A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields

BY BRIAN J. REICH AND MONTSERRAT FUENTES

Department of Statistics, North Carolina State University,
2501 Founders Drive, Box 8203, Raleigh, NC 27695, U.S.A.

Correspondence Author: Brian J. Reich
E-mail: reich@stat.ncsu.edu
Telephone: (919) 513-7686
Fax: (919) 515-7591

February 19, 2007

---

1The research conducted by Reich has been supported by National Science Foundation grant DMS 0354189, and Fuentes has been partly supported by National Science Foundation grant DMS 0353029. The authors thank Professor Lian Xie of the Coastal Fluid Dynamics Laboratory at North Carolina State University and Dr. Kristen Foley of the US EPA for their helpful insight.
A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields

Abstract

Storm surge, the onshore rush of sea water caused by the high winds and low pressure associated with a hurricane, can compound the effects of inland flooding caused by rainfall, leading to loss of property and loss of life for residents of coastal areas. Numerical ocean models are essential for creating storm surge forecasts for coastal areas. These models are driven primarily by the surface wind forcings. Currently, the gridded wind fields used by ocean models are specified by deterministic formulas that are based on the central pressure and location of the storm center. While these equations incorporate important physical knowledge about the structure of hurricane surface wind fields they cannot always capture the asymmetric and dynamic nature of a hurricane. A new Bayesian multivariate spatial statistical modeling framework is introduced combining data with physical knowledge about the wind fields to improve the estimation of the wind vectors. Many spatial models assume the data follow a Gaussian distribution. However, this may be overly-restrictive for wind fields data which often display erratic behavior, such as sudden changes in time or space. In this paper, we develop a semiparametric multivariate spatial model for these data. Our model builds on the stick-breaking prior, which is frequently used in Bayesian modelling to capture uncertainty in the parametric form of an outcome. The stick-breaking prior is extended to the spatial setting by assigning each location a different, unknown distribution, and smoothing the distributions in space with a series of kernel functions. This semiparametric spatial model is shown to improve prediction compared to usual Bayesian Kriging methods for the wind field of Hurricane Ivan.
1 Introduction

Modelling surface wind fields is essential for hurricane forecasting. A wind field gives the wind velocity at any location in the vicinity of the hurricane. The numerical ocean models used to predict the storm surge for coastal areas rely heavily on wind field inputs. Currently, deterministic formulas such as the Holland model (Holland, 1980) are used to generate the wind fields for the storm surge model based on a few meteorological inputs such as the radius and central pressure of the storm.

While the Holland model captures many of the important features of a wind field, Foley and Fuentes (2006) show that this model does not allow for asymmetries often seen in wind fields and that storm surge prediction can be improved by supplementing the Holland model with a Gaussian geostatistical model. Another approach would be to introduce a more sophisticated deterministic wind model. A coupled atmospheric-oceanic numerical model can be used to simulate the surface winds at the boundary layer of the ocean model. However the CPU time required to produce these modeled winds at high enough resolution for coastal prediction (1 to 4 km grids) prevents such model runs from being used in real-time applications. Alternatively, one could write a stochastic version of the deterministic model and approximate the physical model using a stochastic spatial basis. This is the approach of Wikle et al. (2001) for oceanic surface winds.

This paper combines multiple sources of observed data and Holland model output using a semiparametric multivariate spatial model to predict a wind field. Several Gaussian multivariate spatial covariance models have been proposed. For example, Brown et al. (1994) model the joint covariance of the observed multivariate data using an inverse Wishart distribution centered on a separable covariance matrix. Another approach is to represent the
multivariate spatial process as a linear combination of univariate spatial process. Variations of this linear model of coregionization (LMC) have been used by Grzebyk and Wackernagel (1994), Wackernagel (2003), Schmidt and Gelfand (2003), Banerjee et al. (2004), and Gelfand et al. (2004). Foley and Fuentes (2006) apply the LMC to the two orthogonal west/east and north/south components of hurricane wind vectors.

Spatial models often assume the outcomes follow normal distributions. The Gaussian assumption is difficult to verify empirically and may be overly-restrictive for hurricane wind field data, which can display erratic behavior, such as sudden changes in time or space. For example, on the periphery of the map in Figure 1a the wind vectors vary smoothly from one measurement to the next. However, near the eye of the hurricane (center of the plot), the wind vectors are extremely volatile. Traditional Gaussian spatial models tend to oversmooth the area near the eye of the hurricane, resulting in a poor fit. Therefore, in this paper, we develop a new multivariate semiparametric spatial model for these data that avoids specifying a Gaussian distribution for the spatial random effects.

Our semiparametric model avoids assuming normality by extending the stick-breaking prior of Sethuraman (1994) to the multivariate spatial setting. For general (non-spatial) Bayesian modelling, the stick-breaking prior offers a way to model a distribution of a parameter as an unknown quantity to be estimated from the data. The stick-breaking prior for the unknown distribution \( F \) is the mixture

\[
F \overset{d}{=} \sum_{i=1}^{m} p_i \delta(\theta_i), \tag{1}
\]

where the number of mixture components \( m \) may be infinite, \( p_i \) are the mixture probabilities, and \( \delta(\theta_i) \) is the Dirac distribution with point mass at \( \theta_i \). The mixture probabilities “break
the stick’” into \( m \) pieces so the sum of the pieces is one, i.e., \( \sum_{i=1}^{m} p_i = 1 \). The first mixture probability is modelled as \( p_1 = V_1 \), where \( V_1 \sim \text{Beta}(a, b) \). Subsequent mixture probabilities are \( p_i = (1 - \sum_{j=1}^{i-1} p_j)V_i \), where \( 1 - \sum_{j=1}^{i-1} p_j \) is probability not accounted for by the first \( i - 1 \) mixture components, and \( V_i \sim \text{Beta}(a, b) \) is the proportion of the remaining probability assigned to the \( i^{th} \) component. The locations \( \theta_i \sim F_0 \), where \( F_0 \) is a known prior distribution. A special case of this prior is the Dirichlet process prior with \( m = \infty \) and \( a = 1 \) (Ferguson, 1973).

