Modeling and predicting complex space–time structures and patterns of coastal wind fields

Montserrat Fuentes\textsuperscript{1,2,*}, Li Chen\textsuperscript{1}, Jerry M. Davis\textsuperscript{3} and Gary M. Lackmann\textsuperscript{3}

\textsuperscript{1}Statistics Department, North Carolina State University (NCSU), Raleigh, NC 27695-8203, U.S.A.
\textsuperscript{2}U.S. Environmental Protection Agency (EPA)
\textsuperscript{3}Marine Earth and Atmospheric Sciences Department, North Carolina State University

SUMMARY

A statistical technique is developed for wind field mapping that can be used to improve either the assimilation of surface wind observations into a model initial field or the accuracy of post-processing algorithms run on meteorological model output. The observed wind field at any particular location is treated as a function of the true (but unknown) wind and measurement error. The wind field from numerical weather prediction models is treated as a function of a linear and multiplicative bias and a term which represents random deviations with respect to the true wind process. A Bayesian approach is taken to provide information about the true underlying wind field, which is modeled as a stochastic process with a non-stationary and non-separable covariance. The method is applied to forecast wind fields from a widely used mesoscale numerical weather prediction (NWP) model (MM5). The statistical model tests are carried out for the wind speed over the Chesapeake Bay and the surrounding region for 21 July 2002. Coastal wind observations that have not been used in the MM5 initial conditions or forecasts are used in conjunction with the MM5 forecast wind field (valid at the same time that the observations were available) in a post-processing technique that combined these two sources of information to predict the true wind field. Based on the mean square error, this procedure provides a substantial correction to the MM5 wind field forecast over the Chesapeake Bay region. Copyright © 2005 John Wiley & Sons, Ltd.

key words: Bayesian inference; Fourier transform; geostatistics; meteorological mesoscale model (MM5); non-separable models; non-stationary models; wind fields

1. INTRODUCTION

The wind field in coastal regions is influenced by complex physical processes associated with small-scale variations in terrain, and strong gradients in moisture, temperature, and surface roughness. The fundamental diurnal changes in the structure of the atmospheric boundary layer differ greatly from the land surface to the water surface. Recent work by Titlow and McQueen (1999) has indicated that

\*Correspondence to: M. Fuentes, Statistics Department, North Carolina State University (NCSU), Raleigh, NC 27695-8203, U.S.A.
\textsuperscript{1}E-mail: fuentes@stat.ncsu.edu

Contract/grant sponsor: National Science Foundation; contract/grant numbers: DMS 0002790, DMS 0353029.

Received 17 December 2003
Accepted 20 September 2004

Copyright © 2005 John Wiley & Sons, Ltd.
the current versions of the available operational meteorological models are incapable of producing reliable high-resolution wind forecasts during periods of weak synoptic-scale forcing. Although sea-breeze processes have been adequately simulated in research studies (e.g. Rao and Fuelberg, 2000), operational prediction of small-scale processes such as the land–sea-breeze circulation remains problematic for real-time numerical weather prediction (NWP) models available from the National Centers for Environmental Prediction (NCEP). This is true even at high resolution (e.g. 4 km) grid spacing. The emphasis in this article is on the development of a procedure that could potentially be used to improve either the assimilation of surface wind observations into a model initial field, or the accuracy of post-processing algorithms run on meteorological model output.

In the summer of 2001, the National Aeronautic and Space Administration (NASA) and the Defense Threat Reduction Agency (DTRA) sponsored a field experiment (The Chesapeake Bay Numerical Weather Prediction Model Experiment) over the Chesapeake Bay to collect data pertinent to the numerical modeling community. Data were collected over the period 16 to 31 July. These days were chosen in the hope that weakly forced flow conditions would prevail for a good portion of this period. Unfortunately, only the period 21–23 July fully met that criterion. It was hoped that the data collected could be used to improve the ability of mesoscale models to forecast the state of the atmosphere over complex terrain when small scale processes dominate the flow regime, either by improving initial conditions or in the development of a post-processing technique that could be applied to meteorological model output. Limited data collection was continued into the summer of 2002. The criterion of weakly forced conditions was better met in 2002 than in 2001. This study examines data from the 2002 period. During this period both observed data and MM5 (Mesoscale Model Version 5) output fields were available.

