Supplementary materials for this article are available online. Please click the Technometrics link at http://pubs.amstat.org.

Comparing Spatial Predictions

Amanda S. Hering

Department of Applied
Mathematics and Statistics
Colorado School of Mines
Golden, CO 80401-1887
(*ahering@mines.edu*)

Marc G. Genton

Department of Statistics
Texas A&M University
College Station, TX 77843-3143
(*genton@stat.tamu.edu*)

Under a general loss function, we develop a hypothesis test to determine whether a significant difference in the spatial predictions produced by two competing models exists on average across the entire spatial domain of interest. The null hypothesis is that of no difference, and a spatial loss differential is created based on the observed data, the two sets of predictions, and the loss function chosen by the researcher. The test assumes only isotropy and short-range spatial dependence of the loss differential but does allow it to be non-Gaussian, non-zero-mean, and spatially correlated. Constant and nonconstant spatial trends in the loss differential are treated in two separate cases. Monte Carlo simulations illustrate the size and power properties of this test, and an example based on daily average wind speeds in Oklahoma is used for illustration. Supplemental results are available online.

KEY WORDS: Hypothesis test; Kriging; Loss functions; Model validation; Prediction evaluation; Wind power.

1. INTRODUCTION

Making predictions is one of the primary reasons to invest effort in building models that capture the salient features of data. These predictions are used as a guide to make practical decisions. Poor predictions can lead to poor decisions and, ultimately, to abandoning the model used to produce them, whereas good ones can save time, money, and resources. Decision makers are often faced with choosing between the predictions produced by more than one model. Therefore, formally comparing the predictions from competing models is necessary to be confident that the chosen predictive model truly produces superior predictions.

Comparing the accuracy of time series forecasts began with the seminal work of Diebold and Mariano (1995). They introduced a test for the null hypothesis of equal forecast accuracy between two competing models. Their test, hereafter referred to as the DM test, can be used with the researcher's choice of loss functions, makes no distributional assumptions on the forecast errors, and incorporates both serial and contemporaneous correlations in competing forecast errors. Many extensions and improvements to this test have been made. West (1996) and McCracken (2004) addressed the estimation of regression coefficients in the test; Harvey, Leybourne, and Newbold (1997) adjusted the bias of the estimated variance; Dell'Aquila and Ronchetti (2004) proposed a robust version of the test; and finally Giacomini and White (2006) unified much of the preceding theory and generalized the test to include evaluations of point, interval, probability, and density forecasts.

We develop a hypothesis test to compare predictions for spatial data that incorporates unique features of spatial data not encountered in time series, and we show how this test improves upon the DM test in the one-dimensional setting. Spatial predictions are made for many variables such as temperature, precipitation, air pollution, concentration of geological resources such as oil and coal, home prices, and disease concentrations. In the past, authors who have attempted to apply the DM test in a spatial setting have discarded data to create an "independent" dataset (Snell, Gopal, and Kaufmann 2000; Wang et al. 2007). Some have simply noted that no such test is available that incorporates the spatial correlation across predictions (Longhi and Nijkamp 2007). Many give point estimates of predictive accuracy or choose the model that minimizes some loss function, but they may not quantify the uncertainty associated with those estimates or include potential spatial dependence in their estimates; for example, see the works of Atger (2003), Gong, Barnston, and Ward (2003), Kleiber et al. (2011), Willis (2002).

Currently, predicting wind speeds for wind power generation is a particularly important area of application in which such a prediction comparison test would be beneficial (Willis 2002; Genton and Hering 2007). No cost-effective method for storing wind energy exists, so it must be used as soon as it is produced. This variable supply makes it difficult for utility managers to maintain a balance between the supply and demand of electricity. If they fail to maintain this balance, they incur monetary penalties imposed by regulatory agencies. The United States possesses vast regions in which many wind farms have been built. For a given point in time, spatial predictions at these wind farms help utility managers plan for the transmission, purchase, and distribution of electricity. Predictions made by competing models can be evaluated with a unique loss function that incorporates the nonlinear relationship between wind speed and wind power with a penalty for overestimation or underestimation of wind speed (Hering and Genton 2010). The prediction comparison test that is described in this work would be instrumental in determining if a difference in the average loss produced by competing predictions is significant.

The spatial prediction comparison test (SPCT) we describe here is appropriate for testing the null hypothesis that on average there is no significant difference between two sets of spatial predictions. It does not require the prediction errors to be

TECHNOMETRICS, NOVEMBER 2011, VOL. 53, NO. 4
DOI 10.1198/TECH.2011.10136

Gaussian or zero-mean, and it allows for both spatial correlation within the prediction errors and contemporaneous correlation between the prediction errors. Contemporaneous correlation is an important element to consider since many models share sources of information, thereby making simultaneously good or bad predictions at a given location. One final advantage of this type of testing is that loss functions beyond the conventional mean squared error (MSE) are allowed. For example, a researcher may want to penalize overestimation more heavily than underestimation, in which case the loss function could be a piecewise linear function (Gneiting 2011).

To the best of our knowledge, no other method exists that tests the same hypothesis as this proposal. Some methods decompose the total prediction error of a single spatial model into components due to location, magnitude, and pattern errors, such as those in the articles by Ebert and McBride (2000) and Gilleland et al. (2010) and the references therein. Predictions and realizations are also compared in terms of orientation and scale, but they mainly focus on developing measures that quantify the types of prediction errors of a single model. Another approach tests for a difference in spatial signal at every location in the domain, which yields as many hypotheses as there are locations in the dataset. To improve the power when testing many hypotheses, the spatial field is transformed with a discrete wavelet decomposition (Shen, Huang, and Cressie 2002; Sedur, Maxim, and Whitcher 2005; Pavlicová, Santer, and Cressie 2008). In particular, Shen, Huang, and Cressie (2002) and Pavlicová, Santer, and Cressie (2008) outlined enhanced versions of the false discovery rate methodology. The goal of these wavelet-based approaches is to determine if a significant difference between two spatial signals exists at each location and also where in the domain the difference occurs; however, the data structure required for these tests is quite rigid. The data must be on a regular grid, and the grid size must be a dyadic power. For irregularly spaced data, the data must be coerced to a grid, and any missing values must be imputed (Nychka, Wikle, and Royle 2002; Shi and Cressie 2007; Matsuo, Nychka, and Paul 2010). Nonstandard grid sizes need to be padded with zeroes or a combination of multiscale wavelets may be used (Deckmyn and Berre 2005).

The test procedure presented in this article has the advantage of being computationally fast and simple to implement. Only one hypothesis needs to be tested versus as many hypotheses as there are locations for the wavelet-based methods. We proceed in Section 2 by reviewing the DM test of forecast accuracy in time series and compare it with the approach that will be generalized for spatial prediction in Section 3. Size and power properties of the SPCT test are demonstrated with Monte Carlo experiments in Section 4. Section 5 provides an applied example of this test to daily average wind speeds in Oklahoma, and we conclude with some additional discussion of the local wavelet-based approaches and potential extensions of the test in Section 6.

2. THE DIEBOLD–MARIANO TEST

Let $\{\hat{y}_{1t}\}_{t=1}^{T}$ and $\{\hat{y}_{2t}\}_{t=1}^{T}$ be two forecasts of the same time series $\{y_t\}_{t=1}^{T}$. The associated forecast errors are $\{e_{1t}\}_{t=1}^{T}$ and $\{e_{2t}\}_{t=1}^{T}$ where $e_{it} = y_t - \hat{y}_{it}$. The time-$t$ loss associated with a forecast can be an arbitrary function of the realization and the prediction, denoted $g(y_t, \hat{y}_{it})$, $i = 1, 2$, which is often a function of the forecast error. Thus, for simplicity, the loss function will be written as $g(e_{it})$ for $i = 1, 2$. The null hypothesis of equal forecast accuracy for two sets of forecasts is $H_0 : E[g(e_{1t})] = E[g(e_{2t})]$ or $H_0 : E[d_t] = 0$, where $d_t := g(e_{1t}) - g(e_{2t})$ is the loss differential.

The sample path $\{d_t\}_{t=1}^{T}$ is assumed to be covariance stationary and short memory. Thus, the asymptotic distribution of the sample mean loss differential as T goes to infinity, $\bar{d} = \frac{1}{T}\sum_{t=1}^{T}[g(e_{1t}) - g(e_{2t})]$, is $\mathrm{N}(\mu, 2\pi s_d(0))$. Here, μ is the population mean loss differential, and $s_d(0)$ is the spectral density of the loss differential at frequency 0.

The large-sample standard normal test statistic for forecast accuracy is then

$$S_1 = \frac{\bar{d}}{\sqrt{2\pi \hat{s}_d(0)/T}},$$

where $\hat{s}_d(0)$ is a consistent estimator of $s_d(0)$. This consistent estimator is obtained by taking a weighted sum of the available sample autocovariances for a k-step forecast,

$$2\pi \hat{s}_d(0) = \sum_{\tau=-(k-1)}^{k-1} \hat{\gamma}_d(\tau), \tag{1}$$

where

$$\hat{\gamma}_d(\tau) = \frac{1}{T}\sum_{t=|\tau|+1}^{T} (d_t - \bar{d})\big(d_{t-|\tau|} - \bar{d}\big)$$

are the empirical autocovariances. The truncation at lag $k-1$ in Equation (1) is motivated by the result that the optimal k-step forecast errors have significant autocorrelations up to lag $k-1$, which can be checked empirically.