The stick-breaking prior in (1) has been extended to the univariate spatial setting by incorporating spatial information into either the model for the locations \( \theta_i \) or the model for the masses \( p_i \). Gelfand et al. (2005a) model the locations as vectors drawn from a spatial distribution. However, their model requires replication, and thus is not appropriate for analyzing the wind fields data. Griffin and Steel (2006) propose a spatial Dirichlet model that does not require replication. Their model permutes the \( V_i \) based on spatial location, allowing the prior to be different in different regions of the spatial domain.

This paper is the first to extend the stick-breaking prior to the multivariate spatial setting. Our semiparametric multivariate spatial model for a hurricane wind field has bivariate normal priors for the locations \( \theta_i \). Similar to Griffin and Steel, the probabilities \( p_i \) vary spatially. However, rather than a random permutation of \( V_i \), we introduce a series of kernel functions to allow the masses to change with space. This results in a flexible spatial model, as different kernel functions lead to different relationships between the distributions at nearby locations. Our model is also computationally convenient because it avoids reversible jump MCMC steps and inverting large matrices which is crucial for analysis of hurricane wind fields since estimates must be made in real-time.

The paper proceeds as follows. Section 2 describes the various sources of data used to
map the wind field. The semiparametric spatial prior for univariate spatial data is introduced in Section 3. This model is extended to a multivariate model to analyze wind field data in Section 4. The model incorporates both a deterministic wind model and multiple sources of wind observations and allows for potential bias for each data source. A simulation study in Section 5 shows that the semiparametric model outperforms the usual Gaussian Kriging model in terms of prediction in several situations. This model is used to map the wind field of Hurricane Ivan in Section 6. Section 7 concludes.

2 Description of the wind fields data

We model wind fields data from Hurricane Ivan as it passed through the Gulf of Mexico at 12pm on September 15, 2004. The three sources of information used in this analysis are plotted in Figure 1. The first source is gridded satellite data (Figure 1a) available from NASA’s SeaWinds database (http://podaac.jpl.nasa.gov/products/product109.html). These data are available twice daily on a 0.25 x 0.25 degree global grid. Due to the satellite data’s potential bias, measurement error, and course temporal resolution, we supplement our wind fields analysis with data from NOAA’s National Data Buoy Center. Buoy data are collected every ten minutes at a relatively small number of marine locations (Figure 1b). These measurements are adjusted to a common height of 10 meters above sea level using the algorithm of Large and Pond (1981).

In addition to satellite and buoy data, our model incorporates the deterministic Holland model (Holland, 1980). The NOAA currently uses this model alone to produce wind fields for their numerical ocean models. The Holland model predicts that the wind speed at location
\[ H(s) = \left( \frac{B}{\rho} \left( \frac{R_{max}}{r} \right)^B (P_n - P_c) \exp \left[ - \left( \frac{R_{max}}{r} \right)^B \right] \right)^{1/2}, \tag{2} \]

where \( r \) is the radius (km) from the storm center to site \( s \), \( P_n \) is the ambient pressure (mb), \( P_c \) is the hurricane central pressure (mb), \( \rho \) is the air density (kg m\(^{-3}\)), \( R_{max} \) is radius of the maximum wind (km), and \( B \) controls the shape of the pressure profile.

Section 4’s multivariate spatial model decomposes the wind vectors into their orthogonal west/east (\( u \)) and north/south (\( v \)) vectors. The Holland model for the \( u \) and \( v \) components is

\[ H_u(s) = H(s) \sin(\phi) \quad \text{and} \quad H_v(s) = H(s) \cos(\phi), \tag{3} \]

where \( \phi \) is the inflow angle at site \( s \), across circular isobars toward the storm center, rotated to adjust for the storm’s direction. We fix the parameters \( P_n = 1010 \text{mb}, P_c = 939 \text{mb}, \rho = 1.2 \text{ kg m}^{-3}, \) and \( R_{max} = 49 \), and \( B = 1.9 \) using the meteorological data from the national hurricane center (http://www.nhc.noaa.gov) and recommendations of Hsu and Yan (1998). The output from this model for Hurricane Ivan is plotted in Figure 1c. By construction, Holland model output is symmetric with respect to the storm’s center, which does not agree with the satellite observations in Figure 1a.

### 3 The spatial stick-breaking (SSB) prior

In this section, we develop a univariate semiparametric spatial model for data from a single source. The spatial stick-breaking prior developed here is incorporated into our model for the wind fields data in Section 4. Let \( y(s) \), the observed value at site \( s = (s_1, s_2) \), have the
model

\[ y(s) = \mu(s) + x(s)'\beta + \varepsilon(s), \] (4)

where \( \mu(s) \) is a spatial random effect, \( x(s) \) is a vector of covariates for site \( s \), \( \beta \) are the regression parameters, and \( \varepsilon(s) \overset{iid}{\sim} N(0, \sigma^2) \).