The research presented in this article represents a new approach to wind field modeling, analysis and spatial prediction. The difficult challenge of modeling the spatial structure of wind fields over time can be overcome by using separable processes. A spatial–temporal field \( Z(\mathbf{x}, t) \), where \( \mathbf{x} \) represents space and \( t \) time, is separable if \( \text{Cov} \{ Z(\mathbf{x}, t), Z(\mathbf{x}', t') \} = C_1(\mathbf{x}, \mathbf{x}')C_2(t, t') \) for some spatial covariance \( C_1 \) and temporal covariance \( C_2 \). This class of spatial–temporal processes offers enormous computational benefits, because the covariance matrix can be expressed as the Kronecker product of two smaller matrices that arise separately from the temporal and purely spatial processes, and then its determinant and inverse are easily obtained. Stationarity is another common assumption when modeling spatial processes. A random field \( Z \) in \( \mathbb{R}^d \) is called weakly stationary (or stationary) if it has finite second moments, its mean function is constant and it possesses an autocovariance function \( C \), such that \( C(\mathbf{x} - \mathbf{y}) = \text{cov} \{ Z(\mathbf{x}), Z(\mathbf{y}) \} \). Separability and stationarity are desirable properties for spatial–temporal processes, but are usually rather unrealistic assumptions for large spatial–temporal domains. Little has been written about non-separable, non-stationary models. Classes of non-separable stationary spatial–temporal models have been proposed by Cressie and Huang (1999), Gneiting (2002) and Stein (2003). Non-stationary approaches for spatial prediction have been proposed by Sampson and Guttrop (1992), Smith (1996), Haas (1995), Higdon et al. (1999), Wikle et al. (2001), Fuentes (2001) and (2002), Fuentes and Smith (2001), and Nychka et al. (2003), among others. Here, we define a generalized class of non-stationary and also non-separable spatial–temporal covariance models. New fitting algorithms are also developed to estimate the space–time covariance. The true wind field (speed and direction) is a quantity that is not known. It is modeled as a function of a spatial–temporal trend term and a zero mean process, which has a non-separable and non-stationary covariance that might change with location. In this research, the true wind is determined by conditioning on the observed wind field at specified locations and on the wind fields derived from MM5. The observed wind field at any particular location is treated as a function of the true (but unknown) wind and measurement error.
The wind field from the numerical model is treated as a function of a linear and multiplicative bias and a term which represents random deviations with respect to the true wind process. A Bayesian approach is taken to provide information about the true wind field.

This approach provides a representation of the wind field that is potentially more accurate than the one obtained from the operational data assimilation techniques now in use. Thus the procedure presented can be used to generate improved initial fields for use in mesoscale NWP models, or it can be used as a post-processing technique to be applied to mesoscale model output fields. This Bayesian modified model output field would provide a more accurate picture of the atmosphere for subsequent model runs. A similar approach was proposed by Best et al. (2000), who related different spatially varying quantities to an underlying unobservable random field in a regression analysis of health and exposure data. In our method we also have an underlying unobservable process, but the space–time statistical framework we propose is different. We present a new way to relate meteorological variables to an underlying process: in this case the true wind values. Wikle et al. (2001) present a very interesting hierarchical framework for a similar problem. The Wikle et al. approach is a conditional one in which all the spatial quantities are defined through a series of statistical conditional models. We present here a simultaneous representation of the data and the output of numerical models in terms of the underlying truth. In our method we write the output of the models in terms of the underlying truth, taking into account and characterizing the potential bias of the numerical models. In the Wikle et al. approach the numerical models are used as prior processes for the underlying truth or as sources of data without any type of bias.

This article is organized as follows. In Section 2 we describe the algorithm for wind field mapping, and we introduce new classes of non-separable and non-stationary models for space–time processes. In Section 3 we apply the methods described in this article to the wind field data.

2. WIND MAPPING COMBINING MEASUREMENTS AND MODEL OUTPUT

In our statistical approach to wind field mapping, which combines observed data and model output (Fuentes and Raftery, 2004), we do not consider the measurements at the monitoring stations to be the ‘truth’ because of measurement error. Thus, we assume there is an underlying (unobserved) field \( Z(x, t) \), where \( Z(x, t) \) measures the ‘true’ meteorological variable at location \( x \) and time \( t \). At location \( x \), we make an observation \( \hat{Z}(x, t) \) corresponding to the observation at this location at time \( t \), and we assume that

\[
\hat{Z}(x, t) = Z(x, t) + e(x, t)
\]

where \( e(x, t) \sim N(0, \sigma^2) \) represents the measurement error (nugget) at the location \( x \). We assume that the error term \( e(x, t) \) is independent of the true values \( Z(x, t) \).

