kernel covariance, there is no way to make use of the PMC scheme, and we are left with a naive, slowly-mixing scheme for sampling the processes controlling the kernel structure.

The nonparametric GP regression model that I define based on the generalized kernel convolution covariance successfully models a number of experimental datasets, when the dimension of the covariate is between one and three. In one dimension, the method is successful provided the function does not change too sharply, but the BARS method of DiMatteo et al. (2002) is more successful under all conditions examined. In two and three dimensions, the nonstationary GP model outperforms free-knot spline methods generalized to higher-dimensional covariate spaces, as well as a stationary GP model, although the improvement is relatively limited in some cases. While these results are encouraging, successful comparison with standard nonparametric regression methods such as kernel smoothing, wavelet regression, and neural network models, would strengthen my conclusions, and I hope to perform additional comparisons. Based on the current results, for datasets in which a high degree of inhomogeneity is expected, use of a nonstationary GP model may be advantageous, such as in two- and three-dimensional geophysical datasets. However, in

many regression settings, the inhomogeneity may not be drastic and the spline methods (which are nonstationary in nature) and stationary GP model may perform well, with the advantage of being faster computationally and having less complicated parameterizations. The GP model gives smooth function realizations and allows one to model the degree of differentiability of the function, as well as allowing one to place a prior on the degree of flexibility of the function using the trace of the smoothing matrix. In the former aspects, these features of the model are advantageous relative to some of the spline-based models, which must deal with the difficulty of generalizing one-dimensional spline basis functions to higher dimensions. The GP model moves smoothly between functions with varying degrees of flexibility, based on the sizes of the kernels in the nonstationary case and the correlation scale parameters in the stationary case. These parameters control the implicit model dimension and allow one to embed a range of models in one structure without the need for reversible jump MCMC, but this comes at the cost of slow mixing and computation. The model structure implicitly favors smoother functions if these are consistent with the data, because of the Occam's razor effect. This aspect of the Bayesian modelling approach is very attractive, since it encompasses our desire for simpler models, all else being equal. This type of prior information is an aspect of Bayesian modelling that even diehard frequentists may find appealing, since a subjective preference for simpler models is widespread and is usually taken into account implicitly in the choice of likelihood.

I also use the nonstationary correlation function to model correlation in the residual structure of a spatial model designed to jointly assess trends in time of climatological data at multiple spatial locations. The nonstationary model better accounts for the residual covariance structure than do a stationary model and a wavelet-based model. However, probably because the data are pre-smoothed by a deterministic climatological model, the nonstationary model does not predict held-out time points any better than simply using the maximum likelihood estimates. The full Bayesian model does allow one to jointly assess trend significance and compare the results to the frequentist False Discovery Rate (FDR) approach to multiple testing. The Bayesian model shrinks both the point estimates and the estimated uncertainty in those estimates by borrowing strength across space. For one index of storm activity, the Bayesian model indicates the existence of more real trends than does the FDR approach, while for a second index, the Bayesian model indicates

that few locations have strong trends, somewhat fewer than the FDR results. The nonstationary GP method also allows one to perform inference at unobserved locations, unlike some methods for smoothing the empirical correlation structure. Perhaps the most important result from the spatial model analysis is that the comparison of covariance models indicates the difficulty in flexibly fitting residual covariance without overfitting, as I discuss further in Section 6.3. Accounting for residual spatial structure is important in assessing trends in spatial variables and is an area in need of more research.

6.2 Shortcomings and Potential Improvements

Here I describe a number of drawbacks to the nonstationary Gaussian process approach to nonparametric regression and spatial modelling.

The first obvious shortcoming of the model is computational speed. Fitting the model via MCMC involves computing the Cholesky factor of n by n covariance matrices within every iteration of the Markov chain. This is the case even in the regression model in which the function can be integrated out of the model because of the use of GPs to model the eigenstructures of the kernels. Ongoing work in the machine learning community on reduced rank approximations to the covariance may introduce techniques for improving the speed of the models. Another possibility is to use simpler functional forms for processes high in the model hierarchy where the full flexibility of a GP prior may not be needed. The difficulty with GP models is that even though the implicit kernel smoothing that is being performed is essentially local, the computations involve the full covariance matrices and do not make use of sparsity in any sense.

The difficulties involving computational speed are compounded by the slow mixing of GP models. The PMC method introduced here improves the mixing of the hyperparameters in generalized nonparametric regression models, while Langevin updates improve the mixing of the process values. If it can be extended to work with numerically singular covariance matrices, the adaptive reparameterization of Christensen et al. (2003), possibly in conjunction with PMC, may greatly improve mixing, although I suspect that datasets with thousands of observations will still be difficult to fit. However, neither the adaptive reparameterization nor the PMC scheme will work with the GPs used for the kernel covariance structure. This suggests that a parameterization of the kernel

structure that is easier to fit may be worth pursuing.