The spatial effects are each assigned a different prior distribution, i.e., \( \mu(s) \sim F(s) \). The distributions \( F(s) \) are unknown and smoothed spatially. Extending (1) to depend on \( s \), the prior for \( F(s) \) is the potentially infinite mixture

\[ F(s) = \sum_{i=1}^{m} p_i(s) \delta(\theta_i), \] (5)

where \( p_1(s) = V_1(s) \), \( p_i(s) = V_i(s) \prod_{j=1}^{i-1} (1 - V_j(s)) \) for \( i > 1 \), and \( V_i(s) = w_i(s)V_i \). The distributions \( F(s) \) are related through their dependence on the \( V_i \) and \( \theta_i \), which are given the priors \( V_i \sim \text{Beta}(a, b) \) and \( \theta_i \sim N(0, \tau^2) \), each independent across \( i \). However, the distributions vary spatially according to the kernel functions \( w_i(s) \), which are restricted to the interval \([0,1]\). The function \( w_i(s) \) is centered at knot \( \psi_i = (\psi_{1i}, \psi_{2i}) \) and the spread is controlled by the bandwidth parameter \( \epsilon_i = (\epsilon_{1i}, \epsilon_{2i}) \). Both the knots and the bandwidths are modelled as unknown parameters with priors that are independent of the \( V_i \) and \( \theta_i \). The knots \( \psi_i \) are given independent uniform priors over the bounded spatial domain (this is generalized in Section 4). The bandwidths can be modelled as equal for each kernel function or varying across kernel functions following prior distributions.

Although there are many possible kernel functions, Table 1 gives two examples. Uniform kernels offer bounded support. This is an attractive feature when modelling hurricane wind fields because wind behavior may be different in different subregions, e.g., in the center of the storm versus the periphery. We compare uniform kernels with squared-exponential
Table 1: Examples of kernel functions and the induced functions $\gamma(s, s')$, where $h_1 = |s_1 - s'_1| + |s_2 - s'_2|$, $h_2 = \sqrt{(s_1 - s'_1)^2 + (s_2 - s'_2)^2}$, $I(\cdot)$ is the indicator function, and $x^+ = \max(x, 0)$.

<table>
<thead>
<tr>
<th>Name</th>
<th>$w_i(s)$</th>
<th>Model for $\epsilon_{1i}$ and $\epsilon_{2i}$</th>
<th>$\gamma(s, s')$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform</td>
<td>$\prod_{j=1}^2 I(</td>
<td>s_j - \psi_{ji}</td>
<td>&lt; \frac{\epsilon_{ji}}{2})$</td>
</tr>
</tbody>
</table><p>ight)^+$                 |
| Uniform        | $\prod_{j=1}^2 I(|s_j - \psi_{ji}| \leq \frac{\epsilon_{ji}}{2})$      | $\epsilon_{1i}, \epsilon_{2i} \sim \text{Expo}(\lambda)$ | $\exp(-h_1/\lambda)$                                                         |
| Squared exp.   | $\prod_{j=1}^2 \exp\left(-\frac{(s_j - \psi_{ji})^2}{\epsilon_{ji}^2}\right)$ | $\epsilon_{1i}, \epsilon_{2i} \equiv \lambda^2/2$ | $0.5 \exp\left(-\frac{h_2^2}{\lambda^2}\right)$                               |
| Squared exp.   | $\prod_{j=1}^2 \exp\left(-\frac{(s_j - \psi_{ji})^2}{\epsilon_{ji}^2}\right)$ | $\epsilon_{1i}, \epsilon_{2i} \sim \text{IG}(1.5, \lambda^2/2)$ | $0.5/\left(1 + \left(\frac{h_2}{\lambda}\right)^2\right)$                  |</p>

kernels. Squared-exponential kernels decay slowly in space which may be desirable in other applications.

To ensure that the stick-breaking prior is proper, we must choose priors for $\epsilon_i$ and $V_i$ so that $\sum_{i=1}^m p_i(s) = 1$ almost surely for all $s$. Appendix A.1 shows that the SSB prior with infinite $m$ is proper if $E(V_i) = a/(a + b)$ and $E[w_i(s)]$ (where the expectation is taken over $(\psi_i, \epsilon_i)$) are both positive. For finite $m$, we can ensure that $\sum_{i=1}^m p_i(s) = 1$ for all $s$ by setting $V_m(s) \equiv 1$ for all $s$. This is equivalent to truncating the infinite mixture by attributing all of the mass from the terms with $i \geq m$ to $p_m(s)$.

In practice, allowing $m$ to be infinite is often unnecessary and computationally infeasible. Choosing the number of components in a mixture model is notoriously problematic. Fortunately, in this setting the truncation error can easily be accessed by inspecting the distribution of $p_m(s)$, the mass of the final component of the mixture. The number of components $m$ can be chosen by generating samples from the prior distribution of $p_m(s)$. We increase $m$ until $p_m(s)$ is satisfactorily small for each site $s$. Also, the truncation error is monitored by inspecting the posterior distribution of $p_m(s)$, which is readily available from the MCMC samples.

Assuming finite $m$, the spatial stick-breaking model can be written as a mixture model.
where \( g(s) \in \{1, \ldots, m\} \) indicates site \( s \)'s group, i.e.,

\[
y(s) = \theta_{g(s)} + x(s)\beta + \varepsilon(s), \quad \text{where } \varepsilon(s) \overset{iid}{\sim} N(0, \sigma^2)
\]

\[
\theta_j \overset{iid}{\sim} N(0, \tau^2), \quad j = 1, \ldots, m
\]

\[
g(s) \sim \text{Categorical}(p_1(s), \ldots, p_m(s))
\]

\[
p_j(s) = w_j(s) V_j \prod_{k<j} [1 - w_k(s) V_k],
\]

where \( \mu(s) = \theta_{g(s)}, \ V_j \overset{iid}{\sim} \text{Beta}(a,b), \) and \( \prod_{k<j} [1 - w_k(s) V_k] = 1 \) for \( j = 1 \). To complete the Bayesian model, we specify priors for the hyperparameters. The regression parameters \( \beta \) can be given vague normal priors. In the analysis of Hurricane Ivan in Section 6, the mean term \( x(s)'\beta \) is replaced by Holland model output. The parameters that control the beta prior for the \( V_i \), \( a \) and \( b \), have independent Uniform(0,10) priors, and the variances \( \sigma^2 \) and \( \tau^2 \) have InvGamma(0.01,0.01) priors. We also tried InvGamma(0.5,0.005) priors for the variances and found that the prior had little effect. The knots that control the center of the kernel functions, \( \psi_j \), are given uniform priors over the spatial domain and examples of priors for bandwidth parameters \( \epsilon_j \) are given in Table 1. The prior for the bandwidths depend on a range parameter, \( \lambda \), which is given a Uniform(0,\( \lambda_{max} \)) prior. We take \( \lambda_{max} \) to be the maximum distance between any pair of points in the spatial grid.