Diebold and Mariano (1995) claimed that even though the resulting estimator of the spectral density is not guaranteed to be positive semidefinite, since the estimator assigns a large positive weight near the origin, the estimate of $s_d(0)$ is unlikely to be negative. In practice, they treated a negative estimate of $s_d(0)$ as an automatic rejection of the null hypothesis. However, in small samples, it is not unusual to obtain a negative estimate of $s_d(0)$.

We suggest avoiding this problem by fitting a covariance model to the empirical autocovariances that is guaranteed to be positive definite. Instead of truncating the sum in Equation (1), we compute all of the sample autocovariances for lags $L = 0, 1, 2, \ldots, T-1$. Since empirical autocovariances at higher lags are more variable given that fewer observations are available to compute them, only the empirical autocovariances computed up to half of the maximum lag are retained, which is a common rule of thumb in spatial statistics. An exponential covariogram of the form $C(\tau) = \sigma^2 \exp(-3\tau/\theta)$ is fitted to the empirical autocovariances, where the practical range, θ, is the distance beyond which the correlation is less than 0.05. We replace $\hat{\gamma}_d(\tau)$ in Equation (1) with $\hat{C}(\tau)$ for a parametric estimate of the spectral density at zero,

$$2\pi \hat{s}_d^{p}(0) = \hat{C}(0) + 2\sum_{\tau=1}^{T-1} \hat{C}(\tau),$$

which yields a parametrically estimated test statistic

$$S_p = \frac{\bar{d}}{\sqrt{2\pi \hat{s}^p_d(0)/T}}.$$

We compare this method to that described by Diebold and Mariano (1995) by simulating forecast errors as they describe and applying the quadratic loss. The observed size of the test is dramatically improved with S_p. In fact, empirical sizes reach the desired level at $n = 128$ for the DM test but already at $n = 32$ for the parametric test. Full results are available in supplementary materials, which can be accessed online.

3. SPATIAL PREDICTION COMPARISON TEST

We propose a test for spatial prediction comparison following the form of the DM test for time series data. Consider a spatial process $\{Z(\mathbf{s}) \in \mathbb{R} : \mathbf{s} \in \mathcal{D} \subset \mathbb{R}^2\}$ that has been observed at n locations. The observed value is denoted $Z(\mathbf{s}_i)$, for $i = 1, \ldots, n$. The location of each observation is denoted by $\mathbf{s}_i = (x_i, y_i)$. A fraction of these n observed values, ϕ, is reserved to be predicted based on models built from the $\lfloor (1-\phi)n \rfloor$ observations, where $\lfloor x \rfloor$ is the largest integer not greater than x. Let L represent the number of randomly chosen locations to predict, thus $L = \lfloor \phi n \rfloor$. Two sets of spatial predictions are made, denoted by $\{\hat{Z}_1(\mathbf{s}_i)\}_{i=1}^L$ and $\{\hat{Z}_2(\mathbf{s}_i)\}_{i=1}^L$. We assume that the parameters in both models can be estimated well by the data, which is the case when the number of observations used to estimate the parameters is large relative to the number of predictions made (West 1996). We also assume that influential outliers are not present in the data. The associated prediction errors are $\{e_1(\mathbf{s}_i)\}_{i=1}^L$ and $\{e_2(\mathbf{s}_i)\}_{i=1}^L$.

The location-i loss associated with a prediction, say j, could be an arbitrary function of the realization and the prediction, $g(Z(\mathbf{s}_i), \hat{Z}_j(\mathbf{s}_i))$. For example, in many atmospheric applications, the correlation or "skill" between the predictions and the observed values is computed (Gong, Barnston, and Ward 2003). In this setting, the loss function $g(\cdot)$ would be defined as follows:

$$g(Z(\mathbf{s}_i), \hat{Z}_j(\mathbf{s}_i)) = \frac{L}{(L-1)\hat{\sigma}_Z \hat{\sigma}_j}(Z(\mathbf{s}_i) - \bar{Z})(\hat{Z}_j(\mathbf{s}_i) - \bar{Z}_j),$$

where $\bar{Z}$ is the mean of the L observed values, $\bar{Z}_j$ is the mean of the L predictions from model j, $\hat{\sigma}_Z$ is the estimated standard deviation of the observed values, and $\hat{\sigma}_j$ is the estimated standard deviation of the predictions. In this way, the correlation skill of the predictions is $r = (1/L)\sum_{i=1}^L g(Z(\mathbf{s}_i), \hat{Z}_j(\mathbf{s}_i))$ and can be used to compare competing models. However, often the loss function, g, will be a direct realization of the prediction error, $g(e_j(\mathbf{s}_i))$ for $j = 1, 2$, and this abbreviated notation is used in the remaining description.

The spatial process of interest takes the following form:

$$D(\mathbf{s}) = g(e_1(\mathbf{s})) - g(e_2(\mathbf{s})) = f(\mathbf{s}) + \delta(\mathbf{s}), \qquad \mathbf{s} \in \mathcal{D},$$

where $f(\mathbf{s})$ is the mean trend, and $\delta(\mathbf{s})$ is a mean-0 stationary process with unknown covariance function $C(\mathbf{h}) = \mathrm{cov}(\delta(\mathbf{s}), \delta(\mathbf{s}+\mathbf{h}))$. This process has been observed at locations $\{\mathbf{s}_i : i = 1, \ldots, L\}$. We wish to test the null hypothesis of equal predictive ability on average

$$H_0 : \frac{1}{|\mathcal{D}|}\int_{\mathcal{D}} E[D(\mathbf{s})]\, d\mathbf{s} = 0, \tag{2}$$

where $|\mathcal{D}|$ is the area of the domain, $\mathcal{D}$. The process $D(\mathbf{s})$ is referred to as the loss differential, and it is assumed to be isotropic with short range covariance. Requiring that $D(\mathbf{s})$ be stationary implies that the mean trend must be constant in space. When the trend is assumed to be constant in space, $f(\mathbf{s}) = \mu$, then the null hypothesis becomes $H_0 : \mu = 0$. However, in many cases it is unlikely that the mean is the same at every location, and the null hypothesis would then test that the average of the spatially varying mean across all locations is zero.

Based on the two possible forms of $f(\mathbf{s})$, either constant or spatially varying, two versions of the spatial prediction comparison test will be treated separately. When estimating an unknown trend, it becomes important to distinguish between variability in $D(\mathbf{s})$ due to trend and variability due to spatial dependence. If the trend is misspecified as spatial dependence, then the estimate of the variability of $D(\mathbf{s})$ increases. Likewise, including spatial dependence in the trend estimation will reduce the variability of $D(\mathbf{s})$. In the former case, the test for prediction comparison will be undersized, and power will be too low; in the latter case, the test will be oversized, rejecting the null hypothesis too often.

Under increasing domain asymptotics in which the domain is allowed to grow without bound and the spatial covariance approaches zero as the lag distance increases (Park et al. 2009), the sample mean loss differential, $\bar{D} = \frac{1}{L}\sum_{i=1}^L D(\mathbf{s}_i)$, is asymptotically normal,

$$\frac{\bar{D} - \mu}{\sqrt{\mathrm{var}(\bar{D})}} \to \mathrm{N}(0,1)$$

in distribution as L goes to infinity, where

$$\mathrm{var}[\bar{D}] = \frac{1}{L^2}\sum_{i=1}^L \sum_{j=1}^L C(h_{ij}). \tag{3}$$

Here, $C(h_{ij})$ is the covariance function for the loss differential's spatial dependence structure, $\delta(\mathbf{s})$, and h_{ij} is the distance between points $\mathbf{s}_i$ and $\mathbf{s}_j$. As the domain is allowed to grow, the estimation of the spatial dependence improves. All forms of the test statistics we employ to test the hypothesis in Equation (2) utilize some version of Equation (3) in which $C(h_{ij})$ is replaced by an estimate.

Estimating $C(h_{ij})$ is not as straightforward as it may initially seem. First, assume that the trend is constant across space, that is, $f(\mathbf{s}) = \mu$ for μ some constant. The natural empirical estimate of $C(h_{ij})$ is

$$\hat{C}(h_{ij}) = \frac{1}{|N(h_{ij})|}\sum_{N(h_{ij})} (D(\mathbf{s}_i) - \bar{D})(D(\mathbf{s}_j) - \bar{D}), \tag{4}$$

where $N(h_{ij})$ is the set of all pairs of locations that are distance h_{ij} apart. This estimator is generally not positive definite and is biased when the mean must be estimated (Schabenberger and Gotway 2005). In addition, we have the following fact that follows from a similar outcome in time series (Brockwell and Davis 1991; Percival 1993).

Proposition 1. For any set of spatial data, $D(\mathbf{s}_i)$, $i = 1, \ldots, L$, the empirical covariance given in Equation (4) satisfies $\sum_{i=1}^{L} \sum_{j=1}^{L} \hat{C}(h_{ij}) = 0$.