In Chapter 4, I demonstrate that the nonstationary GP model is unable to capture a sharp jump in the regression function without undersmoothing in the neighborhood of the jump. This is inherent in the smooth parameterization of the kernels used to construct the nonstationary covariance function. One possibility that may help somewhat in capturing jumps is the use of asymmetric kernels. This would allow a point to have high correlation with points to one side where the function is smooth and low correlation with points on the other side where the jump occurs. In higher dimensions, this might involve asymmetry across a hyperplane; such non-Gaussian kernels would seem to be difficult to model. Ultimately, the use of a standard kernel shape restricts the types of functions that can be approximated by the GP model. Models that can adaptively add and delete basis functions, such as the spline models and neural network models, may perform better in some cases, although fitting such models and assessing convergence can be difficult, just as in the GP case.

The eigendecomposition model for the kernel structure seems feasible in two and three dimensions, but the number of parameters increases quickly with covariate dimension, even with the simplified scheme of sharing a single correlation structure across the eigenprocesses. However, simpler parameterizations carry the danger of not being as flexible in modelling features of the data. It may be that there are other parameterizations that do a better job of capturing features of the data without having too many parameters or causing mixing problems.

In addition to being faithful to the real features of the data, a regression model should also ignore covariates that appear to be independent of the response. In some initial experimentation, it appeared that my parameterization of the nonstationary GP model did a poor job of ignoring unimportant covariates, in part because this requires that the relevant correlation scale parameter (in the nonstationary model, this is the size of the kernel in the direction of the covariate) become very large, so that the response is highly correlated in the direction of the unimportant covariate. This is difficult to model without causing mixing problems, since as the parameter increases, the likelihood becomes quite flat. One possibility for addressing this shortcoming would be to reparameterize the model so that once that parameter becomes sufficiently large, one enters a discrete part of the parameter space in which the covariate is completely ignored. Such an approach would have the

benefits of ignoring unimportant covariates, allowing one to assess which covariates are relevant, and improving mixing by removing a troublesome part of the covariate space, but would require moving between models of different dimension during the MCMC.

Examination of the empirical correlations in Chapter 5 suggests that the covariance structure of the storm data is complicated; such complicated structure is likely to be present in many spatial datasets. The basis kernel construction of the nonstationary covariance seems to capture only a portion of the local non-negative nonstationary structure apparent in the empirical correlations. This may be because the apparent structure is noise that the nonstationary model is correctly ignoring. However, it seems plausible that there is structure present that the model is unable to capture for some reason. There are structural reasons that the nonstationary model may not be able to capture the true underlying correlation patterns, including its inability, as parameterized here, to account for the long-distance and negative correlations, and the constraints on the correlation structure induced by using the particular kernel form of simple Gaussian kernels. Application of the model in this thesis requires the assumption that these limitations have little effect on the resulting posterior distribution. While this may be reasonable in many applications, methods for dealing with complicated nonstationarity would be desirable.

However, flexibly modelling covariance structure appears to be a difficult task, much more difficult than modelling mean structure. Wikle et al. (1998) build hierarchical models in which the modelling focuses on the mean structure at each stage in the hierarchy in an attempt to avoid joint covariance modelling for a large number of locations. The difficulties in joint covariance modelling are apparent in the model comparison results here. The wavelet model that closely follows the empirical correlation drastically overfits, with very poor generalizability. The smoothed wavelet model does a poor job of predicting test data, both in terms of point predictions and covariance structure, presumably because it is fit without reference to the likelihood or rigorous fitting criteria. This poor predictive ability occurs even though the speed at which the correlations in the wavelet model decay appears to more closely mirror the empirical correlations than does the speed at which the correlations decay in the nonstationary model. While the kernel-based nonstationary model has its problems, as discussed above, the stationary model suffers in comparison. In the stationary model, many of the residual variance estimates are much larger than the ML estimates.