This model can be implemented using WinBUGS. WinBUGS can be freely downloaded from \texttt{http://www.mrc-bsu.cam.ac.uk/bugs/}. Code to implement this model is given in Appendix A.2.

The mixture model in (6) is nowhere continuous unless uniform kernels are selected and
\[ V_i \in \{0, 1\} \text{ for all } i. \] An alternative suggested by a referee is

\[ g(s) = j \text{ where } p_j(s) = \max\{p_1(s), ..., p_m(s)\}. \] (7)

This would result in a piece-wise constant random tessellation model which may be preferred for smooth spatial data. However, to avoid oversmoothing micro-scale phenomena in hurricanes, we use the everywhere discontinuous model in (6).

Figure 2a illustrates the spatially varying weights of the stick-breaking prior for a one-dimensional example with \( m = 6 \) and squared exponential kernel functions. We arbitrarily select knots \( \psi = (0.5, 0.0, 1.0, 0.2, 0.8) \), bandwidths \( \epsilon = (0.1, 0.2, 0.2, 0.2, 0.2) \), and \( V = (0.9, 0.7, 0.7, 0.9, 0.9) \). The first kernel function is centered at \( s = 0.5 \). Since \( V_1 = 0.9 \), the mass for the first component for \( s = 0.5 \) is \( p_1(0.5) = 0.9 \) and decreases as \( s \) moves away from 0.5. The second and third kernel functions are centered at \( s = 0.0 \) and \( s = 1.0 \) respectively, and dominate the probabilities near the edges. For this example, \( p_m(s) \) is as large as 0.2, suggesting \( m \) should be increased to give an acceptable approximation to the infinite spatial stick-breaking prior.

Understanding the spatial correlation function is crucial for analyzing spatial data. Although the spatial stick-breaking prior forgoes the Gaussian assumption for the spatial random effects, we can still compute and investigate the covariance function. Conditional on the probabilities \( p_j(s) \) (but not the locations \( \theta_j \)), the covariance between two observations is

\[ \text{cov}(y(s), y(s')) = \tau^2 \mathbb{P}(\mu(s) = \mu(s')) = \tau^2 \sum_{j=1}^{m} p_j(s)p_j(s'). \] (8)

Figure 2b maps the correlation function induced by the probabilities in Figure 2a. For these
probabilities, the correlation is not simply a function of distance between points, i.e., the correlation is non-stationary. For example, the correlation is near one for all sites in (0.4,0.6) due to the large probability for the first component throughout the region. In contrast, the correlation between nearby sites is smaller near $s = 0.35$ and $s = 0.65$ where several components have substantial probability.

As shown in the Appendix A.3, integrating over $(V_i, \psi_i, \epsilon_i)$ and letting $m \to \infty$ gives

$$\text{Var}(y(s)) = \sigma^2 + \tau^2$$  \hspace{1cm} (9)

$$\text{Cov}(y(s), y(s')) = \tau^2 \gamma(s, s') \left[ \frac{2(a+b+1)}{a+1} - \gamma(s, s') \right]^{-1},$$  \hspace{1cm} (10)

where

$$\gamma(s, s') = \frac{\int \int w_i(s)w_i(s')p(\psi_i, \epsilon_i)d\psi_i d\epsilon_i}{\int \int w_i(s)p(\psi_i, \epsilon_i)d\psi_i d\epsilon_i} \in [0, 1].$$  \hspace{1cm} (11)

Since $(V_i, \psi_i, \epsilon_i)$ have independent priors that are uniform over the spatial domain, integrating over these parameters gives a stationary prior covariance. However, Figure 2 illustrates that the conditional covariance can be non-stationary. Therefore, we conjecture the spatial stick-breaking model is more robust to nonstationarity than traditional stationary Kriging methods. This is demonstrated by the simulation study in Section 5.

If $b/(a+1)$ is large, i.e., the $V_i$ are generally small and there are many terms in the mixture with significant mass, the correlation between $y(s)$ and $y(s')$ is approximately proportional to $\gamma(s, s')$. Table 1 gives the function $\gamma(s, s')$ for several examples of kernel functions and shows that different kernels can produce very different correlation functions. For example, $\gamma(s, s')$ under the uniform kernel with exponential priors for the bandwidth parameters is the familiar exponential correlation function. If the bandwidths are shared across kernel functions, $\gamma(s, s')$ is proportional to a squared exponential covariance under squared exponential kernel
functions. The uniform kernel with common bandwidth parameter $\lambda$ gives compact support, as observations separated by more than $\lambda$ spatial units are uncorrelated.

4  A multivariate spatial model for wind fields data

Let $u(s)$ and $v(s)$ be the underlying wind speed in the west/east and north/south directions, respectively, for spatial location $s$. As described in Section 2, there are two types of observed wind data: $u_T(s)$ and $v_T(s)$ are satellite measurements and $u_B(s)$ and $v_B(s)$ are buoy measurements. Our model for these data is

$$
\begin{align*}
    u_T(s) &= a_u + u(s) + e_{uT}(s) & v_T(s) &= a_v + v(s) + e_{vT}(s) \\
    u_B(s) &= u(s) + e_{uB}(s) & v_B(s) &= v(s) + e_{vB}(s),
\end{align*}
$$

(12)

where $\{e_{uT}, e_{vT}, e_{uB}, e_{vB}\}$ are independent (with each other and with the underlying winds), zero mean, Gaussian errors, each with its own variance, and $\{a_u, a_v\}$ account for additive bias in the satellite and aircraft data. Of course, the buoy data may also have bias, but it is impossible to identify bias from both sources, so we attribute all the bias to the satellite measurements. It is also possible to add multiplicative bias terms, but with the small number of buoy observations it will be difficult to identify both types of bias and Foley and Fuentes (2006) found that the primary source of bias is additive.

