View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Texas A&amp;M Repository

SPACE-TIME FORECASTING AND EVALUATION OF WIND SPEED WITH STATISTICAL TESTS FOR COMPARING ACCURACY OF SPATIAL PREDICTIONS

A Dissertation

by

AMANDA S. HERING

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2009

Major Subject: Statistics

SPACE-TIME FORECASTING AND EVALUATION OF WIND SPEED WITH STATISTICAL TESTS FOR COMPARING ACCURACY OF SPATIAL PREDICTIONS

A Dissertation

by

AMANDA S. HERING

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Marc G. Genton
Committee Members,	Kenneth Bowman
	Raymond J. Carroll
	Mikyoung Jun
	Thomas Wehrly
Head of Department,	Simon J. Sheather

August 2009

Major Subject: Statistics

ABSTRACT

Space-time Forecasting and Evaluation of Wind Speed with Statistical Tests for
Comparing Accuracy of Spatial Predictions.
(August 2009)

Amanda S. Hering, B.S., Baylor University;
M.S., Montana State University
Chair of Advisory Committee: Dr. Marc G. Genton

High-quality short-term forecasts of wind speed are vital to making wind power a more reliable energy source. Gneiting et al. (2006) have introduced a model for the average wind speed two hours ahead based on both spatial and temporal information. The forecasts produced by this model are accurate, and subject to accuracy, the predictive distribution is sharp, i.e., highly concentrated around its center. However, this model is split into nonunique regimes based on the wind direction at an off-site location. This work both generalizes and improves upon this model by treating wind direction as a circular variable and including it in the model. It is robust in many experiments, such as predicting at new locations. This is compared with the more common approach of modeling wind speeds and directions in the Cartesian space and use a skew-t distribution for the errors. The quality of the predictions from all of these models can be more realistically assessed with a loss measure that depends upon the power curve relating wind speed to power output. This proposed loss measure yields more insight into the true value of each model's predictions.

One method of evaluating time series forecasts, such as wind speed forecasts, is to test the null hypothesis of no difference in the accuracy of two competing sets of forecasts.

Diebold and Mariano (1995) proposed a test in this setting that has been extended and widely applied. It allows the researcher to specify a wide variety of loss functions, and the forecast errors can be non-Gaussian, nonzero mean, serially correlated, and contemporaneously correlated. In this work, a similar unconditional test of forecast accuracy for spatial data is proposed. The forecast errors are no longer potentially serially correlated but spatially correlated. Simulations will illustrate the properties of this test, and an example with daily average wind speeds measured at over 100 locations in Oklahoma will demonstrate its use. This test is compared with a wavelet-based method introduced by Shen et al. (2002) in which the presence of a spatial signal at each location in the dataset is tested.

For Alex, whose support never wavers

For Mom, whose prayers never cease

For Dad, whose example I had to follow

ACKNOWLEDGEMENTS

I would like to acknowledge the following groups and individuals:

Committe chair Marc G. Genton directly influenced the details of this research, and his insight has been invaluable.

Committee members Kenneth Bowman, Raymond J. Carroll, Mikyoung Jun, and Thomas Wehrly made comments that guided the direction and improved the quality of this research.

Tilmann Gneiting provided the code and data for the RSTD wind speed forecasting model.

Stel Walker of Oregon State University’s Energy Resources Research Laboratory and Bonneville Power Administration provided ten-minute data of wind speed and direction for Vansycle, Kennewick, and Goodnoe Hills locations.

Oklahoma Mesonet provided the daily meteorological information for the 116 stations in Oklahoma.

Michael Stein and anonymous reviewers made comments that greatly improved the content and presentation of the Chapter III material.

TABLE OF CONTENTS

Page

CHAPTER

CHAPTER Page

LIST OF TABLES

TABLE Page

TABLE Page

TABLE Page

LIST OF FIGURES

FIGURE Page

FIGURE Page

CHAPTER I

INTRODUCTION

Wind energy is a rapidly growing industry that is gaining worldwide public interest. Once a wind farm has been installed, this “green” energy produces no greenhouse gases and is renewable and inexpensive. However, the United States currently supplies less than one percent of its electricity needs with wind power. One of the obstacles to increasing this percentage is that wind power must be distributed to consumers as it is produced. No cost-effective storage system for wind power exists, and utility providers must constantly balance the supply and demand of electricity. While demand is relatively predictable, wind power is by its very nature a variable source. To plan for transmission and scheduling of electricity, for maintenance, and for trading, forecasts of wind power are necessary.

In Chapter II, the current state of wind power is described as well as the issues facing forecasters who seek to minimize the disruptions that utilities experience when they incorporate wind energy into their generation mix. With a forecast of wind speed, a forecast of wind power can be derived for various numbers and types of wind turbines, so most modelers focus on forecasting speed. Wind speed has some unique characteristics, some of which vary from one geographic location to another. But in general, utilizing wind data collected over time and spatially distributed around the location where forecasts are desired will improve forecasts.

Statistical models are just one type of model used to forecast wind speeds, but they are generally the best for one to four hour horizons since they can be made quickly. They

The format and style follow that of *Biometrics*.

also come with a built-in estimate of the variability of the forecast, giving utilities information beyond a simple point forecast. Making improvements in statistical models over the persistence forecast, in which the last observed value is the future forecast, gives utilities more incentive and more confidence in buying wind energy.

In Chapter III, two new models for wind speed forecasting are presented and are tested with wind data from the Pacific Northwest. A model developed by Gneiting, Larson, Westrick, Genton, and Aldrich (2006) that makes two-hour ahead forecasts of the hourly average wind speed at one of the three locations serves as the benchmark. Their model defines two sets of regimes based on the wind direction at one of the sites. Given the regime, the variables in the model change. These regimes are based on unique geographical features of this area, making it difficult to apply in other regions. The first model we propose eliminates the regimes and incorporates wind direction as a variable in the model, and the second model transforms speed and direction to Cartesian coordinates to forecast the wind vector with a bivariate regression with skew-t errors. Both models are regime-free, but the second also forecasts wind direction, which is also needed to obtain an accurate wind power forecast.

The forecasts produced by all three models are compared with a new loss function that reflects the nonlinear relationship between wind speed and wind power. This loss function does not impose any penalties on the forecasts when errors are made in the constant region of the power curve. It also allows underestimation and overestimation of wind speeds to be penalized based on the costs associated with such errors. In evaluating the forecasts with this loss function and various others, the regime-free models demonstrate that they are more flexible and lose no predictive ability. Various experiments, such as predicting at other locations, modeling ten-minute data instead of hourly data, and tuning the penalty for over and under estimation in the loss, demonstrate the robustness of our proposed models.

The modeling approaches in Chapter III suggest the need for a more general modeling strategy in which forecasts can also be made spatially. A statistical test developed by Diebold and Mariano (1995) for comparing the forecast accuracy of competing sets of time series forecasts is applied to each pair of wind speed forecasts in Chapter III, but no comparable type of test is available for spatial data. In Chapter IV, we develop a similar type of test for the null hypothesis that the difference between two sets of spatial forecasts is, on average, zero. This test accounts for contemporaneous correlation, spatial correlation, and non-Gaussianity in two sets of forecasts and allows a general loss function for comparison. Diebold and Mariano (1995) use a truncated sum of the empirical covariances to estimate the variance of their test statistic, but we show that summing the empirical covariances across all spatial lags yields zero. Thus, estimation of the variance of the test statistic is done parametrically with estimation of the semivariogram with weighted least squares producing the best results.

Diebold and Mariano (1995) also did not encounter a time-varying mean since their test is designed exclusively for k-step ahead forecasts. With spatial predictions made at varying lag distances from the nearest neighbor, a spatially varying mean can influence estimation of the test statistic. Simulations show the performance of the test under the null and alternative hypotheses when both a constant spatial mean and a spatially varying trend are present. With a spatially varying mean, the trend must be removed first, and a nonparametric trend estimation routine is proposed. Misspecification of the trend can result in an incorrectly sized test.