The true underlying process \( Z \) is a spatial–temporal process with a non-stationary and non-separable covariance. We represent the large scale structure of \( Z \) using a space–time dynamic statistical model:

\[
Z(x, t) = \sum_{i=1}^{k} \beta_i(x, t)f_i(x, t) + \varepsilon(x, t)
\]

where \( \{f_i\} \) are \( k \) covariates (e.g. sine and cosines and geographic data) of interest with coefficients \( \beta_i \) that vary in space and time. The process \( \varepsilon(x, t) \) has a non-stationary and non-separable covariance (presented in Section 2.2) with a parameter vector that might change with location.
We model the dynamic coefficients $\beta_i$ using a hierarchical model in terms of an overall time component $\gamma_{it}$ and a space–time process $\gamma_i(x, t)$. In our application these are sine and cosine functions,

$$
\begin{align*}
\beta_i(x, t) &= \gamma_{it} + \gamma_i(x, t) \\
\gamma_{it} &= \gamma_{it-1} + u(t) \\
\gamma_i(x, t) &= \gamma_i(x, t-1) + \eta(x, t)
\end{align*}
$$

where $\eta$ and $u$ are independent white noise processes.

We model the output of the meteorological models as follows:

$$
\tilde{Z}(x, t) = a(x, t) + b(x, t)Z(x, t) + \delta(x, t)
$$

Here, the parametric function $a(x, t)$ measures the additive bias of the meteorological models at location $x$ and time $t$, and the parameter function $b(x, t)$ accounts for the multiplicative bias in the meteorological models. The process $\delta(x, t) \sim N(0, \sigma^2_\delta)$ explains the random deviation at location $x$ with respect to the underlying true process $Z(x, t)$. We assume the error term $\delta(x, t)$ is independent of $Z(x, t)$ and $e(x, t)$.

We seek to obtain more reliable wind field spatial predictions by combining model output with observations. Thus, for the prediction of the wind field we simulate values of $Z$ from the posterior predictive distribution of the true underlying wind process given the model output and the observations:

$$
P(Z | \tilde{Z}, \tilde{Z})
$$

This is a calibrated wind prediction. It could also be considered as a data assimilation approach in the sense that we are using the data to improve the wind forecast but, in contrast with the traditional approach of perturbing the model inputs and assimilating data using Kalman-filter methods, we perturb the output using the data to produce a calibrated model output.

### 2.1. Modeling the spatial–temporal dependency structure

We introduce in this section a new class of space–time covariances. We model a space–time process $Z$ as a mixture of (independent) local stationary space–time processes $Z_i$ for $i = 1, \ldots, k$ that explain the space–time dependence structure in subregions of stationarity $S_1, \ldots, S_k$,

$$
Z(x, t) = \sum_{i=1}^{k} K(s - s_i)Z_i(x, t)
$$

where $s_i = (x_i, t_i)$ is the centroid of the $i$th subregion, and for $i = 1, \ldots, k$, $Z_i$ explains the spatial–temporal structure of $Z$ in a subregion of stationarity $S_i$. The corresponding covariance function for the process $Z$ is

$$
\text{cov}\{Z(x_1, t_1), Z(x_2, t_2)\} = \sum_{i=1}^{k} K(s_1 - s_i)K(s_2 - s_i)C_i(x_1 - x_2, t_1 - t_2)
$$
where \( s_1 = (x_1, t_1) \), \( s_2 = (x_2, t_2) \), and each \( C_i \) is a stationary covariance that explains the space–time dependency in a subregion of stationarity \( S_i \).