The underlying orthogonal wind components $u(s)$ and $v(s)$ are modelled as a mixture of a deterministic wind model and a semiparametric multivariate spatial process

$$
\begin{align*}
    u(s) &= H_u(s) + R_u(s) \\
    v(s) &= H_v(s) + R_v(s),
\end{align*}
$$

(13)
where $H_u(s)$ and $H_v(s)$ are the orthogonal components of the deterministic Holland model in (3) and $R(s) = (R_u(s), R_v(s))^\prime$ follows a multivariate extension of the non-Gaussian spatial stick-breaking prior of Section 3. We take $R(s) \sim F(s)$, where $F$ has the stick-breaking prior in (5) modified so the two-dimensional locations $\theta_i$ have multivariate normal priors $\theta_i \overset{iid}{\sim} N(0, \Sigma)$, where $\Sigma$ is a $2 \times 2$ covariance matrix that controls the association between the two wind components. The covariance $\Sigma$ has an $\text{InvWish}(0.1, 0.1 I_2)$ prior and after transforming the spatial grid to be contained in the unit square, the spatial knots $\psi_{s1i}$ and $\psi_{s2i}$ have independent $\text{Beta}(1.5, 1.5)$ priors to encourage knots to lie near the center of the hurricane where the wind is most volatile. Also, we take the spatial range $\lambda \sim \text{Uniform}(0,1)$.

Assuming the same priors for the $p_i(s)$ as in Section 3 and following the same steps as in Appendix A.3, it can be shown that $\text{Cov}(R(s), R(s'))$ is separable, i.e., the product of the a spatial covariate and the cross-dependency covariance matrix $\Sigma$. This could be generalized by allowing the prior covariance of the $\theta_i$ to vary spatially. Alternatively, the spatial stick-breaking prior could be combined with the linear model of coregionalization to give a nonseparable multivariate spatial model by modelling the $u$ and $v$ components of $R(s)$ as linear combinations of univariate spatial terms given spatial stick-breaking priors described in Section 3.

5 Simulation Study

In this section we present the results of a brief simulation study comparing the spatial stick-breaking model of Section 3 with the usual Bayesian Kriging model for univariate geostatistical data. We consider five true mean surfaces. The first is the flat surface with $\mu(s) \equiv 0$ for all $s$. The four non-constant true surfaces are plotted in Figure 3. For each
surface, we randomly generate ten data sets on a 10×10 square grid by adding independent standard normal errors to the true mean surface.

For the spatial stick-breaking prior, we use both the uniform kernel function with exponential priors for the bandwidths and the squared exponential kernel functions with inverse gamma priors for the bandwidths in Table 1. In both cases, the knot components are given independent Uniform(0,1) priors and the range parameter \( \lambda \) has a Uniform(0,1) prior. Also, we assume \( m = 50, a \sim \text{Uniform}(0,10), \) and \( b \sim \text{Uniform}(0,10). \) Using these values, the prior median of the truncation probabilities \( p_m(s) \) is less than 0.001 for all \( s \) for both the uniform and squared exponential kernel functions.

The stick-breaking model is compared with the usual Bayesian Kriging model (Handcock and Stein, 1993) with exponential covariance function, i.e., \( y(s) = \beta_0 + \mu(s) + \epsilon(s), \) where \( \epsilon(s) \overset{iid}{\sim} N(0, \sigma^2) \) and the vector of spatial random effects \( \mu = (\mu(s_1), ..., \mu(s_n)) \) has a multivariate normal prior with mean zero and \( \text{cov}(\mu(s), \mu(s')) = \tau^2 \exp(-||s - s'||/\lambda). \) The intercept \( \beta_0 \) is given a flat prior, the variance parameters \( \sigma^2 \) and \( \tau^2 \) are given independent Gamma(0.01,0.01) priors, and the range parameter \( \lambda \) is given a Uniform(0,1) prior.

We compare the models in terms of root mean squared estimation error, \( RMSE = \sqrt{\sum_{i=1}^{n} (\hat{\mu}(s_i) - \hat{\mu}(s_i))^2} \), where \( \hat{\mu}(s_i) \) is the true mean at site \( s_i \) plotted in Figure 3 and \( \hat{\mu}(s_i) \) is the posterior mean of \( \mu(s_i) \). Also, to measure predictive performance we generate unobserved data \( z(l_1), ..., z(l_{np}) \) on a 20×20 square grid with locations labelled \( l_1, ..., l_{np} \) from the same spatial domain and same model as the original data. The predictive models are evaluated using root mean squared prediction error \( RMSGPE = \sqrt{\sum_{i=1}^{np} (z(l_i) - \hat{z}(l_i))^2} \), where \( \hat{z}(l_i) \) is the posterior mean of the predictive distribution of \( z(l_i) \), and the percentage of sites for which the predictive 95% intervals cover the unobserved value \( z(l_i) \).

Table 2 shows the results of the simulation study. The semiparametric models perform
Table 2: The results of the simulation study. The values in the table are the mean (standard error) across the 10 simulated data sets.