An existing test that can be compared with the one we propose is only applicable in a narrow range of circumstances. Shen, Huang, and Cressie (2002) developed a test for detecting a significant spatial signal at each location in the domain, not on average across all locations as our test does. They apply a discrete wavelet transform to complete data on a dyadic grid and then seek to reduce the number of hypotheses to test by exploiting the structure of the wavelet coefficients. When the spatial mean is constant, the two tests are equivalent, and simulations show that the test by Shen et al. (2002) is oversized when the data is not Gaussian.

For illustration, the spatial forecast accuracy test is applied to a set of daily average wind speeds observed at over one hundred locations in Oklahoma. Finally, a summary of all findings is provided in Chapter V.

CHAPTER II

STATISTICS IN WIND POWER*

Part of the answer to rising energy needs and costs may literally be blowing in the wind. In industrialized countries, flipping on a light switch or booting up a computer is practically an unconscious act, but our dependence on electricity permeates nearly every aspect of life. Among sustainable sources of electricity, only wind energy has the capacity and technology needed to compete in the open marketplace. In fact, the largest onshore wind farm in Europe has begun construction in Scotland, and the largest in the US is planned for southern California. The biggest offshore wind farm production in the world is slated for the Thames Estuary. But, the wind is intermittent. In this work, we explain how advanced statistical techniques will enable wind energy to be more efficiently incorporated into the electrical grid.

2.1 Wind Power Basics

Harnessing the power of wind to benefit humans is not a new concept. Historically, windmills have been used to pump water from wells or to grind grain for centuries. But fast-forwarding into the 21^{st} century, “windmills” are being used to generate electricity. Wind turbines, as they are now commonly called, are enormous structures, generally up to 80 meters tall, which is roughly the equivalent of a 26 story building. With blades up to 40 meters in length and costing up to $2.5 million to manufacture and install a single one (www.eia.doe.gov), the science behind effective wind turbine design has evolved rapidly over the last two decades. Within the wind turbine housing is a gearbox to increase the

* Reprinted with permission from “Blowing in the wind” by Genton, M. G. and Hering, A. S., 2007. *Significance*, **4**, 11-14, Copyright 2007 by Wiley-Blackwell Publishing.

rotational speed and a generator to convert the motion into electricity. A computer in the tower senses the wind direction, points the blades in the optimal direction, and shuts the turbine off in dangerously high winds.

So, can these supercharged wind turbines actually produce enough energy to make a significant contribution to meeting demand? Most modern turbines installed onshore are rated to produce between 1.5 and 1.8 megawatts (MW) of electricity each, which is enough to power 1,000 homes for an entire year (www.bwea.com). Depending on the size and number of turbines, clusters of them situated in windy locations can produce electricity for many thousands of homes. These clusters, as in Figure 1, are called wind farms. Construction of the largest onshore wind farm in Europe started in the fall of 2006 south of Glasgow, Scotland. The construction will take 3 years to complete and will consist of 140 turbines producing 322 MW of electricity, enough for about 200,000 homes. The largest wind farm in the US is planned for a region just north of Los Angeles in California and will produce over 1500 MW of power.

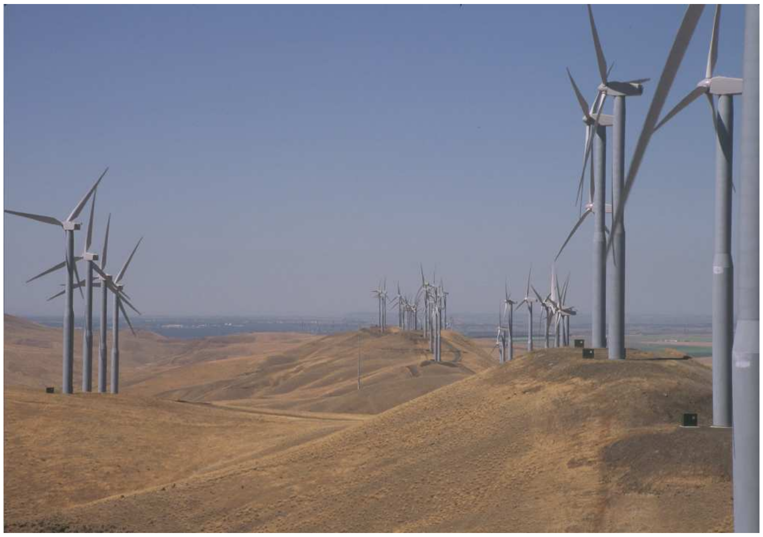


Figure 1: A typical wind farm in the state of Washington, USA.

Figure 2 illustrates the amount of power that can be produced by a typical onshore turbine at various wind speeds. At the cut-in speed, the blades begin to rotate, but the power output increases rapidly even with very small increases in the wind speed. In this range, power is proportional to the cube of wind speed, so small differences in speed can make large differences in power output. The maximum power output of 1.8 MW for this particular turbine occurs at about 30 miles per hour and shuts down at just over 50 miles per hour. However, power depends not only on wind speed but also on variables such as the diameter of the blades, the density of the air, and the direction from which the wind is blowing. Thus, wind power varies from one turbine make and model to another.

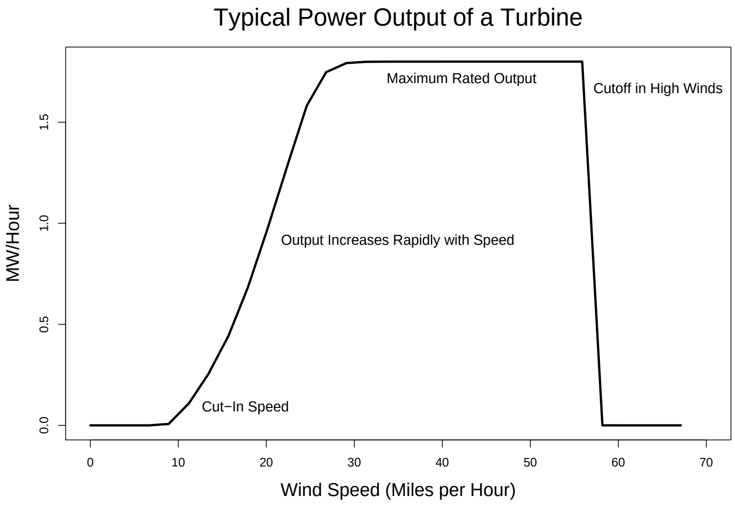


Figure 2: The potential power output of a wind turbine. The data is from a 1.8 MW Vesta V80-1800II turbine.

A large amount of growth and research is now being invested in offshore wind turbines, whose larger sizes (up to 3 MW with 5 and 7 MW machines in development) can take advantage of stronger ocean breezes. Just over 15 offshore wind farms are currently in operation, mainly off the coasts of Denmark, Sweden, and the UK, but many more are in the planning stages. The Thames Estuary scheme announced by the UK Secretary of State in December 2006 will use 341 turbines to generate a planned 1000 MW at a capital cost of £2 billion. Most offshore wind farms are located in water less than 30 meters deep, but engineers feel that they can draw on their experiences with oil platforms and move these farms even farther from land and out of public view.

Compared to traditional power plants fired by coal, natural gas, or nuclear reactions that produce, averaged over the year, 50% of their maximum designed output, wind farms produce on average about 30% of their maximum rated output. In the US, the current cost for a kilowatt hour of wind generated electricity is between \$0.04 and \$0.06, very similar to traditional energy sources which cost between \$0.04 and \$0.055 (www.eia.doe.gov). Opponents to wind energy claim that there are more start-up costs involved with wind energy. Transmission lines to move electricity from windy places, which tend to be remote, must be established, but once a wind farm is operable, it pays for itself in its first 6 to 8 months of operation (www.bwea.com). In addition, decommissioning a wind farm, whose turbines last 20 to 25 years, is simply a matter of disassembling the turbines, removing them, and recycling the materials. This is a much simpler and environmentally friendly process than decommissioning a nuclear power plant, for instance.

Wind farms have other tangible and intangible benefits. Once installed and operable, wind farms produce clean fuel, with no greenhouse gas pollutants or gas emissions. Quantifying the importance of this benefit is difficult but recognized as significant. The Energy Information Administration projects that oil and gas prices will remain high for at least the next 20 years (www.eia.doe.gov). Every hiccup in these prices can send economies into turmoil, so countries who invest in diversifying their energy portfolio, will help to stabilize their economies. Not only will demand for oil and gas decrease, thereby causing a decrease in prices, but more importantly, volatility in energy prices will be reduced.