We propose the following non-separable parametric model for \( C_i \):

\[
C_i(x, t) = \frac{\sigma_i}{2^{\nu_i-1} \Gamma(\nu_i)} \left( \frac{2 \nu_i^{1/2}}{\rho_i} \| (x, \beta_t) \| / \rho_i \right)^{\nu_i} K_{\nu_i} \left( \frac{2 \nu_i^{1/2}}{\rho_i} \| (x, \beta_t) \| / \rho_i \right) \tag{7}
\]

where \( K_{\nu_i} \) is a modified Bessel function and the covariance vector parameter, \( \theta_i = (\nu_i, \sigma_i, \rho_i, \beta_i) \), changes from subregion to subregion to explain the lack of stationarity. The parameter \( \rho_i \) measures how the correlation decays with distance; generally this parameter is called the range. The parameter \( \sigma_i \) is the variance of the random field (i.e. \( \sigma_i = \text{var}(Z_i(x, t)) \)), where the covariance parameter \( \sigma_i \) is usually referred to as the sill. The parameter \( \nu_i \) measures the degree of smoothness of the local stationary process \( Z_i \). The parameter \( \beta_i \) is a conversion factor to take into account the change of units from the temporal to the spatial domain. This parametric model for \( C_i \) is a 3D Matérn type covariance (Matérn, 1960) with an extra parameter (\( \beta_i \)) that can be interpreted as a conversion factor between the units in the space and time domain.

Alternatively, it might be reasonable to assume separability in some subregions of stationarity. Thus, the local processes \( Z_i \) would have a separable spatial–temporal covariance:

\[
\text{cov}(Z_i(x_1, t_1), Z_i(x_2, t_2)) = C_i^{(1)}(x_1 - x_2) \cdot C_i^{(2)}(t_1 - t_2) \tag{8}
\]

where \( C_i^{(1)} \) is a stationary spatial covariance, e.g. a Matérn covariance (Matérn, 1960), and \( C_i^{(2)} \) is a stationary temporal covariance (e.g. the covariance of an autoregressive AR(1) temporal model). Forms such as (8) have a history in spatial–temporal modeling (e.g. Mardia and Goodall, 1993, and references therein). Then, the covariance of \( Z \) can be written as follows:

\[
\text{cov}(Z(x_1, t_1), Z(x_2, t_2)) = \sum_{i=1}^{k} K(s_1 - s_i) K(s_2 - s_i) C_i^{(1)}(x_1 - x_2) \cdot C_i^{(2)}(t_1 - t_2) \tag{9}
\]

This is a non-stationary spatial–temporal covariance. Note that even if the covariances for the processes \( Z_i \) are separable, the global covariance of the process of interest \( Z \) is not separable.

3. APPLICATION

3.1. Description of data

Both observed and model output from a 30-h integration of the MM5 model initialized at 00Z on 21 July 2002 were used in this study. The observed data were collected at 17 locations (see Figure 1) by an independent suite of meteorological stations owned by Weatherflow, Inc. and provided by Jay Titlow of Weatherflow, Inc. The anemometers were location from 9 to 18 m above ground level, depending on the location. Adjustment of the wind speed values to the standard 10-m height was accomplished using Monin–Obukhov similarity theory (Arya, 2001). This adjustment had little effect on the wind speed values given the small distances involved. The MM5 data were furnished by John McHenry at Baron Advanced Meteorological Systems. The MM5 model output fields were generated.
using a 15-km grid. This grid spacing was selected because it is similar to the finest mesh produced by operational models run at NCEP, and based on discussions with Jay Titlow. Winds from the lowest model layer were used after adjustment to the 10-m level.

In this analysis, MM5 model output fields from 21 July 2002 have been examined in detail for the full 24-h period. The complicated flow patterns over the region during this time are evident in Figure 2. The arrows indicate the direction from which the winds are coming, while the length of the stem indicates wind speed. An easterly wind (winds from the east, south-east and north-east) tends to dominate over the majority of the region for this time period. The 3 a.m. (7-h forecast) plot clearly shows an area of confluence (an area where the wind vectors tend to come together) over the Bay. At 9 a.m. (13-h forecast) this area of confluence has been replaced by an area of diffluence (an area where...
the wind vectors tend to spread apart). Diffuence seems to persist over the majority of the Bay at noon (16-h forecast) and for the rest of the period studied. There is little evidence in these plots to suggest that MM5 was capturing the sea breeze circulation, which observations show to be present.