The proof is given in the Appendix and has the following consequences:

1. The estimate of the variance of $\bar{D}$ in Equation (3) is zero when substituting $\hat{C}(h_{ij})$ for $C(h_{ij})$.
2. Since $\hat{C}(0) > 0$ unless $D(\mathbf{s})$ is constant in $\mathbf{s}$, then at least some of the $\hat{C}(h_{ij})$ are constrained to be negative for some lag distances even though the true spatial covariance may not be negative.
3. Using $\hat{C}(h_{ij})$ as a basis for a parametric estimate of $C(h_{ij})$ can yield misleading estimates of the parameters since negative values of $\hat{C}(h_{ij})$ will decrease the strength of the spatial correlation.

This problem does not arise for the DM test since the sum in Equation (3) is truncated at $k-1$ when making k-step forecasts; however, due to the third consequence listed above, $\hat{C}(h_{ij})$ and any models based on $\hat{C}(h_{ij})$ are unreliable and will no longer be considered. An alternative to estimating the covariogram is to estimate the semivariogram, $\gamma(h_{ij})$, up to half of the maximum pairwise distance (Cressie 1993; Schabenberger and Gotway 2005), fit a parametric estimate, and take advantage of the relationship $\gamma(h_{ij}) = C(0) - C(h_{ij})$. Then, $\hat{C}(h_{ij})$ in Equation (3) can be replaced with $\hat{\gamma}(\infty) - \hat{\gamma}(h_{ij})$, where

$$\hat{\gamma}(h_{ij}) = \frac{1}{2|N(h_{ij})|} \sum_{N(h_{ij})} (D(\mathbf{s}_i) - D(\mathbf{s}_j))^2. \tag{5}$$

Standard texts such as the book by Cressie (1993) describe how to fit parametric semivariograms to data. Our approach is to use weighted least squares (WLS), minimizing

$$W_V(\boldsymbol{\theta}) = \sum_{\{h_{ij}>0\}} |N(h_{ij})| \left(\frac{\hat{\gamma}(h_{ij})}{\gamma(h_{ij}|\boldsymbol{\theta})} - 1 \right)^2 \tag{6}$$

according to Cressie (1985a). In Equation (6), the function $\gamma(h_{ij}|\boldsymbol{\theta})$ is the parametric form of the semivariogram, with parameters $\boldsymbol{\theta}$. The maximum distance to which to sum is defined as half of the maximum pairwise distance in the data. Using maximum likelihood to estimate the parameters in the semivariogram is another possible approach, but this requires knowledge of the distribution of the data at each location. A Gaussian model is typically fit in practice (Mardia and Marshall 1984), but the application of the loss function to the prediction errors can change the distribution of the data even when the prediction errors are Gaussian. Therefore, assumptions about the distribution of the data are avoided when fitting the semivariogram with weighted least squares.

Thus, we propose the following test statistic for testing the null hypothesis of equal predictive ability on average under constant trend:

$$S_V = \frac{\bar{D}}{\sqrt{\frac{1}{L^2} \sum_{i=1}^{L} \sum_{j=1}^{L} (\hat{\gamma}(\infty|\hat{\boldsymbol{\theta}}) - \hat{\gamma}(h_{ij}|\hat{\boldsymbol{\theta}}))}}. \tag{7}$$

In application, the assumption of isotropy should be investigated first by plotting the directional semivariograms, and the function `eyefit` in the `R` package `geoR` can be helpful in finding a good-fitting parametric model and starting values for the weighted least squares optimization. If the data are irregularly spaced, then the semivariogram can be computed by grouping observations that fall within a range of distances (Cressie 1993). A general rule of thumb is that the number of pairs of points within each range of distances should be at least thirty.

As mentioned previously, a nonconstant trend can interfere with the estimation of the variance of $\bar{D}$, causing the test to be either undersized or oversized. Diebold and Mariano (1995) did not need to estimate the trend of their loss differential series since all forecasts are for the same forecast horizon. The "horizon" concept does not exist for spatial data, so the trend can be a concern. When the pattern of the trend is known or suspected, then it can be estimated easily from the data, $D(\mathbf{s}_i)$, $i = 1, \ldots, L$, and then the data in Equations (4) and (5) must be replaced with the residuals, denoted $D^r(\mathbf{s}_i) = D(\mathbf{s}_i) - \hat{f}(\mathbf{s}_i)$, and $\bar{D}$ in those equations should be replaced with $\bar{D}^r = (1/L) \sum_{i=1}^{L} D^r(\mathbf{s}_i)$. We denote a parametric semivariogram estimated from detrended data by $\hat{\gamma}^r(h_{ij}|\hat{\boldsymbol{\theta}})$. Then, the test statistic S_V^r is the same as that in Equation (7) with $\hat{\gamma}$ replaced by $\hat{\gamma}^r$.

Of course, if the form of the trend is unknown, but it is likely that a trend exists, it can be estimated nonparametrically. We suggest using a bivariate Nadaraya–Watson estimator with Gaussian product kernel of the following form:

$$\hat{D}_b(x, y) = \frac{\sum_{i=1}^{L} K((x - x_i)/b) K((y - y_i)/b) D(x_i, y_i)}{\sum_{i=1}^{L} K((x - x_i)/b) K((y - y_i)/b)},$$

where b is the bandwidth. Selecting the optimal b when the data are dependent is not straightforward. Hart (1996) and Opsomer, Wang, and Yang (2001) discussed the difficulties and approaches used to select the optimal bandwidth for time series data. Francisco-Fernandez and Opsomer (2005) presented a method for selecting the optimal bandwidth for spatial data, but they used local linear regression and utilized a 2 × 2 matrix of bandwidths. The traditional bandwidth, b_0, selected by minimizing the cross-validation function,

$$\mathrm{CV}(b) = \frac{1}{L} \sum_{i=1}^{L} \bigl(D(\mathbf{s}_i) - \hat{D}_b^{(-i)}(\mathbf{s}_i)\bigr)^2, \tag{8}$$

where $\hat{D}_b^{(-i)}(\mathbf{s}_i)$ is the estimate of $D(\mathbf{s}_i)$ with the observation at location $\mathbf{s}_i$ removed, is too small when the data are positively spatially correlated. This leads to overfitting of the trend, removing too much variability from $D(\mathbf{s})$, an underestimate of the denominator of the spatial predictive comparison test statistic, and a too frequent rejection of the null hypothesis. The traditional bandwidth must be adjusted to account for the presence of spatial correlation. Similarly to the adjustment for time series data (Hart 1996), the adjustment for spatial data is

$$b_a = \left[\sum_{i=1}^{L} \sum_{j=1}^{L} C(h_{ij})/C(0) \right]^{1/5} b_0, \tag{9}$$

with the obvious circular problem of needing an estimate of the covariance structure to properly estimate the trend which is needed to properly estimate the covariance.

For a rough estimate of this adjusted bandwidth, we suggest an iterated procedure. Begin by substituting $C(h_{ij}) = 1$ if $i = j$

and $C(h_{ij}) = 0$ if $i \neq j$ into Equation (9) to get an initial adjusted bandwidth, $b_a^1 = (L)^{1/5} \cdot b_0$. Estimate the trend nonparametrically based on b_a^1, remove this trend from $D(\mathbf{s})$, estimate $C(h)$ using either WLS estimation of the empirical covariogram or semivariogram. Update the bandwidth and continue iterating until the bandwidth stabilizes. Use this stabilized bandwidth to estimate the trend, remove this trend from the data, and compute S_V^r.

4. MONTE CARLO SIMULATION STUDY

4.1 Data Simulation

To demonstrate the size properties of the test, we vary the grid size, the spatial correlation, the contemporaneous correlation, and the loss function. The basic simulation outline is to generate two sets of prediction errors in space, each with a certain spatial correlation and with a particular correlation to each other, apply the loss function, and then compute each test statistic and p-value. Prediction errors are generated instead of the original data, $Z(\mathbf{s}_i)$, $i = 1, \ldots, n$. If the simulation begins at the data level, then prediction models would need to be built in order to obtain prediction errors and thereby the loss differential. An infinite number of possible prediction models could be considered, and since predictions can be produced even from deterministic models, the prediction errors are the random quantity we consider.

First, a realization of a bivariate Gaussian random field on an $r \times c$ grid is drawn. To do so, the random field is generated using a linear model of coregionalization (LMC) (Gelfand et al. 2004). This model allows each set of prediction errors to have a unique spatial correlation and to be correlated to each other. Let σ_1^2 and σ_2^2 represent the variability of the first and second set of prediction errors, respectively, and both are set to 1 in the simulation. The contemporaneous correlation between the prediction errors is denoted by ρ. Then, for

$$\mathbf{T} = \mathbf{A}\mathbf{A}' = \begin{bmatrix} \sigma_1^2 & \rho \\ \rho & \sigma_2^2 \end{bmatrix},$$

the LMC model specifies the cross-covariance matrix between the two sets of prediction errors as

$$\begin{aligned}\mathbf{C}(h) &= \sum_{j=1}^{2} \exp(-3h/\theta_j)\mathbf{a}_j\mathbf{a}_j' \\ &= \begin{bmatrix} \sigma_1^2 \exp(-3h/\theta_1) & \rho\exp(-3h/\theta_1) \\ \rho\exp(-3h/\theta_1) & \sigma_2^2\exp(-3h/\theta_2) + (\frac{\rho^2}{\sigma_1^2})(\exp(-3h/\theta_1) - \exp(-3h/\theta_2)) \end{bmatrix},\end{aligned} \tag{10}$$

for $\mathbf{a}_j$ the jth column of $\mathbf{A}$. We then generate the bivariate random field from a Gaussian distribution with mean zero and the cross-covariance function given in Equation (10). For a given distance, h, the upper left entry in Equation (10) describes the spatial correlation of the first set of prediction errors; the lower right entry describes the spatial correlation of the second set of prediction errors; and the diagonal entries give the cross-correlation between the two sets of predictions. Note that the spatial range of the first set of prediction errors is θ_1, but the spatial range of the second set of errors depends upon θ_1 and ρ. Only when either $\rho = 0$ or $\theta_1 = \theta_2$ is the spatial range of the second set of prediction errors equal to θ_2.