Worldwide, only 1% of electricity is generated from wind, but the growth rate has been rapid—24% overall in 2005, with a stunning 48% increase in Asian markets. The

World Wind Energy Association expects that over 120,000 MW of wind power have been installed through the end of 2008 (www.wwindea.org). Many countries already boast a large proportion of wind generated electricity. The pioneering countries of Denmark and Germany who generate over 20% and 8% of their total electricity needs from wind, respectively, have set an example to others who plan to integrate wind electricity into their utility systems. Countries such as the US and the UK (both currently generating 1% of their electricity needs from wind) are aggressively developing their abundant wind resources. Figure 3 shows how much electricity of the worldwide total is produced by each of the top 6 countries.

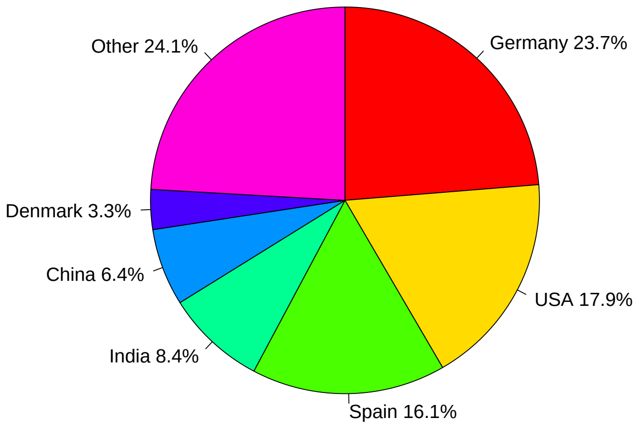


Figure 3: The percentage of worldwide wind capacity generated by each country in 2007 (www.wwindea.org).

With all of the advantages of and interest in wind produced electricity, barriers to widespread usage still exist. Indeed, utility companies must manage a delicate balance between electricity supply and demand. In larger markets, excess electricity can be sold, and deficits can be bought. But, depending on regulations in each particular market, monetary penalties can be imposed when energy is wasted. In smaller markets, such as those on islands, with no one else available to buy or sell electricity, there is little room for error.

Electricity demand by consumers varies in a nearly deterministic fashion based on outdoor temperature, daylight hours, and holidays. Thus, demand can be predicted, but when wind power is added as a source of electricity, supply becomes unpredictable (Giebel, Brownsword, and Kariniotakis, 2003). Wind power is intermittent—obviously the wind does not blow at a constant speed but is variable. No cost-effective solution to storing wind energy has been found, so wind energy must be used immediately when it enters the grid. A utility company consequently needs to schedule how much energy it needs to “order” from its traditional plants so that supply will equal demand. Gas turbine plants need at least 20 minutes notice to begin production, but large coal and oil plants require at least 8 hours to come online. Markets with slow-start production units would benefit the most from accurate wind power forecasts.

2.2 Statistical Solutions

Utilities cannot rely completely on wind energy because of its uncontrollable and intermittent nature. Given certain information, electricity dispatchers do not have to make blind decisions without any knowledge of how much electricity the wind will produce during a critical stage of decision-making. Statistical modeling to predict wind speeds or wind power can improve on our “best guess” estimate, which is the current wind speed, called the persistence model.

The number of hours ahead that a forecast is needed is called the forecast horizon and can vary depending on the reason for the prediction. The maximum horizon needed would be for 2 to 5 days ahead to schedule maintenance of the turbines during slow wind days. Otherwise, 24 and 48 hour forecasts are needed for trading in the electricity market. For scheduling and dispatch, a typical horizon is between 3 and 10 hours, but in systems whose conventional sources generate electricity quickly, the horizon can be under 3 hours.

Both physical and statistical models for predicting wind power have been proposed and are currently in use, but both approaches follow similar strategies. The available data that the models will be built with must be scaled to the hub height of the turbine. For instance, the wind speeds for the past 24 hours may be available at a height of 10 meters above ground level, but as the altitude increases, wind speed also increases in a logarithmic fashion. As a result, doubling the altitude can increase the wind speed by 10% and the power output by 34%.

The next step is deciding whether to predict the wind speed or jump straight to predicting wind power output, which is the bottom line for utilities. If wind speed is predicted, then an additional step of translating that into power output for the particular types and numbers of turbines in use must be done. However, solely predicting power for a particular region may make it difficult to predict power output for a nearby wind farm with different turbines. In statistical models used to date, it has been found that modeling the wind speed itself is most efficient for horizons up to 8 hours and then modeling the power output thereafter is sufficient (Giebel et al., 2003).

Finally, predictions can be upscaled for an entire region. This is especially important for areas like the UK and Europe where wind farms are geographically dense, and utility companies may manage several wind farms located in close proximity to each other.

Most physical models used to predict wind speed or power incorporate output from Numerical Weather Prediction (NWP) models. The basic premise of these models is the same–use a finer and finer grid of information to get a more complete picture of terrain and air flow. NWP based models can cover thousands of kilometers horizontally with grid resolutions from 5 to 25 km, but they are computationally extremely expensive to run. Models can require up to 4 hours of computer time and therefore cannot generate fast, reliable forecasts for short horizons (Gneiting, Larson, Westrick, Genton, and Aldrich, 2006). These short horizons are the typical time needed to schedule transmissions and dispatch. Thus, physical models are more effective for 24 hour predictions. Ensemble models, averaging many different physical models together or combining them with statistical models, are also becoming popular.

Statistical models are the most competitive for short forecast lead times. Neural networks, fuzzy logic, local regression, and time series methods have all been applied to the problem of wind speed prediction. Many of these models improve when additional information from the wind farm is included, such as wind direction, time of day, atmospheric pressure, and even physical model output (Gneiting et al., 2006). The best statistical models, however, do not use a “black box” approach but also incorporate expert knowledge of the wind characteristics of a particular region (Gneiting et al., 2006). It also makes sense that allowing parameters in these models to vary seasonally can result in improvements since a variable’s influence may change throughout the year.

A growing area of emphasis has been to incorporate off-site observations into statistical models (Gneiting et al., 2006; Larson and Westrick, 2006). Changes in wind speeds may be detected at upwind locations before reaching the wind farm and can improve predictions. An argument against this methodology is that sites “upwind” of a wind farm can change as the wind direction changes (Kretzschmar, Eckert, and Cattani, 2004), and no single off-site location may exist that has consistently high correlation with wind speeds at the prediction site. The ANEMOS Project group (a consortium in Europe whose goal is to improve wind forecasting) found that with information on 23 off-site locations, predictions could be improved using 3 to 5 of these sites whose meteorological conditions were most representative of the region (http://anemos.cma.fr). Even physical models have been shown to benefit from the use of additional spatial information.

With the plethora of models being proposed and tested, a consistent way to compare them is needed but not straightforward. Differences in complexity of the terrain, forecast and data resolution (10 minute, hourly, daily), and size and number of wind turbines at a farm can all affect model comparison. A common way to evaluate a model is to compute some function of the error–the actual observation minus what was forecast–such as the root mean square error (RMSE). RMSE will vary from one dataset to another; a skill score is used to remove the inherent variability in the observations. It is defined as the difference between the RMSE of a reference forecast and the RMSE of a model, divided by the RMSE of the reference forecast. The skill score can only be computed if a reference forecast (a model currently in place or the persistence forecast) is available. Even though RMSE is the most common measure to quantify error, it is not sensitive enough to reflect improvements in prediction quality. In addition, comparisons made only against the persistence model may be overly optimistic since improving upon the persistence forecast can be accomplished with the simplest of statistical techniques. The ANEMOS project group has also suggested that errors be normalized with respect to the installed capacity of the wind farm (http://anemos.cma.fr).

Besides the most obvious problem of forecasting the wind speed or wind power for a particular horizon, more detailed information about the quality of the forecasts is also desired. Statistical forecasts have a built in probabilistic error rate based on sampling distributions. These error bands around the predictions, or confidence bands, give dispatchers an idea of how certain the forecasts are. Very wide bands may indicate an unpredictable forecast, and smaller bands may indicate a more reliable estimate. Ensemble predictions can also give a sense of the forecast uncertainty (http://anemos.cma.fr). If the predictions from several different models are similar, then the collective prediction is more certain than if the forecasts vary dramatically. It is also of interest to identify conditions that lead to unpredictable power output or dramatic changes in power. When those conditions occur, utilities can protect themselves by carrying larger rolling reserves from traditional energy sources.