3.2. The non-stationary for wind fields

As a first empirical attempt to deal with the non-stationary inherent in these kinds of environmental data, we divided the spatial domain into two broad categories: land and water. There are major differences in these two regimes with regard to surface roughness, the surface radiation and energy budgets, stability regimes and the overall diurnal variations in the structure of the atmospheric boundary layer. All of these factors have a profound influence on the surface wind field (Stull, 1988). Using the $u$ wind component (the east–west wind), the $v$ wind component (the north–south wind) and the wind speed from the MM5 model output at noon on 21 July 2002, the $K$-means cluster procedure was used to further subdivide these two regions to help identify domains of stationarity. We iterated this process until the Akaike Information Criteria (AIC) or Schwarz Bayesian Information Criterion (BIC) suggested that there was no significant improvement in the estimation of the covariance parameters for the wind speed anomalies (wind value at each location minus the mean over time at that location) using the theoretical covariance model presented in Section 2.2 (6). The optimal number of clusters suggested by the AIC was five. The five subregions are shown in Figure 3. This final regional arrangement of clusters appears reasonable considering atmospheric and oceanic processes that are occurring in the boundary layer on this day.

![Subregions of stationarity](image)

Figure 3. Subregions of stationarity
There are 48 meteorological variables available in the first sigma layer in our MM5 output. A
number of these variables or combinations of them (e.g. \( h/L \), a boundary layer stability parameter,
where \( h \) is the height of the boundary layer and \( L \) is the Monin–Obukhov length) can be used to
identify regions of homogeneous wind field regimes. We looked at many combinations of these
variables, and found that the ones we used provided us with regions where, based on the physical
processes involved, we would expect to find spatial stationarity in the wind fields. This is not a trivial
problem given the complex space/time variations that are part of the daily cycles that occur in the
atmospheric boundary layer. We acknowledge that there are multiple methods for accomplishing this,
and that this is an area of possible future improvement. However, using the AIC criteria combined with
the \( K \)-means cluster method does provide an objective technique as opposed to arbitrarily subdividing
the domain based purely on geographical factors, for example. It is reasonable to assume that there will
be some changes in the configuration of these regions as time passes and the boundary layer structure
changes. For the purposes of this work, we assumed that the changes were negligible. However, for
the other times that we examined on that day, the clusters remained reasonably stable.
Therefore, the underlying true process for wind speed was modeled as a mixture of five orthogonal
stationary spatial–temporal processes corresponding to the five subregions obtained above.

3.3. Spatial–temporal trend

A spatial–temporal trend with two components was fitted using the the space–time dynamic model
proposed in Section 2.1. One component is the overall temporal trend for the process \( Z \) (the truth); it
does not change with location. The other part is the temporal variation that changes with location. The
covariates used here are sine and cosine functions with respect to two different periods (1/frequency),
24 h and 12 h, which capture the diurnal and half-diurnal cycles. These two components of the spatial–
temporal trend are shown in Figure 4. This trend is fitted using a Bayesian hierarchical framework. The
parameters are estimated using the mean of the posterior distribution. Figure 4(a) shows the overall
wind speed trend over time. From Figure 4(a), it appears that the highest wind speeds over the region
as a whole occur near sunset, while the minimum wind speeds on that day (21 July) occur several hours
before sunrise. Figure 4(b) shows the temporal variation for each subregion. The anomalies for
subregions 1 and 3 are generally negative over time, while for subregions 4 and 5 they are generally
positive. Subregion 2 starts out positive but then becomes negative. The large positive values in
subregion 5 indicate that the wind speeds in that subregion are higher than the average during the mid-
day hours, while the large mid-day negative values for subregion 1 indicate that the wind speeds are
lower than average during the mid-day hours in that subregion.

3.4. Spatial–temporal covariance structure

3.4.1. Empirical covariance analyses. We start with an empirical analysis of the covariance structure.
We draw empirical covariance graphs for each subregion for the wind speed anomalies. Figure 5
clearly shows pronounced ridges in the empirical covariances for subregions 1, 2, 3 and 5. These type
of ridges appear in separable covariance structures. On the other hand, the empirical covariance plot
for subregion 4 is smoother; the lack of a strong ridge effect suggests a non-separable covariance
structure. The tests for stationarity, separability and isotropy proposed by Fuentes (2004) where
performed within each subregion. Stationarity was found to hold within each subregion. Separability
was found to hold within all subregions except for subregion 4. The results from the formal tests agree
with the empirical covariance analyses. Furthermore, anisotropy was found to hold within subregions
1 and 5.
Figure 4. Spatial–temporal trend for wind speed: (a) overall temporal trend; (b) mean temporal variation for each subregion