(a) Root mean squared error (RMSE) at observed locations

<table>
<thead>
<tr>
<th></th>
<th>SSB uniform</th>
<th>SSB squared exponential</th>
<th>Gaussian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat</td>
<td>0.12 (0.02)</td>
<td>0.14 (0.01)</td>
<td>0.26 (0.02)</td>
</tr>
<tr>
<td>SW/NE trend</td>
<td>0.40 (0.02)</td>
<td>0.43 (0.02)</td>
<td>0.35 (0.03)</td>
</tr>
<tr>
<td>Waves</td>
<td>0.66 (0.04)</td>
<td>0.65 (0.02)</td>
<td>0.60 (0.03)</td>
</tr>
<tr>
<td>Steps</td>
<td>0.45 (0.03)</td>
<td>0.60 (0.03)</td>
<td>0.71 (0.03)</td>
</tr>
<tr>
<td>Wind field</td>
<td>0.66 (0.02)</td>
<td>0.67 (0.02)</td>
<td>0.84 (0.01)</td>
</tr>
</tbody>
</table>

(b) Root mean squared prediction error (RMSPE) at unobserved locations

<table>
<thead>
<tr>
<th></th>
<th>SSB uniform</th>
<th>SSB squared exponential</th>
<th>Gaussian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat</td>
<td>1.01 (0.01)</td>
<td>1.01 (0.01)</td>
<td>1.03 (0.01)</td>
</tr>
<tr>
<td>SW/NE trend</td>
<td>1.05 (0.02)</td>
<td>1.04 (0.02)</td>
<td>1.08 (0.02)</td>
</tr>
<tr>
<td>Waves</td>
<td>1.28 (0.04)</td>
<td>1.35 (0.03)</td>
<td>1.19 (0.04)</td>
</tr>
<tr>
<td>Steps</td>
<td>1.36 (0.02)</td>
<td>1.37 (0.02)</td>
<td>1.58 (0.10)</td>
</tr>
<tr>
<td>Wind field</td>
<td>1.32 (0.02)</td>
<td>1.65 (0.02)</td>
<td>1.41 (0.08)</td>
</tr>
</tbody>
</table>

(c) Coverage frequencies for 95% prediction intervals at unobserved locations

<table>
<thead>
<tr>
<th></th>
<th>SSB uniform</th>
<th>SSB squared exponential</th>
<th>Gaussian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat</td>
<td>0.943 (0.007)</td>
<td>0.943 (0.006)</td>
<td>0.939 (0.007)</td>
</tr>
<tr>
<td>SW/NE trend</td>
<td>0.950 (0.006)</td>
<td>0.948 (0.005)</td>
<td>0.945 (0.007)</td>
</tr>
<tr>
<td>Waves</td>
<td>0.942 (0.007)</td>
<td>0.933 (0.007)</td>
<td>0.917 (0.011)</td>
</tr>
<tr>
<td>Steps</td>
<td>0.964 (0.007)</td>
<td>0.971 (0.006)</td>
<td>0.897 (0.013)</td>
</tr>
<tr>
<td>Wind field</td>
<td>0.971 (0.004)</td>
<td>0.967 (0.004)</td>
<td>0.907 (0.007)</td>
</tr>
</tbody>
</table>
better than the fully-Gaussian model for the simulation with a flat true mean surface. The RMSE is the smallest for the spatial stick-breaking model with uniform kernel function. Inspecting the simulation output shows that the uniform kernel model performs well in this situation because there was often one component of the mixture with \( w_i(s) = 1 \) for most of the spatial domain that dominates the mixture probabilities and has \( \theta_i \) near zero.

The fully-parametric geostatistical model has the smallest RMSE for the “SW/NE trend” and “waves” surfaces plotted in Figures 3a and 3b, respectively. These surfaces vary gradually in space so it is not surprising that the parametric model that assumes smoothness outperforms the semiparametric models. However, the semiparametric models perform fairly well in these settings. The root mean squared prediction error (Table 2b) for the “SW/NE trend” data sets is actually smaller for the semiparametric models and the stick-breaking model with uniform kernel has the smallest RMSE for 3 of the 10 simulated “waves” data sets. Also, the SSB models have slightly higher predictive coverage probabilities for the “waves” data sets than the parametric model.

The “steps” and “wind fields” surfaces in Figures 3c and 3d have either discontinuities or areas with increased volatility, causing the Bayesian Kriging model to perform poorly. For example, the coverage probabilities for the 95% predictive intervals for the Kriging model are only about 90% for these surfaces. In contrast, the coverage frequencies are over 95% for the semiparametric models for both the “step” and “wind fields” surfaces. Also, there is considerable improvement over the parametric model in terms of mean squared error at observed and unobserved locations.

In summary, the parametric model was superior to the spatial stick-breaking model for smooth true surfaces. However, the semiparametric models were fairly competitive for the smooth surfaces and outperformed the parametric model for true surfaces with discontinuities.
and highly volatile regions. Also, for these surfaces the stick-breaking model with uniform kernel functions performed better than the model with squared exponential kernel functions.

6 Analysis of Hurricane Ivan’s wind field

In this section, we fit the multivariate extensions of the three models used in Section 5’s simulation study to 182 satellite observations and 7 buoy observations for the Hurricane Ivan. We use the multivariate SSB model in Section 4 with both uniform and squared-exponential kernels. Also, to illustrate the effect of relaxing the normality assumption, we also fit a fully-Gaussian Bayesian Kriging model that replaces the stick-breaking prior for $\mathbf{R}(s) = (R_u(s), R_v(s))'$ in (13) with a zero-mean Gaussian prior with separable covariance

$$
\text{Var}(\mathbf{R}(s)) = \Sigma \quad \text{and} \quad \text{Cov}(\mathbf{R}(s), \mathbf{R}(s')) = \Sigma \times \exp\left(-||s - s'||/\lambda\right),
$$

where $\Sigma$ controls the dependency between the wind components at a given location and $\lambda$ is a spatial range parameter. The covariance parameters $\Sigma$ and $\lambda$ have the same priors as the covariance parameters in Section 4.