2.3 Future Work

Predicting wind speed and power is a blossoming area of research. Besides the issues previously mentioned, predictions at offshore wind farms add another dimension to the process. The vertical wind profile (and thus the relationship between wind speed at an observed height and the turbine hub height) differs offshore due to nonlinear interaction between the wind and waves, surface heating, and the land-sea interface that modifies the air flow. Understanding the wake effect behind massive offshore turbines will influence turbine orientation and spacing. A wake is the decrease in wind speed since some energy is lost after moving through the turbine blades. They differ from one turbine to another and can decrease power output by up to 10% (http://anemos.cma.fr).

Predictions both on and offshore may benefit from the use of more advanced statistical techniques. Many statistical methods are built on the assumption that the variable of interest is normally (symmetric and bell-shaped) distributed. This is decidedly untrue for wind speeds that are constrained to be positive and for which large values occur less frequently than small ones, as illustrated in Figure 4. Nonnormality should be incorporated into statistical models. In addition, placement criteria for new wind farms and for turbines within a wind farm can be evaluated and aided with the use of spatial statistics.

Improved statistical forecasting has already had an influence in increasing wind energy production. As the industry continues to expand, the end wants of utilities will only grow in number and complexity; they will need longer forecasts, more accurate forecasts, measures of forecast quality, and good tools for forecasting. As statisticians and scientists work together to provide these tools, the power in the wind will be harnessed and become a mainstream solution to energy demands.

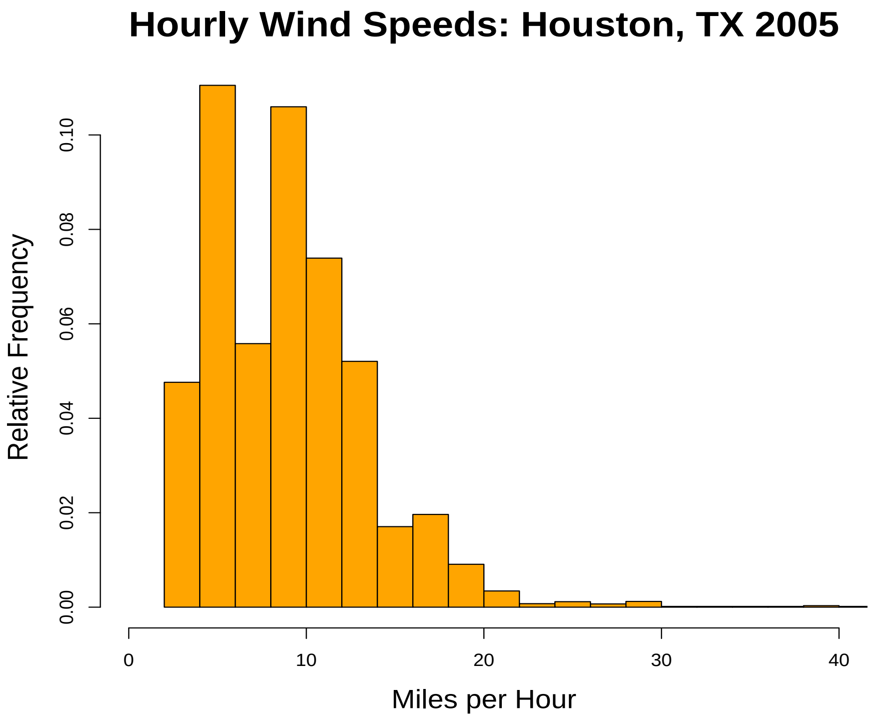


Figure 4: The wind speed recorded hourly during the year 2005 at Houston International Airport in Houston, Texas, USA.

CHAPTER III

SPACE-TIME WIND SPEED FORECASTING AND EVALUATION

3.1 Introduction

3.1.1 Wind Energy Background

The history of harnessing the power in wind for the benefit of man is long and diverse, yet wind energy's current role is evolving rapidly. Throughout the world, the number of installed megawatts increased in 2008 from 2007 by 29%. More facts and information on the role of statistics in wind power can be found in Genton and Hering (2007) and the references therein. Wind farms capable of powering many thousands of homes are springing up both on land and sea. Since the cost of a kilowatt (kW) of wind powered electricity is now nearly the same as a kW produced by coal or nuclear energy, many users are switching to this green energy that produces no greenhouse gases or harmful byproducts. Uneven heating of the earth's surface by the sun produces wind and guarantees that this natural resource will never be diminished or depleted.

Despite its many advantages, utilizing wind energy also presents its share of challenges. The windiest places tend to be the most remote, requiring transmission lines to carry electricity to populated areas. Some complain that the wind turbines ruin the scenery of pristine lands and interfere with bird migration. But by far, the biggest challenges are: (1) the wind is not a steady, constant supply of energy, and (2) no cost-effective method for storing its power currently exists. Its intermittent nature can create a problem for those managing the electrical grid, which is where the supply and demand of electricity meet and must be balanced. Electrical demand is easily predictable based on weather patterns, daylight hours, and holidays or work days. Usually, an equal amount of electricity is ordered to meet this demand from traditional sources. Wind-powered electricity must be used as soon as it enters the electrical grid, so the amount of additional electricity to order from traditional sources becomes unpredictable. Ordering too much or too little electricity can carry severe penalties and fines in utility markets.

Making accurate predictions of the future wind speed reduces the variability and risk that the electrical grid faces once it accepts wind energy as a source (Smith, Parsons, Acker, Milligan, Zavadil, Schuerger, and DeMeo, 2007). For a range of wind speeds, the amount of energy that can be produced from a wind turbine is proportional to the cube of wind speed, so small improvements in predicting wind speed lead to larger improvements in predicting wind energy. Predictions of wind energy could be made directly, but these are highly dependent upon the types, sizes, and number of wind turbines in operation. A prediction of wind speed, on the other hand, can be used to derive a prediction of wind energy for a given wind farm. The typical forecast horizon needed for scheduling transmission and dispatch is two to four hours. Longer horizons, such as two to three days, are useful for scheduling maintenance of the turbines, and numerical weather prediction models are best for this purpose.

Statistical models, especially those that incorporate expert knowledge of wind characteristics and geography, are unmatched in making short-term predictions (Giebel et al., 2003). However, this area of application has not been exhaustively explored by statisticians (Kestens and Teugels, 2002). Gneiting, Larson, Westrick, Genton, and Aldrich (2006) have recently proposed several models for predicting the two-hour ahead average wind speed near a wind farm in northern Oregon. Their best model, called the Regime-Switching Space-Time Diurnal (RSTD) model, accounts for the diurnal, non-negative, and volatile nature of wind speed. It takes advantage of the topography of the Columbia River Gorge in which winds are generally channeled in either an easterly or westerly direction to define two regimes. The regimes switch based on whether the wind direction at a point west of the wind farm is blowing from the west or from the east.

3.1.2 New Models and Evaluation Tools

In this work, two new models are introduced that eliminate the RSTD regimes, a loss measure to assess the quality of the predictions in terms of power is proposed, and experiments demonstrate the robustness of the new models. The two new models highlight differences in how the wind speed and direction variables may be treated—either in Polar coordinates or in Cartesian coordinates. In the first model, the Trigonometric Direction Diurnal (TDD) model, the wind direction is not simply used to determine the regimes. It is incorporated directly into the predictive mean function of the RSTD model by treating it as a circular variable and using its sine and cosine. Weisberg (2005) found that including the sine and cosine of wind direction did not improve wind speed prediction, but his model building approach is not systematic. The TDD model is more general than the RSTD model and has similar predictive ability.