We generate data on square grids of sizes 5 × 5, 8 × 8, 10 × 10, 16 × 16, and 20 × 20, such that data on two adjacent cells are one unit apart. Thus, as the grid increases in size, more points are added at greater distances, which corresponds to an increasing domain asymptotic framework. With a prediction fraction of $\phi = 0.40$, the number of randomly selected locations for each grid size is $L = 10$, 25, 40, 102, and 160. We consider values of the contemporaneous correlation parameter, ρ, to be 0, 0.5, or 0.9, where $\rho = 0$ indicates that the forecast errors are completely unrelated, and $\rho = 0.9$ represents forecast errors that share a large proportion of simultaneously large or small values. The spatial correlation parameters vary among $\theta_1 = \theta_2 = 3$; $\theta_1 = \theta_2 = 6$; and $\theta_1 = 3, \theta_2 = 9$, and these generally represent the practical range, or the distance at which the spatial correlation is 5% or less. As the grid size increases, estimating this range becomes better. The variance of each process is set to 1 by dividing each simulated set of prediction errors by the square root of $C(0) = \sigma_1^2 + \sigma_2^2 - 2\rho$, and the tests are performed at the $\alpha = 0.05$ level. Two loss functions are evaluated: the quadratic loss, $g_1(e(\mathbf{s})) = (e(\mathbf{s}))^2$, and the absolute loss, $g_2(e(\mathbf{s})) = |e(\mathbf{s})|$. Unless otherwise stated, for each combination of parameters, 2500 simulated datasets are generated.

4.2 Constant Trend

In this section, both the size and power properties of the spatial prediction comparison test are explored when $f(\mathbf{s}) = \mu$, for μ some constant. For reference, the simulated true variance of $\bar{D}$ for each combination of sample size, spatial and contemporaneous correlation in the quadratic and absolute loss functions is found through simulation of 20,000 datasets. The test statistic with this simulated or true variance is denoted $S_T = \bar{D}/\sqrt{\hat{\sigma}_{\bar{D}}^2}$. When the true simulated variance of $\bar{D}$ is used, the proper size of the test is attained for every sample size, contemporaneous correlation, and spatial correlation. This simply illustrates that if one can estimate the variance of $\bar{D}$ accurately regardless of sample size or distribution of $\bar{D}$, then the spatial prediction comparison test is correctly sized.

The results for the test statistic S_V are given in Table 1, and upon examination, several points become clear:

- The contemporaneous correlation appears to have little influence on the size of the test except when the spatial range of each set of prediction errors differs.
- The size of the test is strongly influenced by the strength of the spatial correlation. As the spatial range increases, the null hypothesis is rejected more often than it should be.
- Except for the smallest grid, when the spatial ranges of the errors differ, the size is larger than when the spatial correlation is the same for both sets of prediction errors.
- As the sample size increases, the size of the test improves.
- The size can also be influenced by the type of loss function that is used. The empirical size is slightly larger for the absolute loss.

Table 1. Empirical size of test for loss functions under WLS semivariogram estimate of variance of $\bar{D}$ for the spatial prediction comparison test, S_V. All tests are reported at the 5% level, and 2500 Monte Carlo replications are performed

			Quadratic loss			Absolute loss		
Grid	L	ρ	$\theta_{i,j}=3$	$\theta_{i,j}=6$	$\theta_{i,j}=3,9$	$\theta_{i,j}=3$	$\theta_{i,j}=6$	$\theta_{i,j}=3,9$
5	10	0.0	5.00	9.72	8.44	6.36	10.72	10.56
		0.5	5.12	8.00	9.08	6.52	9.72	10.72
		0.9	4.64	8.52	8.12	5.92	10.44	7.88
8	25	0.0	5.08	7.04	9.00	5.72	8.16	9.84
		0.5	4.32	6.56	8.04	5.24	7.04	8.96
		0.9	4.40	6.40	5.72	4.64	7.32	6.24
10	40	0.0	4.20	5.64	9.24	4.88	6.44	9.72
		0.5	4.36	6.00	8.00	4.60	7.12	7.92
		0.9	5.20	6.32	5.96	5.36	6.80	6.16
16	102	0.0	4.08	5.84	8.72	4.92	6.16	8.48
		0.5	4.92	5.52	8.32	4.92	6.4	7.88
		0.9	4.88	5.76	5.92	4.64	6.32	6.24
20	160	0.0	5.32	5.36	9.20	4.68	5.96	9.16
		0.5	4.88	5.20	9.20	4.96	5.68	8.24
		0.9	5.56	5.56	5.64	5.68	6.12	6.24

NOTE: Standard errors of values in the table are between 0.4% and 1.0%.

The effect of the quadratic loss on the spatial correlation can be explained theoretically. From the book by Cressie (1993), if $\mathbf{Z} = (X, Y)^T$ is a bivariate normal random vector with mean $\boldsymbol{\mu} = (0, 0)^T$ and covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix},$$

then $\mathrm{Corr}(X^2, Y^2) = \rho^2$. Thus, positive spatial correlation is reduced under the quadratic loss. A similar effect occurs for a general loss $g(\cdot)$ for which $g'(0) = 0$ and $g''(0) \neq 0$ (Cressie 1985b). The absolute loss is not twice differentiable, but the spatial correlation is also reduced by it.

The power of the test using the test statistic S_V is given in Figure 1 for all combinations of ρ, θ_1, and θ_2 in the 16×16 grid size. (Other grid sizes gave similar results.) The mean, μ, of $f(\mathbf{s})$ is allowed to vary from 0 to 7 in increments of 0.5. For a given value of μ, the power increases with an increase in sample size and from Figure 1, we see that:

- Power rapidly approaches 100% as the mean increases.
- The stronger the spatial correlation, the longer it takes the power to reach 100%.
- Contemporaneous correlation does not appear to have much effect on the power.

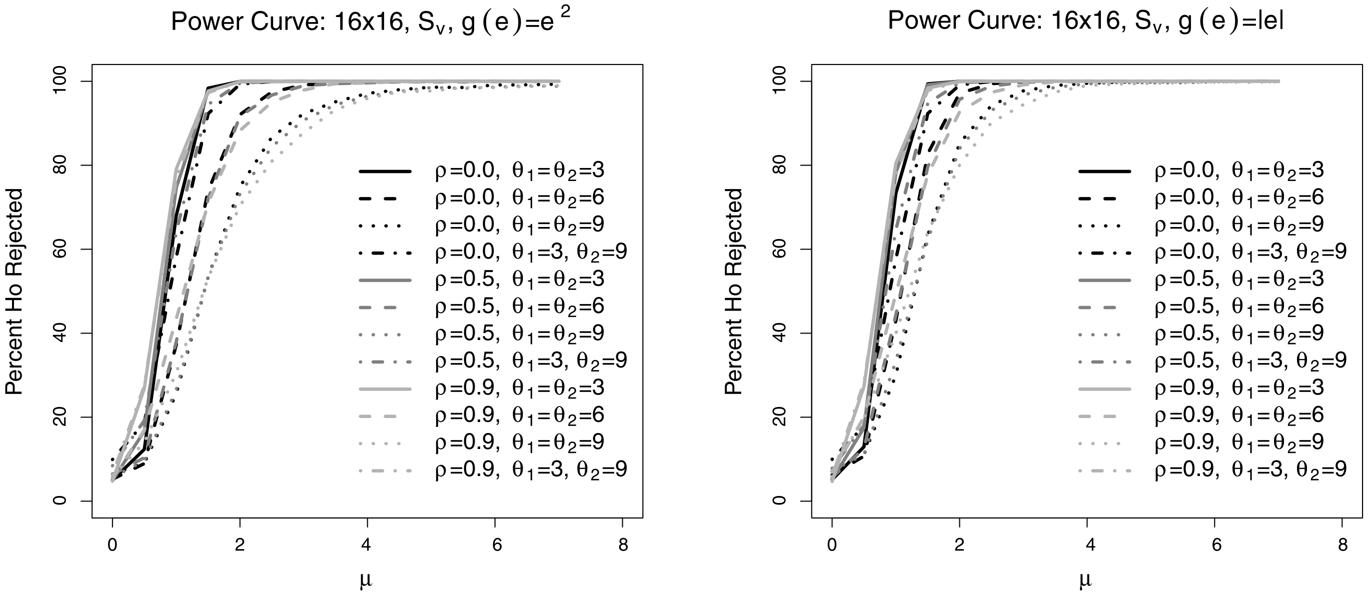

Figure 1. Power curves for the weighted semivariogram estimate of variance of $\bar{D}$ for the SPCT method on grids of size 16 × 16. The left panel is quadratic loss, and the right panel is absolute loss.