Figure 5. Space–time empirical covariance
3.4.2. Theoretical covariance model. We model the covariance of $Z$ with five subregions of stationarity (Figure 3) using model (6) proposed in Section 2.2. The weight function in (6) is modeled as $K(x - x_i) = 1/h_i^2 K_0((x - x_i)/h_i)$, where the location $x_i$ is the centroid of the $i$th subregion. The bandwidth $h_i$ is defined as half of the maximum distance for the $i$th subregion. The function $K_0$ is modeled as $K_0(u) = 3/4(1 - u_1^2)_+ + 3/4(1 - u_2^2)_+$, which is a quadratic weight function for $u = (u_1, u_2)$. For subregion 4 we use the non-separable model (7). For the separable spatial–temporal covariances in the other subregions, we use Matérn models for the spatial covariance and exponential models for the temporal covariance. We fit the parameters for the spatial–temporal covariance of $Z$ using a Bayesian framework. In our covariance model, we allowed for geometric anisotropy. Subregions 1 and 5 are anisotropic, which means that the spatial covariance depended not only on distance but also direction. A linear transformation was used to transform the co-ordinates $x = (x_1, x_2)$. The new co-ordinates are $(x'_1, x'_2) = (x_1, x_2)RT$, where $R$ is a rotation matrix and $T$ is a shrinking matrix. $R$ and $T$ are defined as follows:

$$R = \begin{pmatrix} \cos A & \sin A \\ -\sin A & \cos A \end{pmatrix} \quad T = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{r} \end{pmatrix}$$

The parameter $A$ describes the angle of rotation and $r$ is used to stretch the co-ordinates. These anisotropic parameters are estimated using their posterior distributions, as are the other covariance parameters.

The priors for the sill parameter, spatial range parameter and temporal range parameter are inverse Gamma with infinite variance. The prior distribution for the smoothness parameter is a uniform distribution with support $(0, 2)$. The prior distributions for the anisotropic parameters are uniform distributions with support $[0, 2\pi]$ for $A$ and $(0, 5)$ for $r$. The support for the smoothness parameter is a conservative interval based on our previous experience analyzing similar datasets.

The posterior distribution is determined by the likelihood function and the prior distribution. In order to show how sensitive the posterior distribution is with respect to the prior distribution, we plot the posterior distribution for the sill parameter corresponding to different priors. In Figure 6, we show results for subregions 1 and 4. Similar analyses were done for all subregions and all parameters, leading to the same conclusions. Since the sill parameter is positive, the following three different priors are used: (a) inverse Gamma $(0.4, 2)$; (b) reciprocal; and (c) lognormal $(5, 1)$. Figure 6 indicates that the different choices for the prior have little effect on the posterior distribution.

The posterior densities for the sill parameter of each subregion are shown in Figure 7. The differences in the subregions (see Figure 3) are also reflected in the posterior distributions for the sill parameter (Figure 7). The largest mean of the posterior distributions was found for subregion 4, which encompasses a large variety of mainly water surfaces. In addition, the region is characterized by a large variety of coastal orientations. These large variations in the surface roughness characteristics could contribute to large variations in wind speed across the region. Stability differences would exist between land and water surfaces due in large part to the differential heating experienced by these surfaces. This could also contribute to the spatial variations in the wind speed. Subregion 4 is the least homogeneous of the five subregions; the second most spatially diverse subregion is subregion 1, whose sill value reflects this diversity. Moreover, the significant difference in the sill parameter is evidence for the non-stationary exhibited over the spatial–temporal domain.

The posterior distribution for the spatial range (Figure 8) indicates that subregion 5 has the highest posterior mean. Given the nature of this mainly deeper water subregion, one would expect strong...
spatial continuity. The posterior mean is lower for the other subregions, where the spatial continuity is weaker. Unlike the sill and the spatial range parameters, the spatial smoothness parameter does not change much from subregion to subregion, the mean of the posterior distribution for this parameter always being between 0.5 and 1.5. For subregion 4, the non-separable spatial–temporal covariance model (7) is fitted.