The second model is called the Bivariate Skew-T (BST) model and uses the 2-dimensional Cartesian wind vector at different locations and lags in time to model the wind speed at the location of interest. The errors in this bivariate regression model are not distributed according to a normal distribution but with a skew-t distribution which is normal in a special case; see the review paper by (Azzalini, 2005). The skew-t distribution has additional parameters that are flexible for capturing skewness and heavy tails. Predictions of wind speed are ultimately for the purpose of predicting power; thus, assessing the quality of wind speed predictions should link speed and power (Lange, 2005; Lange and Focken, 2005). Typical measures such as Root Mean Squared Error (RMSE) or Mean Absolute Error (MAE) for gauging the quality of predictions do not make this link. Power curves describe the relationship between speed and power, and we develop a new loss measure that depends upon this curve. For various ranges of wind speeds, the power output is either constant or proportional to the cube of wind speed. Using a wind power curve for a standard turbine, penalties are assigned to each prediction in terms of power output. Finally, empirical evidence has shown that underestimating wind power averages a higher economic cost than overestimating it does (Pinson, Chevallier, and Kariniotakis, 2007). Therefore, the penalties are weighted based on the ratio between costs for over versus underproducing, and the effect of the weight on model performance is investigated.

The robustness of these new models is investigated in various experiments. The RSTD predictions are examined when another site besides the most westerly one is chosen to determine the regimes. In fact, choosing a different site with northerly/southerly regimes produces predictions that are as good as those produced with the RSTD, and choosing a poor set of regimes can deteriorate the predictions. This example illustrates that complex decisions involved in selecting regimes can impact the predictions. Each model is rebuilt to make predictions at other sites in the dataset, and the TDD model is found to perform significantly better than the RSTD model. Finally, the models are rebuilt on data observed at the ten-minute scale instead of data that have been aggregated to the hourly scale. These data are more variable, but the TDD model performs significantly better than the RSTD model.

This work is organized as follows. In Section 3.2, the RSTD, TDD, and BST models are described in detail. Section 3.3 introduces the power curve loss measure. Predictive performance of each model and robustness in several experiments are reported in Section 3.4. We conclude in Section 3.5.

3.2 Predictive Wind Models

3.2.1 Data Description

The data used in this study were collected at 3 meteorological towers near the Columbia River which runs along the Oregon-Washington border. The wind speed and direction were recorded every ten minutes. Vansycle, Oregon is located near the Stateline wind energy center and is the location where prediction is desired. Goodnoe Hills, Washington lies 146 km west of Vansycle, and Kennewick, Washington lies 39 km northwest of Vansycle. Figure 5 shows the approximate relative locations of the three stations. The time series of wind speed and direction are simultaneously recorded at all 3 locations for 55 days from September 4, 2002 to October 28, 2002 (used for training) and also for 279 days from February 25, 2003 to November 30, 2003 (used for testing). Wind speed and direction densities for the 2002 training data are in Figure 5. Each point on the circular histograms represents an observed wind direction. A point at the 0 angle indicates that the wind is blowing from the east to the west, a $\pi/2$ observation means the wind is blowing from the north toward the south and so on. For complete details on the dataset and site information, the reader is referred to Gneiting et al. (2006).

Many characteristics of the wind vector must be considered in building a model. Inherent in this dataset is spatial correlation. As weather systems move through the area, the site upwind of the others will be affected first, and the current wind conditions at that site will soon prevail at the other sites (Alexiadis, Dokopoulos, and Sahsamanoglou, 1999). Of course, which site is upwind of the others will change depending on the orientation of the weather system, but this can be addressed in the modeling. Strong temporal correlation is also present in the data with significant correlations in both the speed and direction lasting for over 24 hours. The wind speed and wind direction are also strongly linked. Martin, Cremades, and Santabárbara (1999) note the strong correlation between wind speed and direction but then ignore it and model the two variables separately. There is a diurnal pattern in the wind speeds, and seasonal differences do exist (Klink, 1999) but are more difficult to model with this limited amount of data. Finally, the wind speed variance varies in time as wind speeds change rapidly and with high frequency, which will be referred to as conditional heteroscedasticity.

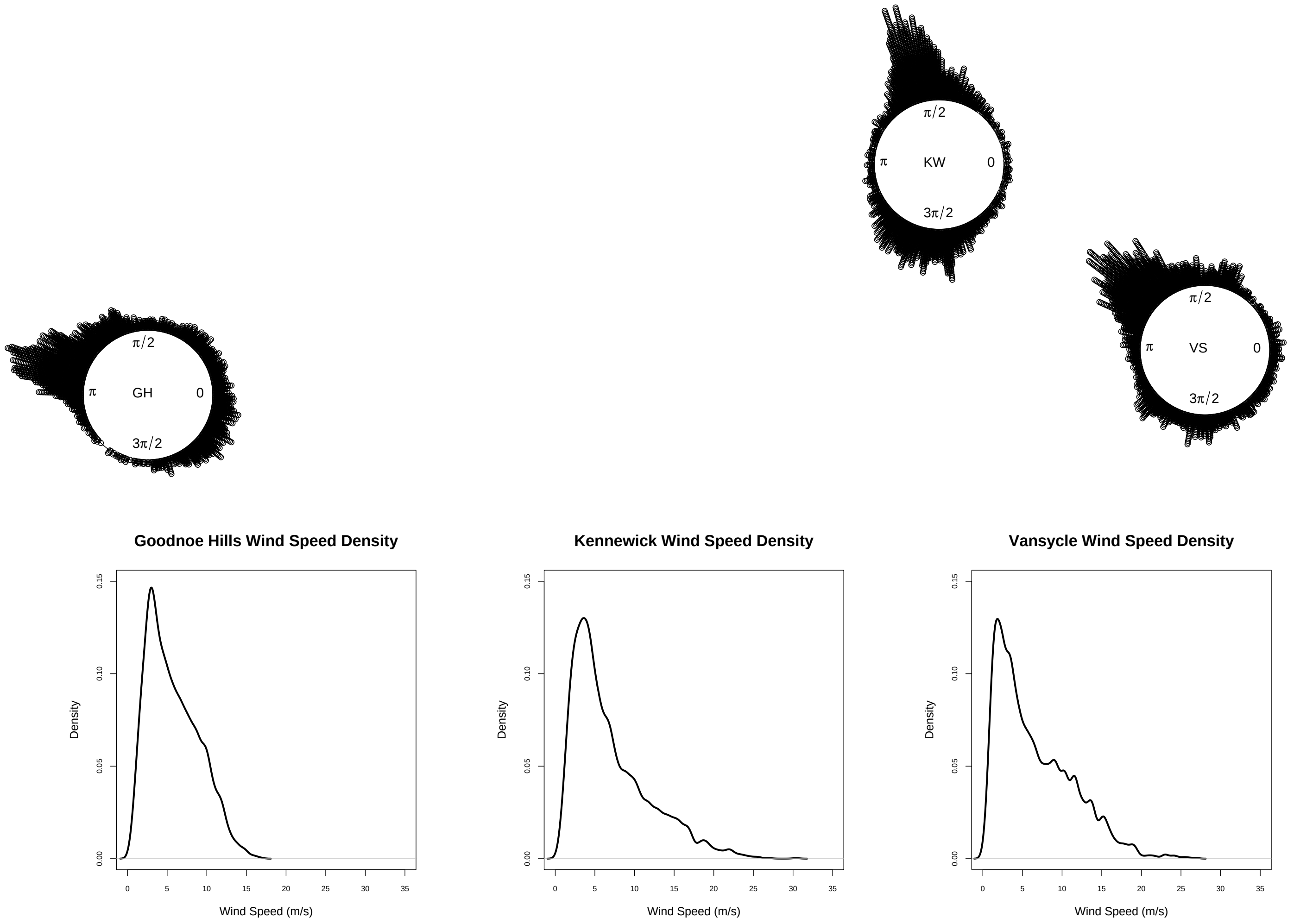


Figure 5: GH, KW, and VS denote Goodnoe Hills, Kennewick, and Vansycle, respectively. The locations of each circle indicate the relative location of each tower to the others. Each point on the circular histograms at the top represents a wind direction from the 2002 training data. For example, at Vansycle the majority of the wind directions blow from the northwest towards the southeast. The bottom panels are nonparametric density estimates of the 2002 wind speed data.