4.3 Spatially Varying Trend

In this section, results are given under both the null and possible alternatives when the mean function $f(\mathbf{s})$ is not assumed to be constant. Under the null hypothesis and with no information about the form of the trend, the trend can be estimated nonparametrically, where the bandwidth is adjusted to account for the spatial correlation in $\delta(\mathbf{s})$. When the bandwidth is not adjusted for the spatial correlation, the size of the test is more than double the sizes reported in Table 1 (results not shown). In other words, the traditional bandwidth, b_0, is used, and it is evident that the test is severely oversized. The selected bandwidth is too small, and the trend is overfit, making the test reject more often than it should. Table 2 displays the sizes using the test statistic S_V^r when the bandwidth is adjusted, even with a rudimentary iterative method. The size of the test does becomes worse as the spatial correlation increases, but it does better, even in small samples, when the spatial correlation is low, and the size improves as the sample size increases. Thus, avoiding trend estimation, if possible, is advisable. If the semivariogram for $D(\mathbf{s})$ rises and then reaches a constant sill as h grows, then the mean can be assumed to be constant, and trend estimation is not required.

Many types of spatially varying means for the alternative hypothesis could be imagined. For a 16×16 size grid with $\rho = 0.50$ and $\theta_1 = \theta_2 = 6$, two different types of trends for $f(\mathbf{s})$ will be examined, a split and a linear trend. The split mean function at location $\mathbf{s} = (x, y)$ is $f(\mathbf{s}) = v$ if $x \leq 8$ and $f(\mathbf{s}) = -v$ if $x > 8$. A nonparametric estimate may capture this trend effectively, but a linear fitted trend will be a poor fit. Then, a linear trend with mean function, $f(\mathbf{s}) = 20 + 0.5x - 0.5y$, is investigated. Of course, fitting a linear trend should work well for such a mean function, but the nonparametric estimate should also do well since the trend is a smooth function.

The results for the split trend are given in Table 3, and an additional loss function, the simple loss defined as $g(e_j(\mathbf{s}_i)) = e_j(\mathbf{s}_i)$, is also reported. The simple loss function illustrates how the prediction comparison test is only designed to detect a difference in two sets of spatial predictions on average and will not be successful in detecting local differences between two sets of predictions. Since the positive and negative values of $D(\mathbf{s})$ sum to zero in the split trend, the expected outcome for the simple loss for any value of v is the size of the test. Based on results not reported here, when $\theta_1 = \theta_2 = 6$, the empirical size of the test will be around 10%. However, in Table 3 when the true mean is removed, the size grows as v increases. The key to understanding this phenomenon lies in the random selection of locations to keep in the simulation. When 102 locations are selected out of the 256 grid locations, they are selected at random from across the entire grid. As v grows, the effect of not selecting an equal number of locations with positive and negative values on $\bar{D}$ grows. For example, if 40 locations are chosen with mean v and 62 are chosen with mean $-v$, then for large v, $\bar{D}$ will be far from zero. If 51 locations with mean v and 51 locations with mean $-v$ are selected instead, then the sizes remain around 10%. The dramatic result of failing to remove the trend or estimating the trend poorly with linear or iteratively reweighted generalized least squares (IRWGLS) trends (Schabenberger and Gotway 2005) results in a mean loss differential that is far more variable, and as a result the null is rarely rejected. The size with the nonparametric trend removed is the closest to what is expected under the null.

Under the quadratic and absolute loss functions in the split pattern, the alternative hypothesis is true. The spatial prediction comparison test does very well for these losses but can reject too little for $v = 1$. The linear trend rejects too infrequently for $v = 1$ and $v = 2$, and the IRWGLS approach performs similarly. Again, the nonparametric trend with adjusted bandwidth does the best job of filtering out the trend independently with similar results to those obtained when removing the true trend.

Table 2. Empirical size of test for loss functions under the weighted semivariogram estimate of variance of $\bar{D}$ for the spatial prediction comparison test with an iteratively adjusted bandwidth used in the nonparametric trend estimation, S_V^r. All tests are reported at the 5% level, and 2500 Monte Carlo replications are performed

			Quadratic loss			Absolute loss		
Grid	L	ρ	$\theta_{i,j} = 3$	$\theta_{i,j} = 6$	$\theta_{i,j} = 3, 9$	$\theta_{i,j} = 3$	$\theta_{i,j} = 6$	$\theta_{i,j} = 3, 9$
5	10	0.0	10.76	21.06	19.60	12.16	21.56	20.00
		0.5	10.96	18.64	18.32	12.00	20.60	19.68
		0.9	11.56	19.80	17.44	11.44	21.64	17.28
8	25	0.0	6.48	14.20	16.48	6.32	14.04	17.00
		0.5	7.12	15.56	15.24	7.36	15.24	14.80
		0.9	6.48	14.64	11.92	6.44	13.40	9.96
10	40	0.0	5.64	13.96	15.84	5.64	12.68	15.60
		0.5	6.04	14.16	14.40	5.72	13.48	14.04
		0.9	7.00	15.48	11.56	6.48	12.76	9.24
16	102	0.0	5.52	12.60	14.88	5.24	12.12	14.48
		0.5	5.84	12.36	13.60	5.92	11.08	12.68
		0.9	5.36	11.76	9.68	4.92	10.28	8.44
20	160	0.0	4.80	11.76	14.00	4.52	11.00	13.92
		0.5	5.08	13.04	12.64	5.60	12.20	11.36
		0.9	5.40	13.48	9.64	4.84	11.60	9.44

NOTE: Standard errors of values in the table are between 0.4% and 1.0%

Table 3. Percent of null hypotheses (under weighted semivariogram estimation) rejected in 2500 simulated datasets for the split trend datasets

		Intensity				
	Loss	$v=1$	$v=2$	$v=3$	$v=4$	$v=5$
True trend	Simple	9.52	12.20	16.72	23.20	25.28
	Quadratic	75.20	99.84	100.00	100.00	100.00
	Absolute	72.96	99.84	100.00	100.00	100.00
No trend	Simple	0.08	0.00	0.00	0.00	0.00
	Quadratic	42.68	95.60	99.64	99.92	100.00
	Absolute	48.12	97.52	99.92	100.00	100.00
Linear trend	Simple	6.84	1.20	0.40	0.04	0.00
	Quadratic	52.44	98.16	99.92	100.00	100.00
	Absolute	57.08	99.20	100.00	100.00	100.00
IRWGLS trend	Simple	5.40	0.60	0.28	0.00	0.00
	Quadratic	49.88	96.72	99.56	99.84	99.92
	Absolute	55.16	98.08	99.84	99.92	100.00
Nonparametric trend	Simple	21.24	13.44	14.60	16.60	17.64
	Quadratic	69.48	99.56	100.00	100.00	100.00
	Absolute	71.40	99.76	100.00	100.00	100.00

NOTE: Standard errors of values in the table are between 0.4% and 1.0%.

For the linear trend with results shown in Table 4, treating the true trend as if it is known and removing it gives very good results. However, ignoring the trend results in the rejection of almost none of the null hypotheses. This again illustrates how failure to remove a strong trend can negatively influence the test. After the quadratic and absolute losses are applied to the prediction errors, the trend is no longer linear, so a quadratic trend is fit for the quadratic and absolute losses. Fitting these types of trends yields results much closer to removal of the true mean. The IRWGLS fit yields nearly the same results. Finally, the nonparametric fitted trend does not fare as well as the linear and quadratic trends do.

5. OKLAHOMA WIND SPEED DATA

The Oklahoma Mesonet provides meteorological information at a network of over 100 stations across the state of Oklahoma and can be accessed at *http://www.mesonet.org*. The daily average wind speed is the quantity we wish to predict, but the daily averages of temperature, pressure, humidity, dew point, and rainfall are recorded as well. The latitude, longitude, and elevation of each site is given. While many years of data are available, the day we choose to focus on is September 10, 2008. The Oklahoma Gas and Electric utility serves the majority of the state and delivers electricity across an interconnected transmission and distribution system. Having one model which produces significantly better daily global average wind speed predictions would simplify and streamline their operations. Two non-nested spatial models are built based on 70 locations to predict the daily average wind speed at 46 reserved locations. Figure 2 gives a plot of these locations across the state. One time series model is also built based on three years of daily wind speed averages collected at each of the 46 sites.

Table 4. Percent of null hypotheses (under weighted semivariogram estimation) rejected in 2500 simulated datasets for the linear trend dataset with mean $f(\mathbf{s}) = 20 + 0.5x - 0.5y$

	Loss	
Trend removed	Quadratic	Absolute
True trend	100.00	100.00
No trend	0.08	0.28
Quadratic Trend	100.00	100.00
IRWGLS trend	99.96	100.00
Nonparametric trend	71.80	60.36

NOTE: Standard errors of values in the table are between 0.4% and 1.0%.