The variances seem to be accommodating the inability of the stationary model to fit the correlation structure. One difficulty that may play an important role in these results is that when modelling many locations simultaneously, the response at some locations can be nearly a linear combination of the response at other locations. When this constraint is part of a covariance structure and is violated, the fit can be very poor, even if the covariances look reasonable to the eye and on an element by element basis. It is possible that the quicker decay with distance in the correlations in the GP models relative to the wavelet models serves to minimize the inclusion of such linear combinations in the covariance matrix, at the cost of modelling other features of the covariance structure less closely. A possible alternative to the methods used here is to employ the wavelet approach with a specific criteria for fitting, such as making use of the likelihood and performing the thresholding in a Bayesian context. This might produce a flexible covariance model that does a better job of prediction and gets around the difficulty in choosing the thresholding in the current approach, although it would seemingly obviate many of the computational advantages of the method.

6.3 Future Work

In addition to the ideas mentioned in the previous section, there are some general areas of potential future work involving GP models that may be fruitful. My discussion here focuses first on generalizing the Gaussian process approach in two ways and then comments on including nonlinear time structure in the spatial model.

Previous approaches to modeling non-normal data using Gaussian processes have generally assumed independence of the observations conditional on a Gaussian process prior for a function determining location (Diggle et al. 1998). In many spatial datasets with replicated data, including the discrete-valued storm count index of Paciorek et al. (2002), residual correlation suggests this assumption of independence is violated. Here I suggest a generalized Gaussian process distribution for non-normal data with a single observation per location. To describe the distribution of the data, I show how to generate a single sample at each of n covariate values. Let $Z(\cdot)$ be a Gaussian process with covariance function $C(\cdot,\cdot)$. For each covariate, $\boldsymbol{x_i}$, let $p(\boldsymbol{x_i}) = \Phi(Z(\boldsymbol{x_i}))$, the standard normal CDF transformation of the Gaussian process at each covariate; $p(\cdot)$ might be called the quantile process. Next let $H_{\mu(\cdot)}$ be an appropriate parametric distribution function indexed by a

mean/location process, $\mu(\cdot)$, which could be a function of additional covariates or have a Gaussian process prior itself. For count data, one would likely choose H to be Poisson, while for binary data, it would be Bernoulli. Then the data are generated such that $Y_i = H^{-1}_{\mu(\boldsymbol{x_i})}(p(\boldsymbol{x_i}))$, i.e. the $p(\boldsymbol{x_i})$ quantile of $H_{\mu(\boldsymbol{x_i})}$. If the variance terms in the covariance function of $Z(\cdot)$ are equal to one, then the marginal distribution of $Z(\cdot)$ at each covariate is standard normal and therefore $p(\cdot)$ is $U(0,1)$, which means that $Y_i \sim H_{\mu(\boldsymbol{x_i})}$. If the variance at a covariate is less (greater) than one, then the distribution of the observation is under(over)-dispersed with median equal to that of $H_{\mu(\boldsymbol{x_i})}$. This approach mimics the inherent separation of mean and covariance in a Gaussian distribution by modelling the mean/location separately from the correlation structure.

The storm indices of Paciorek et al. (2002) are closely related to each other, measuring different aspects of the same phenomenon, yet I have modelled them independently, in part because I do not have a reasonable model for the joint distribution of the indices as a function of spatial location. To my knowledge little work has been done to define models for regression problems involving multiple responses. As one approach to defining a covariance structure for two responses, consider the kernel convolution of Higdon et al. (1999), but introduce a cross-covariance matrix process, $C'_{a,b}(\cdot)$, that relates the responses, a and b, at any location. Now if we have two separate kernel matrix processes, one for each response, we can do the convolution,

$$C_{a,b}(x,y) = \int_u K^a_x(u) K^b_y(u) C'_{a,b}(u) du,$$

to produce the covariance between the two responses at any pair of locations (x,y). The parameterizations for smoothly spatially-varying covariance matrices in Chapter 3 allow one to model the kernel matrix processes as before, as well as the newly-introduced cross-covariance matrix processes. Note that the covariance between the responses at the same location, $C_{a,b}(x,x)$ is not determined solely by the cross-covariance at x; this makes the underlying cross-covariance matrices, $C'_{a,b}(\cdot)$, somewhat difficult to interpret. In principle, this approach extends simply to more than two response variables. Other parameterizations may also be useful.

A final area for future research, which I discuss in more detail in Section 5.8.2, involves modelling nonlinear trends at multiple locations while accounting for spatial structure. The goal is to be able to understand the broad-scale time trends while allowing for nonlinearities that are likely to be present. Even if a reasonable temporal model can be constructed and fit, presenting the results

in an easily accessible manner will be a challenge because one is working with high-dimensional quantities in addition to the two-dimensional covariate space. Of course as the temporal modelling becomes more complicated and additional parameters are introduced, accounting for spatial structure also becomes more of a challenge, and I have found that doing this even in the linear trend case is difficult.