3.2.2 Regime-switching Space-time Diurnal Model

The best model that Gneiting et al. (2006) build incorporates many of the variable characteristics discussed in Section 3.2.1. This particular model will be presented briefly here for clarity, but the reader should see the original paper for the most complete description. In this model, the ten-minute observations of wind speed are averaged over each hour to yield a single hourly observation. The hourly wind speed at Vansycle is modeled with the truncated normal distribution, $N^+(\mu, \sigma^2)$, whose mean and α-quantile are given by

$$\mu^+ = \mu + \sigma \cdot \phi\left(\frac{\mu}{\sigma}\right) / \Phi\left(\frac{\mu}{\sigma}\right) \tag{3.1}$$

and

$$z_\alpha^+ = \mu + \sigma \cdot \Phi^{-1}\left[\alpha + (1-\alpha)\Phi\left(-\mu/\sigma\right)\right], \tag{3.2}$$

respectively, where ϕ and Φ denote the density and distribution function of a standard normal random variable. The key to the RSTD model is in choosing a structure for the predictive center, μ, and for σ, the predictive spread. The direction that the wind is blowing during the last ten-minute observation of the hour is used to switch the regimes. When the wind at Goodnoe Hills is blowing from the west to the east (i.e., the wind is westerly or in the westerly regime), the mean hourly wind speed at a particular location, D_s, is regressed on two pairs of harmonics as

$$D_s = d_0 + d_1 \sin\left(\frac{2\pi s}{24}\right) + d_2 \cos\left(\frac{2\pi s}{24}\right) + d_3 \sin\left(\frac{4\pi s}{24}\right) + d_4 \cos\left(\frac{4\pi s}{24}\right),$$

for $s = 1, 2, \ldots, 24$. Then the least squares fit from the wind speed series at each location is removed, resulting in residual series without a diurnal cycle. V_t^r, K_t^r, and G_t^r denote the residual series at time t for Vansycle, Kennewick, and Goodnoe Hills, respectively. Then, the predictive center is modeled by

$$\mu_{t+2} = D_{s+2} + \mu_{t+2}^r. \tag{3.3}$$

D_{s+2} is the fitted diurnal component at Vansycle, and μ^r_{t+2} is a linear combination of the present and past values of the residual series at the three sites

$$\mu^r_{t+2} = a_0 + a_1 V^r_t + a_2 V^r_{t-1} + a_3 K^r_t + a_4 K^r_{t-1} + a_5 G^r_t. \tag{3.4}$$

When the wind is easterly (blowing from the east to the west) at Goodnoe Hills, removing the diurnal variability from the wind speed series does not result in improvement, so the predictive center is modeled as

$$\mu_{t+2} = a_0 + a_1 V_t + a_2 K_t, \tag{3.5}$$

where V_t and K_t are the original time series.

For the westerly regime, the conditional heteroscedasticity is incorporated by modeling σ as a linear function of the volatility value with

$$\sigma_{t+2} = b_0 + b_1 v_t. \tag{3.6}$$

The coefficients b_0 and b_1 are constrained to be non-negative, and the volatility value, v_t, is

$$v_t = \left(\frac{1}{6} \sum_{i=0}^{1} \left((V^r_{t-i} - V^r_{t-i-1})^2 + (K^r_{t-i} - K^r_{t-i-1})^2 + (G^r_{t-i} - G^r_{t-i-1})^2 \right) \right)^{1/2}. \tag{3.7}$$

This reflects the magnitude of the most recent changes in the wind speed. In the easterly regime, the residual series in Equation (3.7) are replaced by the original wind series. The parameters in Equations (3.4), (3.5), and (3.6) are estimated numerically by minimizing the Continuous Ranked Probability Score (CRPS) for a truncated normal distribution (Gneiting and Raftery, 2007).

The 2002 data is used for building and developing the predictive mean structures in Equations (3.4) and (3.5), and the model is tested during the last 214 days of the 2003 series. A window of days in 2003 is used to estimate the parameters in the model before making the first prediction, and this window is rolled ahead by one observation after each two-hour prediction is made, the parameters are estimated again, and so on. Based on experiments performed by Gneiting et al. (2006), the window length that yields the best predictions is 45 days.

3.2.3 Trigonometric Direction Diurnal Model

Much of the structure of the RSTD model is retained in the TDD model, but Figure 6 clearly shows that the distribution of wind directions at Goodnoe Hills changes from the spring to fall months. It is less clear for months such as October and November if two regimes are sufficient. If not, it is even more difficult to determine how many regimes would be necessary and where the boundaries for these regimes would be. Instead of making a subjective decision about the number and position of the regimes, the TDD model eliminates the regimes but includes the wind direction, possibly at all three locations, as a covariate in the predictive mean function. Since wind direction is a circular variable, we include it in the model as the sine or cosine of the wind direction, following the suggestion by Mardia and Jupp (2000). We also use the hourly average of the ten-minute observations of wind direction instead of the last observed wind direction of each hour.

We build the predictive mean function from the pool of variables listed in Table 1 with the Bayesian Information Criterion, or BIC (Schwarz, 1978). Only lags up to three hours are shown since none greater are selected with this criterion. Using the 2002 data to build the model, the wind speed at Vansycle two hours ahead is regressed on the first variable, Vansycle's wind speed at the current time. If the BIC of this model is less than the model including only an intercept, then V_t is retained in the model. Then, V_{t-1} is added to the regression. If the BIC is reduced, then it is also added to the model. If BIC increases, then we do not include V_{t-1} in the model and skip the remaining lags of Vansycle wind speed. Next, both the sine and cosine of the current wind direction at Vansycle are added simultaneously to make the model invariant with respect to the axes, and they are

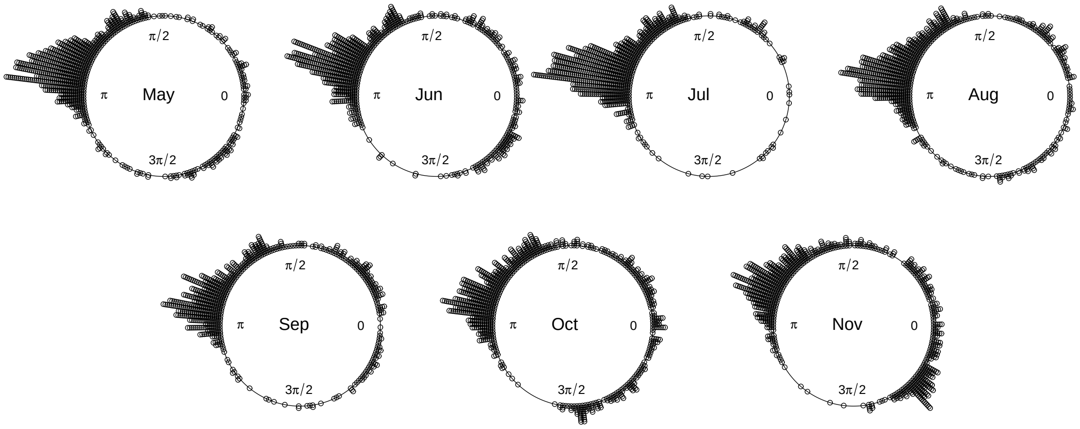


Figure 6: Circular histograms of wind directions at Goodnoe Hills for each month when predictions are made in the year 2003. Easterly winds are defined as those on the right-hand side of the circle between $3\pi/2$ and $\pi/2$. Westerly winds are on the left-hand side.

Table 1: This table contains correlations between the variables listed and the hourly wind speed two hours ahead at Vansycle. They are based on the 2002 training data and are used to build the TDD model. V, K, and G indicate the hourly wind speed at one of the three locations, and θ_V, θ_K, and θ_G represent the corresponding hourly wind direction for each location. Values in bold correspond to variables selected in the TDD model.