The first spatial model, called S1, uses the latitude, longitude, and elevation as covariates for the trend. This type of model might be used in a situation where the meteorological tower is off-line, and no other meteorological information is available. In the second spatial model, S2, the covariates of temperature, pressure, humidity, and dew point are included. For both models, the spatial dependence is modeled by an exponential covariance with a nugget. Parameters are estimated in both cases using the IRWGLS procedure. The preceding three years of daily average wind speed data before September 10, 2008 are used to build a time series model, T, at each of the 46 locations where a prediction is desired. At each location, a smoothed monthly mean and a smoothed monthly standard deviation are used to standardize the data. These smoothed values are obtained by regressing the monthly means and monthly standard deviations on a pair of harmonics. Then, the order, p, of an AR(p) model is selected with BIC.

Predictions are made at each of the 46 locations based on these three models. These predictions are compared to the observed daily average wind speeds using Mean Squared Error (MSE), Mean Absolute Error (MAE), correlation skill (COR), and Power Curve Error (PCE). The PCE was introduced by Hering and Genton (2010) as a more realistic assessment of wind speed predictions in the context of wind power generation. It incorporates information about the power curve that transforms wind speed to wind power and it allows the user to specify an asymmetric penalty for overestimation versus underestimation.

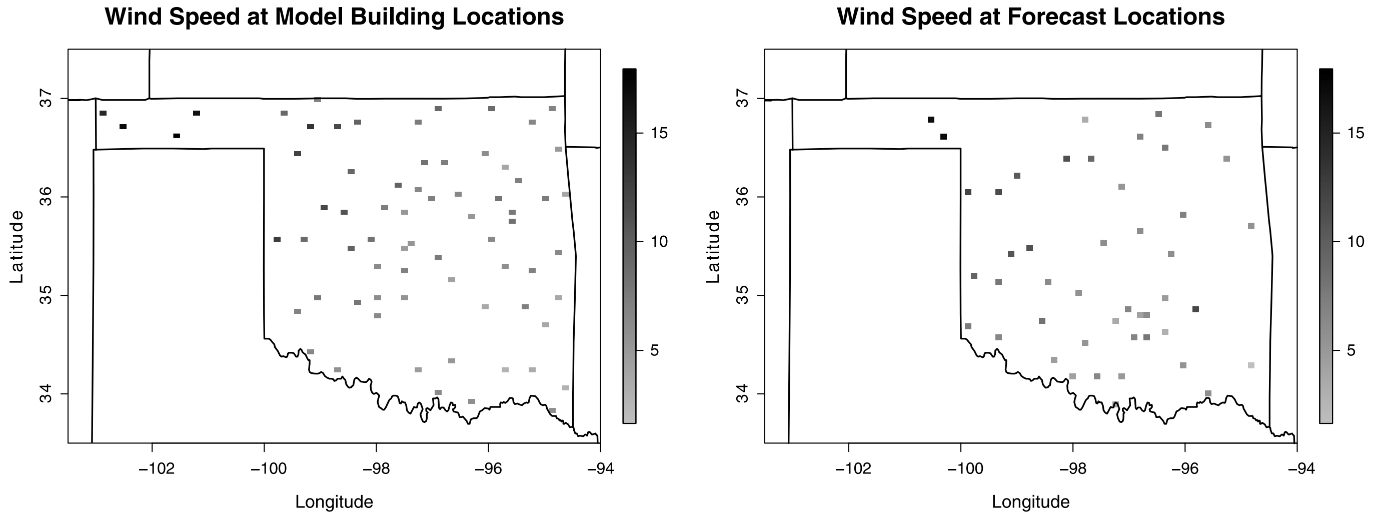
Wind Speed at Model Building Locations
Wind Speed at Forecast Locations
Latitude
Longitude
37
36
35
34
−102
−100
−98
−96
−94
15
10
5

Figure 2. Average daily wind speeds in miles per hour at 116 locations in Oklahoma on September 10, 2008. On the left are the 70 locations used to build the spatial models. On the right are the 46 locations where predictions are made.

Like the correlation skill, it is an example of a loss function that cannot be written in terms of the errors alone. Table 5 gives the values of MSE, MAE, COR, and PCE for each of the three models. The spatial model S2 produces predictions with the most favorable values of MSE, MAE, COR, and PCE, and the time series predictions have the least favorable values.

The top left plot in Figure 3 shows the loss differential of the power curve errors at each location comparing the time series predictions with the S2 predictions. The empirical semivariogram of the loss differential does not level off, which indicates that a trend is present in the data. With no knowledge of the trend, estimating the trend nonparametrically is likely the best option. The bandwidth is first selected by minimizing the cross-validation function, Equation (8); the top right plot shows where the curve is minimized. This initial bandwidth is then adjusted to account for spatial correlation, and the adjusted bandwidth is used to estimate the differenced field at a fine grid of points (bottom left plot). The biggest difference between the prediction errors of the time series and spatial models occurs in the northwestern region of the state, and the time series predictions only have smaller errors in isolated regions of the state. Finally, the empirical semivariogram of the detrended loss differential is computed, and a Gaussian semivariogram model is fitted (bottom right plot).

The estimates of the numerator and denominator of the spatial prediction comparison test, test statistics, and *p*-values are given in Table 6. The time series predictions and the predictions produced by model S1 are not significantly different from each other on average in terms of MSE, MAE, COR, or PCE. Even though S1 has MSE, MAE, COR, and PCE that are 1.28, 0.30, −5.8, and 24.02 less than the time series predictions, respectively, the variability in the errors is quite large. The S1 and S2 models also do not differ significantly from each other on average in either MSE, MAE, COR, or PCE. However, the S2 model does produce significantly different (and better) predictions on average in terms of MSE, MAE, and PCE than the time series model does. This would lead a researcher to conclude that when covariates such as average temperature, humidity, pressure, and dew point are available, they can produce on average a significantly superior prediction.

Table 5. MSE, MAE, COR, and PCE of each set of predictions for the Oklahoma wind speed dataset

Model	MSE	MAE	COR	PCE
T	5.29	1.77	73.4%	121.81
S1	4.01	1.47	79.2%	97.79
S2	2.51	1.34	87.8%	72.84

6. DISCUSSION

6.1 Local versus Global Methods

One advantage of local methods that the spatial prediction comparison test lacks is that they estimate the location and magnitude of a statistically significant spatial signal. Shen, Huang, and Cressie (2002) proposed a method called the Enhanced False Discovery Rate (EFDR), which is based on controlling the False Discovery Rate (FDR), for determining if there is a significant spatial signal at each location in the domain. Thus, for each location, there is a hypothesis to test, and to reduce the number of hypotheses that must be tested, the model is represented in the wavelet domain. Taking advantage of the "spatial" structure of the wavelet coefficients that is likely present under the alternative hypotheses, more power is gained by observing that the "large" wavelet coefficients of a pure signal typically cluster both within each scale and across different scales. This spatial structure allows researchers to predict whether a wavelet coefficient of the signal is 0 or not from observing its neighbors. This can be used to identify individual hypotheses that should be removed before applying the FDR procedure. We refer to this method as SHC, and note that the purpose of the Pavlicová, Santer, and Cressie (2008) work is the same, but a different type of wavelet coefficient thresholding is applied. Further details can be found in the original articles.

These methods are intended to be used on a complete grid of dyadic data, such as fMRI or climate model output data.