Since our primary objective is to predict wind vectors at unmeasured locations to use as inputs for numerical ocean models, we compare models in terms of expected mean squared prediction error (Laud and Ibrahim, 1995; Gelfand and Ghosh, 1998), i.e.,

$$
EMSP\ E = E\left(\sum_s (u_T(s) - \tilde{u}_T(s))^2 + (v_T(s) - \tilde{v}_T(s))^2 + \sum_s (u_B(s) - \tilde{u}_B(s))^2 + (v_B(s) - \tilde{v}_B(s))^2\right),
$$

where, say, $\tilde{u}_T(s)$ is viewed as a replicate of the observed $u$ component of the satellite measurement at site $s$, the summation is taken over all observation locations, and the expectation is
taken over the full posterior of all the parameters in the model. This model selection criteria favors predictive models centered near the observed data with small predictive variances.

The EMSPE is smaller for the semiparametric model uniform kernels (EMSPE=3.46) than for the semiparametric model squared exponential kernels (EMSPE=4.19) and the fully-Gaussian model (EMSPE = 5.17). Figures 4a and 4b show that the squared residuals from the fully-Gaussian fit are near zero for most of the spatial domain but are large near the center of the hurricane for both components. The Gaussian model oversmooths in this area with high volatility in the underlying wind surface. In contrast, the semiparametric model with uniform kernel functions is able to capture the peaks near the eye of the hurricane and the squared residuals (Figures 4c and 4d) show less spatial structure than the residuals from the Gaussian model.

Figure 5 summarizes the posterior from the spatial stick-breaking prior with uniform kernel functions. The fitted values in Figures 5a and 5b vary rapidly near the center of the storm and are fairly smooth in the periphery. After accounting for the Holland model, the correlation between the residual $u$ and $v$ components $R_u(s)$ and $R_v(s)$ ($\Sigma_{12}/\sqrt{\Sigma_{11} \Sigma_{22}}$, where $\Sigma_{kl}$ is the $(k,l)$ element of $\Sigma$) is generally negative (Figure 5c), confirming the need for a multivariate analysis. Figure 5d plots the posterior of the parameter that controls the size of bandwidths, $\lambda$. The posterior median of $\lambda$ is 0.17, so on average the uniform kernels span about 17% of the spatial domain.

The satellite data are significantly biased relative to the buoy data. The 95% posterior intervals for the bias terms $a_u$ and $a_v$ are (-6.91, -2.16) and (0.04, 4.38) respectively. The biases seem to be driven by the third buoy’s wind vector in Figure 1b, which is quite different from the nearby satellite observations in Figure 1a.

To show that the semiparametric model with uniform kernel functions fits the data well,
we randomly (across $u$ and $v$ components and buoy and satellite data) set aside 10% of the observations and compute 95% predictive intervals for the missing observations. The prediction intervals contain 94.7% (18/19) of the deleted $u$ components and 95.2% (20/21) of the deleted $v$ components. These statistics suggest that our model is well-calibrated.

7 Discussion

Modeling hurricane wind fields is an important and challenging problem. This paper presents a semiparametric multivariate spatial model for these data. Gaussian models with highly-structured covariance functions, e.g., Wikle et al. (2001) and Fuentes et al. (2005), are an alternative. However, our non-parametric model offers greater flexibility by allowing for non-stationarity and non-normality which is advantageous when building an automated procedure. The semiparametric model has smaller mean square error than the stationary Gaussian spatial model for simulated data with discontinuities and areas with high volatility. The semiparametric model also avoids oversmoothing near the center of Hurricane Ivan’s wind field, resulting in a superior predictive model.

In the statistical model for wind fields data, the spatial random effects with the spatial stick-breaking prior are mixed with independent normal errors. An extension of this model would be to replace the independent normal effects with a Gaussian spatial process. This would give a mixture of two spatial terms: a semiparametric term to handle discontinuities and a Gaussian process which performs well in smooth areas. A mixture of this nature has been considered by Lawson and Clark (2002), who propose a fully-parametric mixture of spatial models for disease mapping with areal spatial data. As Lawson and Clark point out, it can be difficult to identify the contribution of each component of the mixture, but using
a combination of spatial terms can lead to an improvement in fit.

This paper focused on estimating the wind field at a single time point because satellite data are only available twice daily. However, the spatial stick-breaking prior developed here could be extended to the spatiotemporal setting to improve real-time estimates. One possibility is to use three-dimensional kernel functions in space and time. An alternative spatiotemporal model would be an extension of the dynamic linear model of Gelfand et al. (2005b), i.e., \( R(s, t) = BR(s, t - 1) + \Delta(s, t) \), where \( R(s, t) \) is the vector of residual wind components at location \( s \) at time \( t \), \( B \) is diagonal with \( B_{ii} \in [-1, 1] \), and \( \Delta(s, t) \) is the vector of changes from time \( t - 1 \) to time \( t \). The spatial stick-breaking prior could be applied to the mean at the first time point, \( R(s, 1) \), and each \( \Delta(s, t) \).