	Time Lag			
Variable	t	$t-1$	$t-2$	$t-3$
V	$\mathbf{0.90}$	$\mathbf{0.85}$	0.80	0.75
$\cos(\theta_V)$	$-\mathbf{0.55}$	-0.53	-0.51	-0.48
$\sin(\theta_V)$	$-\mathbf{0.21}$	-0.20	-0.18	-0.16
K	$\mathbf{0.74}$	$\mathbf{0.72}$	0.69	0.66
$\cos(\theta_K)$	$-\mathbf{0.63}$	-0.63	-0.62	-0.61
$\sin(\theta_K)$	$-\mathbf{0.02}$	-0.01	-0.00	0.01
G	$\mathbf{0.60}$	0.60	0.58	0.56
$\cos(\theta_G)$	$-\mathbf{0.33}$	-0.33	-0.34	-0.35
$\sin(\theta_G)$	$-\mathbf{0.45}$	-0.43	-0.42	-0.41

retained if their addition reduces the BIC. This process is repeated with the remaining variables in Table 1. The wind speed variables selected by this process are the same as the ones included in the RSTD westerly regime in Equation (3.4). In addition, several wind direction components are also included—both the sine and cosine of the current Vansycle wind direction, the sine and cosine of the current Kennewick wind direction, and the sine and cosine of the current Goodnoe Hills wind direction. We denote the wind direction at these locations and times as $\theta_{V,t}$, $\theta_{K,t}$, and $\theta_{G,t}$.

Removing the diurnal component of the wind speed was helpful in the RSTD model, and a strong diurnal component in wind direction is also detected. In Figure 7, the fitted values of a linear model regressing the hourly mean for speed and the hourly circular mean for direction (Fisher, 1993) on a pair of harmonics is plotted against the hour of the day for each location. If there were no diurnal trend, then the lines would be flat. All three locations show a clear cyclical pattern in the wind direction, so the fitted hourly mean direction is subtracted from each of the wind direction series. Thus, the predictive mean is modeled as in Equation (3.3), where D_{s+2} is still the fitted diurnal component of the wind speed at Vansycle, and

$$\begin{aligned}\mu^r_{t+2} \quad = \quad & a_0 + a_1 V^r_t + a_2 V^r_{t-1} + a_3 K^r_t + a_4 K^r_{t-1} + a_5 G^r_t + a_6 \sin(\theta^r_{V,t}) + a_7 \cos(\theta^r_{V,t}) \\ & + \; a_8 \sin(\theta^r_{K,t}) + a_9 \cos(\theta^r_{K,t}) + a_{10} \sin(\theta^r_{G,t}) + a_{11} \cos(\theta^r_{G,t}). \qquad (3.8)\end{aligned}$$

The scale of the truncated normal distribution is modeled as a linear function of the volatility value as in Equation (3.6).

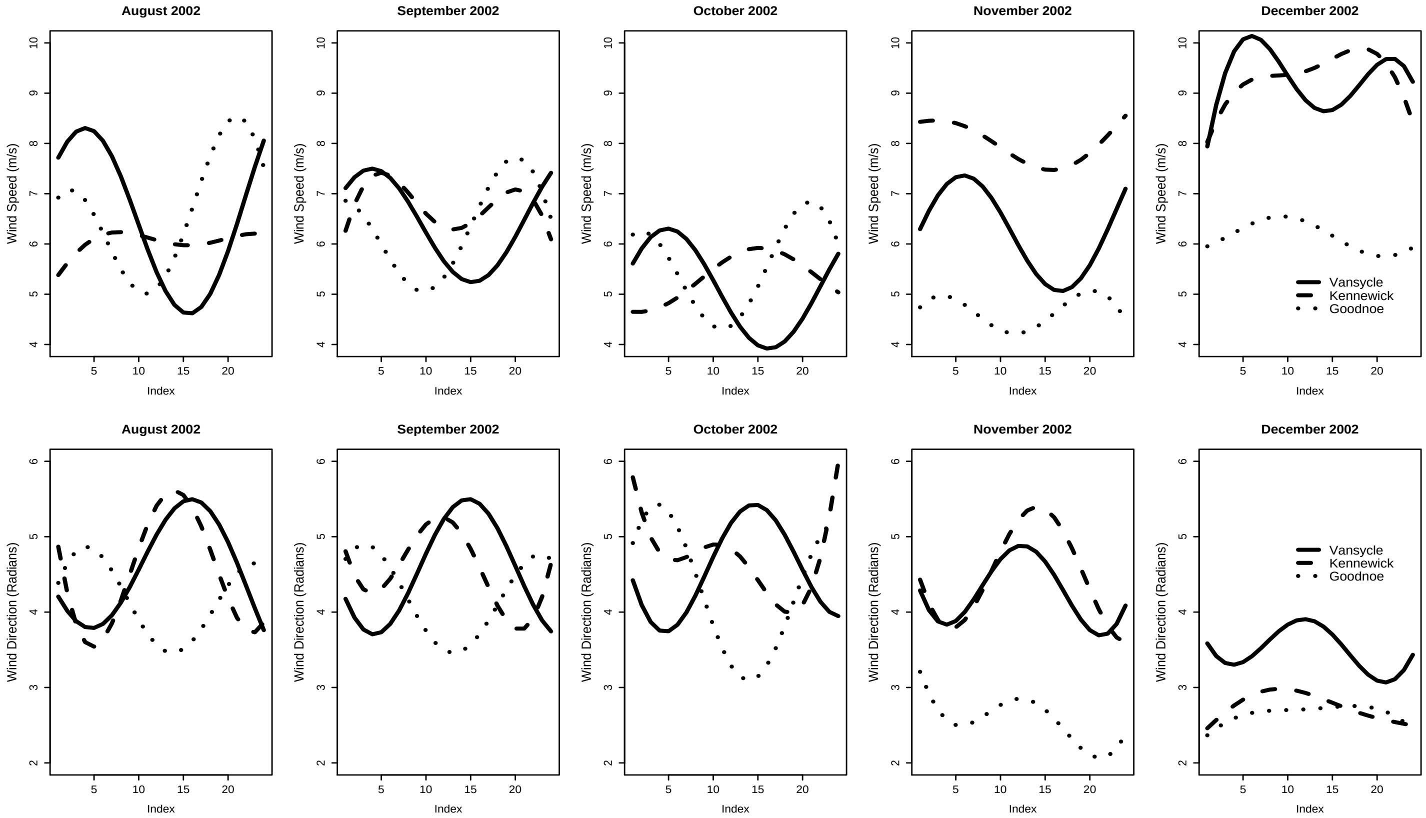


Figure 7: The top panels plot the fitted diurnal model of wind speed at each hour of the day at all 3 sites. It is clearly diurnal in nature for every month. The bottom panels plot the fitted diurnal model of wind direction at each hour of the day at all 3 sites. The diurnal nature of the directions is strong for every month except December which had 10 days of missing data.

3.2.4 Bivariate Skew-t Model

The BST model differs substantially from either the RSTD or TDD models. Instead of using hourly wind speed and hourly direction directly, these variables are converted into Cartesian components with $x = r\cos(\theta)$ and $y = r\sin(\theta)$ for r a wind speed and θ a wind direction. Let $\mathbf{V}_t = (V_{t,x}, V_{t,y})'$ denote the Cartesian components of the wind vector at Vansycle at time t. Here, $V_{t,x}$ is the east-west component, and $V_{t,y}$ is the north-south component. Let $\mathbf{K}_t$ and $\mathbf{G}_t$ denote similar vectors of values at Kennewick and Goodnoe Hills. The diurnal cycle is again removed from each component at each location by fitting a pair of harmonics to each set of hourly means, denoted by $D_{s,x}$ and $D_{s,y}$. Then, each component is standardized by dividing by an overall standard deviation computed at each location, denoted σ_x and σ_y (Brown, Katz, and Murphy, 1984). For example at Vansycle, the series is transformed by

$$\mathbf{V}_t^r = \left(V_{t,x}^r, V_{t,y}^r\right)' = \left(\frac{V_{t,x} - D_{s,x}}{\sigma_x}, \frac{V_{t,y} - D_{s,y}}{\sigma_y}\right)'.$$

These centered and standardized residual series will be denoted as $\mathbf{V}_t^r$, $\mathbf{K}_t^r$, and $\mathbf{G}_t^r$.