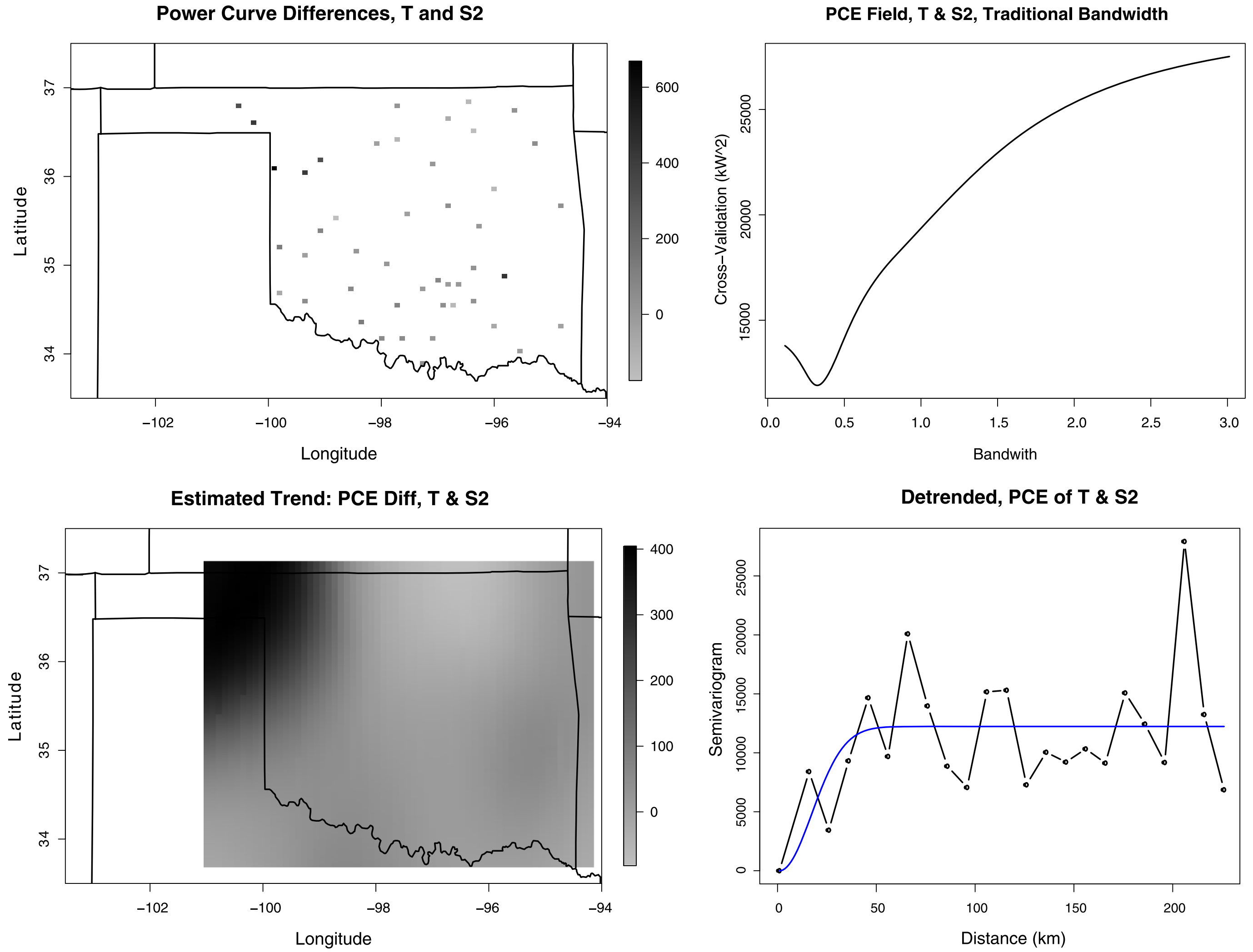

Figure 3. Plots of (upper left) power curve errors loss differential (time series errors minus S2 errors), (upper right) traditional bandwidth selection ignoring spatial correlation as given by the criteria in Equation (8), (lower left) reconstructed spatial field using nonparametric smooth and adjusted bandwidth, and (lower right) empirical semivariogram of detrended field with fitted Gaussian semivariogram overlaid. The online version of this figure is in color.

They were not introduced in the spatial prediction comparison context, but the loss differential can be thought of as the random field from which we wish to detect a spatial signal. Shen, Huang, and Cressie (2002) tested their method in simulations with normally distributed data and spatial independence. For comparative purposes, we present results from a small simula-

Table 6. SPCT comparison of the time series prediction, T, with two sets of spatial predictions, S1 and S2, for the Oklahoma wind speed dataset

Comparison	Loss	Numerator	Denominator	Test Statistic	*p*-value
T versus S1	MSE	1.28	1.21	1.05	0.2935
	MAE	0.29	0.16	1.89	0.0590
	COR	5.77	17.64	0.33	0.7435
	PCE	24.02	14.85	1.62	0.1057
T versus S2	MSE	2.77	0.38	7.41	<0.001
	MAE	0.43	0.13	3.19	0.0014
	COR	14.37	11.37	1.26	0.2063
	PCE	48.97	17.99	2.72	0.0065
S1 versus S2	MSE	1.50	0.86	1.75	0.0799
	MAE	0.13	0.21	0.66	0.5113
	COR	8.59	7.97	1.08	0.2809
	PCE	24.95	18.88	1.32	0.1864

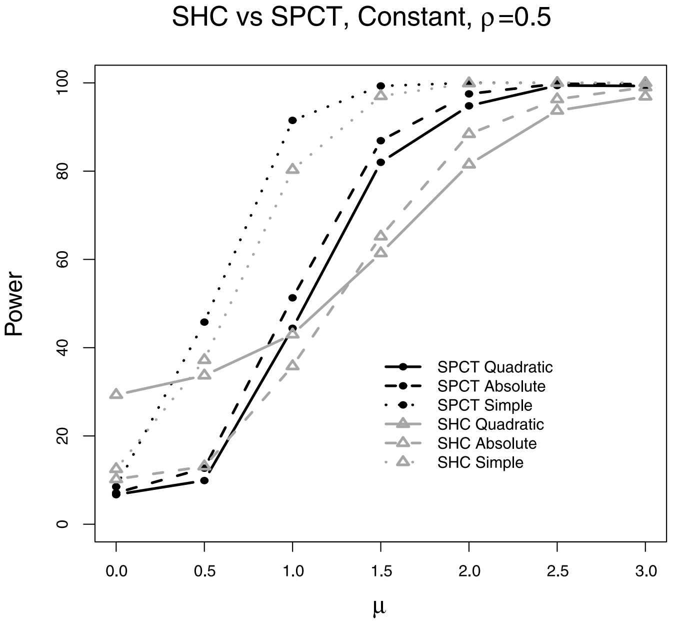

Figure 4. Comparing size and power of SHC and SPCT methods on a 16×16 grid with constant trend and $\theta_1 = \theta_2 = 6$.

tion experiment comparing SHC to the SPCT. The data that are generated in this simulation are multivariate normal, but once the quadratic and absolute loss functions are applied to the errors, the data are no longer normal. In fact, for interesting loss functions, the resulting loss differential will rarely be normally distributed.

When $f(\mathbf{s})$ is constant, then the SHC method and SPCT are testing the same null hypothesis that the mean of the spatial process is zero. To compare the two methods, we need to make adjustments to the SPCT setting so that the SHC method will work. Thus, 16×16 dyadic grids are generated in 1000 simulated datasets, and the full field of data is retained since the SHC method is only defined for data on a regular grid with no missing values. A constant mean alternative is generated in which the trend is $f(\mathbf{s}) = \mu$ for $\mu = 0, 0.5, 1, 1.5, 2, 2.5, 3$. The spatial range is $\theta_1 = \theta_2 = 6$, and both the SHC method and SPCT are applied to the quadratic, absolute, and simple loss differentials. The results are given in Figure 4. It is immediately evident that when $\mu = 0$, which is the null hypothesis, the SHC method is oversized for the quadratic loss. In fact, for the absolute and simple losses, the size is still about 5% too large. Thus, it is not sensible to compare the powers at the remaining values of μ since the SHC method is not correctly sized.

Local methods for signal detection such as those of Pavlicová, Santer, and Cressie (2008) and Shen, Huang, and Cressie (2002) require a user to be familiar with both wavelets and programming to implement these tests. The time to run the tests, approximately 5 minutes for a dataset of size 256 on a dual-core 2.67 GHz processor, is not prohibitive, but the SPCT can be implemented with standard geostatistical software and takes a few seconds to run. In addition, maps of the estimated trend, $\hat{f}(\mathbf{s})$, produced when detrending the data in the SPCT does give qualitative information about where significant differences may exist, but reducing the domain of interest to detect regional differences may be a better quantitative solution. Benjamini and Heller (2007) argued that in analyzing fMRI data differences in signals at the individual locations are not as important as detecting differences in clusters of voxels. This suggests that one solution to detecting regional differences in predictions could be to apply the SPCT in local regions of interest instead of across the entire set of predictions.

6.2 Outlook

Several versions of the spatial prediction comparison test have been proposed in this work. Test statistics under parametric estimation of the semivariogram for constant and spatially varying trends have been studied. When a spatially varying trend is present, estimating and removing this trend is crucial to maintain an accurately sized test. Overall, the spatial prediction comparison test is simple to compute, accounts for the presence of spatial correlation among the errors of a given loss function and for contemporaneous correlation, and allows flexible loss functions.

This work highlights promising directions for future research. Some are evident, such as a prediction comparison test for multivariate spatial predictions or for space–time forecasts. In fact, many of the most interesting examples involve predicting multiple variables over space, and the wind speed prediction example makes it clear that spatial forecasts made through time are necessary. In the space–time setting, it would be prudent to follow the example of Giacomini and White (2006) in which they proposed both conditional (for a given forecast horizon) and unconditional (averaged over all forecast horizons) tests of forecast accuracy. More generally, an improved and optimal method for selecting the bandwidth in the nonparametric estimate of the trend would have important applications beyond the spatial prediction comparison test; see the work of Bliznyuk et al. (2012). In summary, the SPCT used as a tool in model evaluation can help researchers determine if the difference they see in the average losses of two competing models is significant or not, which yields a more informed picture of the predictions.