Figure 6. Posterior, prior and likelihood function for the sill parameter, with different choices for the prior: (a) inverse Gamma (0.4, 2); (b) reciprocal; (c) lognormal (5, 1)

Figure 7. Triplots for the sill parameter
The posterior means of the covariance parameters are listed in Table 1. In Figure 9 we show the MCMC traces for the sill (in the logarithmic scale), range and smoothness parameters in one of the subregions. We iterate through the algorithm discussed in the Appendix till there is evidence that convergence has been achieved.

Table 1. Estimated covariance parameters

<table>
<thead>
<tr>
<th>Process</th>
<th>Sill (m²/s²)</th>
<th>Spatial range (km)</th>
<th>Temporal range (h)</th>
<th>Smoothness</th>
<th>A</th>
<th>R</th>
<th>Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reg. 1</td>
<td>0.1615</td>
<td>14.44</td>
<td>2.02</td>
<td>0.70</td>
<td>2.78</td>
<td>1.11</td>
<td></td>
</tr>
<tr>
<td>Reg. 2</td>
<td>0.1275</td>
<td>17.62</td>
<td>1.64</td>
<td>0.88</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reg. 3</td>
<td>0.1200</td>
<td>16.24</td>
<td>1.92</td>
<td>0.99</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reg. 4</td>
<td>0.3073</td>
<td>16.61</td>
<td>1.42</td>
<td>13.24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reg. 5</td>
<td>0.0944</td>
<td>25.64</td>
<td>1.73</td>
<td>0.58</td>
<td>1.18</td>
<td>1.17</td>
<td></td>
</tr>
</tbody>
</table>

Figure 8. Posterior distribution for the range parameter

Figure 9. MCMC traces for the sill (in the logarithmic scale), range and smoothness parameters in subregion 3
3.5. Wind field mapping

The original MM5 output wind speed map is shown in Figure 10(a). The colors in the image represent wind speeds from MM5, and the value on top of the image is the observed wind speed. Figure 10(a) indicates that there is a substantial difference between the MM5 forecast and the observed data. With our approach, we simulate values of the wind speed from the posterior predictive distribution of the underlying true wind process given the model output and the observations. Figure 10(b) shows an improved wind field map obtained by combining model output and observed wind data.

In Figure 10(b), the color of the background indicates the mean of posterior predictive distribution. The improved map agrees better with the observed data. In order to quantify this improvement, we calculated

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (\hat{r}_i - r_{oi})^2,$$

where $r_{oi}$ is the observed wind speed at location $i$, $\hat{r}_i$ is the predicted wind speed at location $i$ (without using $r_{oi}$), and $n$ is the number of observations. The MSE for the MM5 forecast is $10.32 \text{ m}^2/\text{s}^2$, and with our approach we reduced it to $2.45 \text{ m}^2/\text{s}^2$. We also calculated MSEs for the 3D-Var technique, which is a three-dimensional variational assimilation approach (Kalnay, 2003). This is a Kalman-filter based technique used by meteorologists to improve the numerical model output. We obtained a larger MSE (compared to our approach) of $9.17 \text{ m}^2/\text{s}^2$. Similar comparisons were made at other times giving promising results.

4. DISCUSSION

In this article we introduce new statistical models for the spatial prediction of wind fields using different sources of data. We combine spatial data by relating the spatially varying variables to an underlying unobservable true wind field process, and we predict this latent process. The general statistical models introduced in this article for spatial–temporal processes allow for the lack of
stationarity and isotropy, and they do not assume separability. From a meteorological viewpoint, this technique could be used to improve the assimilation of the surface wind field observations into a meteorological model initial field. In addition, it could serve as a post-processing algorithm which would be run on meteorological model output fields. The meteorological community has been working on the data assimilation problem for many years. This research has produced some very sophisticated methodologies. For a current review of this literature see Kalnay (2003). The work by Daley (1991) and Talagrand (1997) should also be consulted.