References


Appendix A.1 – Propriety of the SSB prior

For infinite \( m \), Ishwaran and James (2001) show that \( \sum_{i=1}^{m} p_i(s) = 1 \) almost surely if and only if \( \sum_{i=1}^{\infty} E(\log(1 - V_i(s))) = -\infty \). Applying Jensen’s inequality,

\[
E[\log(1 - V_i(s))] \leq \log[E(1 - V_i(s))] = \log[1 - E\{w_i(s)\}E(V_i)].
\]

If both \( E\{w_i(s)\} \) and \( E(V_i) \) are positive, \( \log(1 - E\{w_i(s)\}E(V_i)) \) is negative and

\[
\sum_{i=1}^{\infty} E[\log(1 - V_i(s))] \leq \sum_{i=1}^{\infty} \log(1 - E\{w_i(s)\}E(V_i)) = -\infty.
\]

Appendix A.2 – WinBUGS code

Below is the WinBUGS code for the univariate spatial stick-breaking prior with finite \( m \) described in (6). The inputs to this program include the outcome vector \( y \) and spatial coordinates \( s_1 \) and \( s_2 \).

```plaintext
model{
  for(i in 1:n){
    y[i] ~ dnorm(mu[i],taue)
    mu[i] <- int+theta[g[i]]
    g[i] ~ dcat(p[i,])
  }
  int ~ dnorm(0,0.01)
  taue ~ dgamma(0.01,0.01)
  for(j in 1:m){theta[j] ~ dnorm(0,taus)}
  taus ~ dgamma(0.01,0.01)
  for(i in 1:n){
    p[i,1]<-uin[i,1]
    for(j in 2:m)p[i,j]<-uin[i,j]*prod(uout[i,1:(j-1)])
  }
  for(j in 1:(m-1)){
    v[j] ~ dbeta(a,b)
    knot1[j] ~ dunif(min1,max1)
    knot2[j] ~ dunif(min2,max2)
    invbw[j] ~ dgamma(1.5,lambda)
  }
}
```
for(i in 1:n){
    uin[i,j]<-exp(-.5*invbw[j]*(pow(knot1[j]-s1[i],2) + pow(knot2[j]-s2[i],2)))*v[j]
    uout[i,j]<-1-uin[i,j]
}
}

for(i in 1:n){
    uin[i,m]<-1
}

lambda ~ dunif(0,lmax)
a ~ dunif(amin,amax)
b ~ dunif(bmin,bmax)

Appendix A.3 – Cov(µ(s), µ(s'))

Due to the discrete nature of the stick-breaking prior, Cov(µ(s), µ(s')) = \tau^2 \text{Prob(µ(s) = µ(s'))}.

\text{Prob(µ(s) = µ(s')|V, \psi, \epsilon)} = \sum_{i=1}^{\infty} p_i(s)p_i(s')

= \sum_{i=1}^{\infty} \left[w_i(s)w_i(s')V_i^2 \prod_{j<i} \left(1 - (w_j(s) + w_j(s'))V_j + w_j(s)w_j(s')V_j^2\right)\right].

Integrating over the (V, \psi, \epsilon) gives

\text{Prob(µ(s) = µ(bs')) = c_2\tilde{v}_2 \sum_{i=1}^{\infty} [1 - 2c_1\tilde{v}_1 + c_2\tilde{v}_2]^{i-1}}

where c_1 = \int \int w_i(s)p(\psi_i, \epsilon_i)d\psi_i d\epsilon_i, c_2 = \int \int w_i(s)w_i(s')p(\psi_i, \epsilon_i)d\psi_i d\epsilon_i, \tilde{v}_1 = E(V_1) = a/(a + b), and \tilde{v}_2 = E(V_1^2) = a(a + 1)/[(a + b)(a + b + 1)]. Since 1 - 2c_1\tilde{v}_1 + c_2\tilde{v}_2 = E [(1 - p_i(s))(1 - p_i(s'))] \in [0, 1] we apply the formula for the sum of a geometric series and simplify, leaving

\text{Prob(µ(s) = µ(s')) = \frac{c_2\tilde{v}_2}{2c_1\tilde{v}_1 - c_2\tilde{v}_2} = \frac{\gamma(s,s')}{2 \left(1 + \frac{b}{a+1}\right) - \gamma(s,s')}},

where \gamma(s,s') = c_2/c_1.
Figure 1: Plot of various types of wind field data/output for Hurricane Ivan on September 15, 2004.

(a) Satellite data

(b) Buoy data

(c) Holland model output

(d) Satellite - Holland model output
Figure 2: Example to illustrate the spatial stick-breaking prior. In this example, the spatial domain is the one-dimensional interval (0,1) and the model has Gaussian kernels with knots $\psi = (0.5, 0.0, 1.0, 0.2, 0.8)$, bandwidths $\epsilon = (0.1, 0.2, 0.2, 0.2, 0.2)$, and $V = (0.9, 0.7, 0.7, 0.9, 0.9)$. Panel (a) shows the masses $p_i(s)$ and Panel (b) shows the correlation between $\mu(s)$ and $\mu(s')$. 

(a) Probabilities $p_i(s)$

(b) Cor($\mu(s), \mu(s')$)
Figure 3: True mean surfaces for the simulation study.

(a) SW/NE trend

(b) Waves

(c) Jumps

(d) Wind fields
Figure 4: Squared residuals (value-posterior mean) for the $u$ and $v$ components of the Gaussian and spatial stick-breaking model with uniform kernels. The “⋆” represents the storm’s center.

(a) Gaussian model, $u$

(b) Gaussian model, $v$

(c) Spatial stick-breaking model, $u$

(d) Spatial stick-breaking model, $v$
Figure 5: Summary of the posterior of the spatial stick-breaking model with uniform kernels. Panels (a) and (b) give the posterior mean surface for the $u$ and $v$ components, Panel (c) shows the posterior of the cross-correlation between the residual wind components $R_u(s)$ and $R_v(s)$ ($\Sigma_{12}/\sqrt{\Sigma_{11}\Sigma_{22}}$), and Panel (d) plots the posterior of the parameter that controls the average kernel bandwidth $\lambda$ assuming the spatial grid has been transformed to lie in the unit square. The “⋆” represents the storm’s center.

(a) Posterior mean of $u(s)$

(b) Posterior mean of $v(s)$

(c) Cross-correlation

(d) Average bandwidth $\lambda$