APPENDIX

Proof of Proposition 1

Consider an $L \times L$ matrix, $\mathbf{S} = (s_{ij})$, where L is the number of locations in a spatial dataset whose (i,j)th entry is $s_{ij} = (D(\mathbf{s}_i) - \bar{D})(D(\mathbf{s}_j) - \bar{D})$ for $1 \le i, j \le L$:

$$\mathbf{S} = \begin{bmatrix} (D(\mathbf{s}_1) - \bar{D})(D(\mathbf{s}_1) - \bar{D}) & (D(\mathbf{s}_1) - \bar{D})(D(\mathbf{s}_2) - \bar{D}) & \dots & (D(\mathbf{s}_1) - \bar{D})(D(\mathbf{s}_L) - \bar{D}) \\ (D(\mathbf{s}_2) - \bar{D})(D(\mathbf{s}_1) - \bar{D}) & (D(\mathbf{s}_2) - \bar{D})(D(\mathbf{s}_2) - \bar{D}) & \dots & (D(\mathbf{s}_2) - \bar{D})(D(\mathbf{s}_L) - \bar{D}) \\ \vdots & \vdots & \ddots & \vdots \\ (D(\mathbf{s}_L) - \bar{D})(D(\mathbf{s}_1) - \bar{D}) & (D(\mathbf{s}_L) - \bar{D})(D(\mathbf{s}_2) - \bar{D}) & \dots & (D(\mathbf{s}_L) - \bar{D})(D(\mathbf{s}_L) - \bar{D}) \end{bmatrix}.$$

The sum of the elements in this matrix is zero. This can be seen since the sum of any row in the matrix is zero. Then, by expanding and collecting like terms in $\sum_{i=1}^{L}\sum_{j=1}^{L} \hat{C}(h_{ij})$, it is easy to show that it is equal to the sum of the elements in the matrix $\mathbf{S}$, $\sum_{\{i,j\}} s_{ij}$.

SUPPLEMENTRY MATERIALS

Table for DM/SPCT Comparison: This pdf file contains the table of simulation results that are referenced at the end of Section 2. (supp_table.pdf)

ACKNOWLEDGMENTS

This research was partially supported by NSF grants DMS-1007504, CMG-0621118, and Award No. KUS-C1-016-04, made by King Abdullah University of Science and Technology (KAUST). The authors also thank the editor, associate editor, and two anonymous reviewers whose constructive comments have greatly improved the presentation of the article.

[Received September 2010. Revised May 2011.]

REFERENCES

Atger, F. (2003), "Spatial and Interannual Variability of the Reliability of Ensemble-Based Probabilistic Forecasts: Consequences for Calibration," *Monthly Weather Review*, 131, 1509–1523. [414]

Benjamini, Y., and Heller, R. (2007), "False Discovery Rates for Spatial Signals," *Journal of the American Statistical Association*, 102, 1272–1281. [424]

Bliznyuk, N., Carroll, R. J., Genton, M. G., and Wang, Y. (2012), "Variogram Estimation in the Presence of Trend," *Statistics and Its Interface*, to appear. [424]

Brockwell, P. J., and Davis, R. A. (1991), *Time Series: Theory and Methods*, New York: Springer-Verlag. [416]

Cressie, N. (1985a), "Fitting Variogram Models by Weighted Least Squares," *Journal of the International Association for Mathematical Geology*, 17, 563–586. [417]

——— (1985b), "When Are Relative Variograms Useful in Geostatistics," *Journal of the International Association for Mathematical Geology*, 17, 693–702. [419]

——— (1993), *Statistics for Spatial Data*, New York: Wiley. [417,419]

Deckmyn, A., and Berre, L. (2005), "A Wavelet Approach to Representing Background Error Covariances in a Limited-Area Model," *Monthly Weather Review*, 133, 1279–1294. [415]

Dell'Aquila, R., and Ronchetti, E. (2004), "Robust Tests of Predictive Accuracy," *Metron*, 62, 161–184. [414]

Diebold, F. X., and Mariano, R. S. (1995), "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 13, 253–263. [414-417]

Ebert, E. E., and McBride, K. (2000), "Verification of Precipitation in Weather Systems: Determination of Systematic Errors," *Journal of Hydrology*, 239, 179–202. [415]

Francisco-Fernandez, M., and Opsomer, J. D. (2005), "Smoothing Parameter Selection Methods for Nonparametric Regression With Spatially Correlated Errors," *The Canadian Journal of Statistics*, 33, 279–295. [417]

Gelfand, A. E., Schmidt, A. M., Banerjee, S., and Sirmans, C. F. (2004), "Nonstationary Multivariate Process Modeling Through Spatially Varying Coregionalization," *Test*, 13, 263–312. [418]

Genton, M. G., and Hering, A. S. (2007), "Blowing in the Wind," *Significance*, 4, 11–14. [414]

Giacomini, R., and White, H. (2006), "Tests of Conditional Predictive Ability," *Econometrica*, 74, 1545–1578. [414,424]

Gilleland, E., Ahijevych, D. A., Brown, B. G., and Ebert, E. E. (2010), "Verifying Forecasts Spatially," *Bulletin of the American Meteorological Society*, 91, 1365–1373. [415]

Gneiting, T. (2011), "Quantiles as Optimal Point Predictors," *International Journal of Forecasting*, 27, 197–207. [415]

Gong, X., Barnston, A. G., and Ward, N. M. (2003), "The Effect of Spatial Aggregation on the Skill of Seasonal Precipitation Forecasts," *Journal of Climate*, 16, 3059–3071. [414,416]

Hart, J. D. (1996), "Some Automated Methods of Smoothing Time-Dependent Data," *Journal of Nonparametric Statistics*, 6, 115–142. [417]

Harvey, D., Leybourne, S., and Newbold, P. (1997), "Testing the Equality of Prediction Mean Squared Errors," *International Journal of Forecasting*, 13, 281–291. [414]

Hering, A. S., and Genton, M. G. (2010), "Powering up With Space–Time Wind Forecasting," *Journal of the American Statistical Association*, 105, 92–104. [414,421]

Kleiber, W., Raftery, A. E., Baars, J., Gneiting, T., Mass, C. F., and Grimit, E. (2011), "Locally Calibrated Probabilistic Temperature Forecasting Using Geostatistical Model Averaging and Local Bayesian Model," *Monthly Weather Review*, 139, 2630–2649. [414]

Longhi, S., and Nijkamp, P. (2007), "Forecasting Regional Labor Market Developments Under Spatial Autocorrelation," *International Regional Science Review*, 30, 100–119. [414]

Mardia, K. V., and Marshall, R. J. (1984), "Maximum Likelihood Estimation of Models for Residual Spatial Covariance in Spatial Regression," *Biometrika*, 71, 135–146. [417]

Matsuo, T., Nychka, D., and Paul, D. (2010), "Nonstationary Covariance Modeling for Incomplete Data: Monte Carlo EM Approach," *Computational Statistics and Data Analysis*, 55, 2059–2073. [415]

McCracken, M. W. (2004), "Parameter Estimation and Tests of Equal Forecast Accuracy Between Non-Nested Models," *International Journal of Forecasting*, 20, 503–514. [414]

Nychka, D., Wikle, C., and Royle, J. A. (2002), "Multiresolution Models for Nonstationary Spatial Covariance Functions," *Statistical Modelling*, 4, 315–331. [415]

Opsomer, J., Wang, Y., and Yang, Y. (2001), "Nonparametric Regression With Correlated Errors," *Statistical Science*, 16, 134–153. [417]

Park, B. U., Kim, T. Y., Park, J., and Hwang, S. Y. (2009), "Practically Applicable Central Limit Theorem for Spatial Statistics," *Mathematical Geosciences*, 16, 555–569. [416]

Pavlicová, M., Santer, T. J., and Cressie, N. (2008), "Detecting Signals in FMRI Data Using Powerful FDR Procedures," *Statistics and Its Interface*, 1, 23–32. [415,422,424]

Percival, D. B. (1993), "Three Curious Properties of the Sample Variance and Autocovariance for Stationary Processes With Unknown Mean," *The American Statistician*, 47, 274–276. [416]

Schabenberger, O., and Gotway, C. A. (2005), *Statistical Methods for Spatial Data Analysis*, Boca Raton, FL: Chapman & Hall. [416,417,420]

Sedur, L., Maxim, V., and Whitcher, B. (2005), "Multiple Hypothesis Mapping of Functional MRI Data in Orthogonal and Complex Wavelet Domains," *IEEE Transactions on Signal Processing*, 53, 3413–3426. [415]

Shen, X., Huang, H. C., and Cressie, N. (2002), "Nonparametric Hypothesis Testing for a Spatial Signal," *Journal of the American Statistical Association*, 97, 1122–1140. [415,422-424]

Shi, T., and Cressie, N. (2007), "Global Statistical Analysis of MISR Aerosol Data: A Massive Data Product From NASA's Terra Satellite," *Environmetrics*, 18, 665–680. [415]

Snell, S. E., Gopal, S., and Kaufmann, R. K. (2000), "Spatial Interpolation of Surface Air Temperatures Using Artificial Neural Networks: Evaluating Their Use for Downscaling GCMs," *Journal of Climate*, 13, 886–895. [414]

Wang, W., Anderson, B. T., Entekhabi, D., Huang, D., Su, Y., Kaufmann, R. K., Potter, C., and Myneni, R. B. (2007), "Intraseasonal Interactions Between Temperature and Vegetation Over the Boreal Forests," *Earth Interactions*, 11, 1–30. [414]

West, K. D. (1996), "Asymptotic Inference About Predictive Ability," *Econometrica*, 64, 1067–1084. [414,416]

Willis, H. L. (2002), *Spatial Electric Load Forecasting*, New York: Marcel-Dekker. [414]