In areas where there are many observing sites, the observed data have the most influence on the model initial field, while in areas where observational data are scarce the model first guess field is important since it can build on the observational data from data rich areas. The procedure outlined in this article provides a powerful technique for combining the first guess field with the available observations to create an improved field of initial conditions. In order to rigorously quantify the improvement offered by this technique, controlled experimental model forecasts could be compared: the control run with an unperturbed initial condition, and the experimental run with an adjusted initial condition. The application presented in this article is a model output post-processing correction. Here, corrections are applied to a 16-h NWP forecast using uniquely available coastal wind observation. The improvement obtained in the mean square error when the meteorological model output and the observed data are combined is very good. However, these results are based on a very limited time period and have not been rigorously compared to other currently available methods. While the techniques presented in this article have great promise, both further testing on larger data sets and comparison with existing techniques will show their true worth.

ACKNOWLEDGEMENTS

We would like to thank two referees for their helpful comments and suggestions. This research was sponsored by a National Science Foundation grant DMS 0002790 and DMS 0353029, and by a U.S. Defense Threat Reduction Agency (DTRA) grant.

APPENDIX: ALGORITHM FOR COMBINING DATA

We describe the steps in the algorithm used in this article to combine numerical model output and observational data taking into account the lack of stationarity and separability.

First, we choose the number of spatial subregions of stationarity. We start with two categories: land and water. Then, we further divide these two subregions (land and water) using the K-means cluster procedure, and we apply AIC to determine whether there is a significant improvement in the estimation of the covariance parameters for the wind speed anomalies by adding more subregions of stationarity. We iterate this process until the AIC indicates that there is no significant improvement in the estimation of the covariance parameters by increasing the number of subregions of stationarity. The optimal number of subregions suggested by AIC in this application was five. Then, using these five subregions, we fit the new model presented in the Section 2 to the wind data in a fully Bayesian framework.

Let \( \boldsymbol{\theta} \) be the collection of all parameters, \( \boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3) \), where \( \boldsymbol{\theta}_1 \) are the parameters in the statistical model for the point observations (1), \( \boldsymbol{\theta}_2 \) are the parameters in the statistical model for the numerical model output (3), and finally \( \boldsymbol{\theta}_3 \) is a collection of all parameters used in (2) to model the
underlying true process. The posterior distribution for $\theta$ given the observations $\mathbf{Z}$ and the model output $\mathbf{Z}$, $\theta | \mathbf{Z}$, is

$$ \theta | \mathbf{Z}, \mathbf{Z} \propto \int \mathbf{Z}, \mathbf{Z}, \mathbf{Z} | \theta, \mathbf{Z} d\mathbf{Z}, $$

where the brackets denote a density function. We know that

$$ \mathbf{Z}, \mathbf{Z}, \mathbf{Z} | \theta, \mathbf{Z} = \mathbf{Z}, \mathbf{Z} | \theta, \mathbf{Z} \mathbf{Z} | \theta, \mathbf{Z} = \mathbf{Z}, \mathbf{Z} | \theta, \mathbf{Z} \mathbf{Z} | \theta, \mathbf{Z} = \mathbf{Z}, \mathbf{Z} | \theta, \mathbf{Z} \mathbf{Z} | \theta, \mathbf{Z} \mathbf{Z} | \theta, \mathbf{Z} $$

We use a multiple-stage Gibbs sampling approach. We cycle through three stages. In Stage 1, we estimate the posterior distribution of the mean and covariance parameters for the true wind process ($\theta_3$). In Stage 2, we obtain the posterior distribution of the parameters that explain the measurement error in the observed data ($\theta_1$) given the underlying true wind fields, and in Stage 3 we estimate the posterior distribution of the bias in the numerical model output ($\theta_2$) given $\mathbf{Z}$. Thus, we obtain $\{\theta^{(i)}\}_{i=1}^N$, which are $N$ simulated values from the posterior distribution of the vector parameter $\theta$.

The posterior predictive distribution of the true wind process $\mathbf{Z}$ at the location $x_0$ and time $t_0$ given all the available data $\mathbf{Z} = (\mathbf{Z}, \mathbf{Z})^T$ is

$$ P(\mathbf{Z}(x_0, t_0) | \mathbf{Z}) \propto \int P(\mathbf{Z}(x_0, t_0) | \mathbf{Z}, \theta) P(\theta | \mathbf{Z}) d\theta $$

This posterior predictive distribution is approximated by

$$ P(\mathbf{Z}(x_0, t_0) | \mathbf{Z}) = \frac{1}{N} \sum_{i=1}^N P(\mathbf{Z}(x_0, t_0) | \mathbf{Z}, \theta^{(i)}) $$

REFERENCES