The residual series at time $t+2$ at Vansycle is modeled by

$$\mathbf{V}_{t+2}^r = \mathbf{A}_0 + \mathbf{A}_1\mathbf{V}_t^r + \mathbf{A}_2\mathbf{V}_{t-1}^r + \mathbf{A}_3\mathbf{K}_t^r + \mathbf{A}_4\mathbf{K}_{t-1}^r + \mathbf{A}_5\mathbf{G}_t^r + \boldsymbol{\epsilon}_t, \tag{3.9}$$

where $\mathbf{A}_0$ is a 2-dimensional vector of constants, $\mathbf{A}_i$ is a 2×2 matrix of coefficients for $i = 1, \ldots, 5$, and $\boldsymbol{\epsilon}_t$ follows a bivariate skew-t distribution. Then the random vector $\mathbf{V}_{t+2}^r$ follows a skew-t distribution whose location parameter is $\boldsymbol{\xi} = \mathbf{A}_0 + \mathbf{A}_1\mathbf{V}_t^r + \mathbf{A}_2\mathbf{V}_{t-1}^r + \mathbf{A}_3\mathbf{K}_t^r + \mathbf{A}_4\mathbf{K}_{t-1}^r + \mathbf{A}_5\mathbf{G}_t^r$, with scale matrix $\boldsymbol{\Omega}$, shape parameters $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)'$ to model skewness, and degrees of freedom ν to model kurtosis (Azzalini, 2005). In short, $\mathbf{V}_{t+2}^r \sim ST_2(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha}, \nu)$. The variables in the model in Equation (3.9) are selected using a BIC procedure similar to that used for the TDD model. The parameters are estimated using maximum likelihood estimation with the `R` package `sn` (Azzalini, 2006). Figure 8 shows that the skew-t distribution for the errors is a much better fit to the 2002 training data than the normal distribution for the errors is.

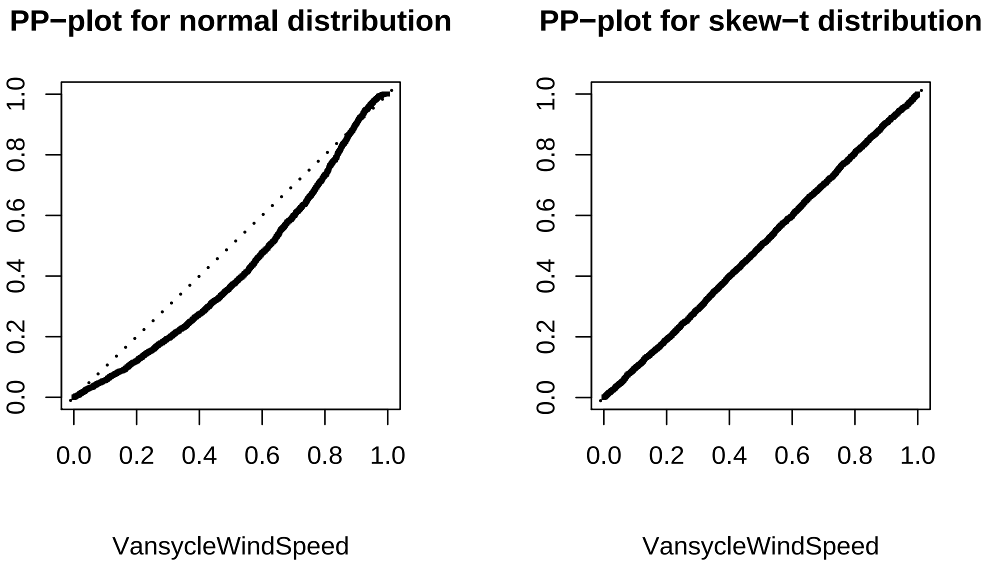


Figure 8: Comparison of the BST with normal errors (left plot) and with skew-t errors (right plot) on the 2002 training data.

Then, the predicted vector of Cartesian components at Vansycle two hours ahead is given by

$$\hat{\mathbf{V}}_{t+2} = \hat{\mathbf{\Sigma}}\hat{\mathbf{V}}^r_{t+2} + \mathbf{D}_{s+2},$$

where $t = 1, 2, 3, \ldots$ and $s = ((t-1) \mod 24) + 1$. The $\hat{\mathbf{A}}_i$, $i = 0, 1, \ldots, 5$, in $\hat{\mathbf{V}}^r_{t+2}$ are estimated from a 45 day window of data before the desired two-hour ahead prediction; $\hat{\mathbf{\Sigma}}$ is a matrix with the standard deviations of the x components and the y components estimated from the 45 day window on the diagonal and zeroes on the off-diagonal; and $\mathbf{D}_{s+2} = (D_{s+2,x}, D_{s+2,y})'$ is the fitted diurnal mean of the x and y components at Vansycle. Thus, the linear transformation of $\mathbf{V}^r_{t+2}$ gives $\mathbf{V}_{t+2}$ a $ST_2\left(\mathbf{\Sigma}\boldsymbol{\xi} + \mathbf{D}_{s+2}, \mathbf{\Sigma}\mathbf{\Omega}\mathbf{\Sigma}', \mathbf{\Sigma}^{-1}\boldsymbol{\alpha}, \nu\right)$ distribution (Azzalini and Capitanio, 2003). The predictive distribution of the wind speed requires taking the norm of $\mathbf{V}_{t+2}$, so the norm of 50,000 observations drawn from a skew-t

distribution with parameters estimated from each 45 day window of data is taken as the simulated predictive distribution. A large number of observations can be simulated quickly and easily and ensures that the behavior in the tails of the distribution is accurately characterized. The 45 day window length is chosen since it yields slightly better predictions than 30 and 60 day windows. This window length also makes the BST model easier to compare with the RSTD and TDD models, which also use 45 day windows.

3.3 Power Curve Loss Measure

Wind speed predictions from different models are commonly compared with RMSE and MAE, but these are not necessarily the appropriate loss functions in the wind forecasting paradigm. A better loss function should relate predicted wind speeds to the wind power since predicting power is the ultimate goal (Madsen, Pinson, Kariniotakis, Nielsen, and Nielson, 2005). The power depends on several factors, such as the air density ρ, the radius swept by the turbine blades r, and the wind speed v as follows

$$P = \tfrac{1}{2}\,\alpha\rho\pi r^2 v^3, \tag{3.10}$$

where α is an efficiency constant. As a baseline power curve, we use the GE 1.5 megawatt (MW) manufacturer's power curve (black dots in Figure 9) with fixed air density. The relationship between speed and power is not perfectly predictable, potentially even depending on the wind direction (Potter, Gil, and McCaa, 2007), but for practical purposes, we assume here that it is.

Four zones of the power curve are defined by the cut-in speed, the rated speed, and the cut-out speed. The cut-in speed is the speed at which the turbine blades begin to rotate. The rated speed is the lowest wind speed at which the maximum power output of the turbine is achieved. The cut-out speed is the speed at which the blades stop rotating to protect the turbine from damage. Zone 2 in Figure 9 is where the relationship in Equation (3.10) holds,

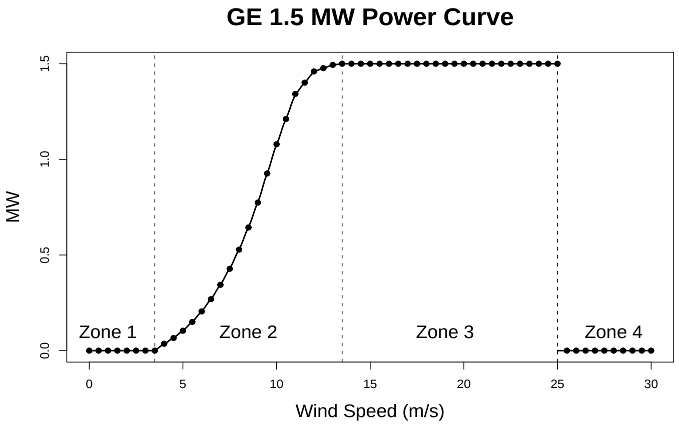


Figure 9: The GE 1.5 MW power curve. The black dots are the manufacturer's data. The solid curve in Zone 2 is a nonparametric fit to those data. It has a cut-in speed of 3.5 m/s, a rated speed of 13.5 m/s, and a cut-out speed of 25 m/s. These values change from one type of turbine to another.

and the solid curve in this region is a nonparametric Nadaraya-Watson type of estimate (Nadaraya, 1964; Watson, 1964) fitted with bandwidth $h = 0.025$. Small changes in the wind speed here can result in large differences in power output since power depends on the wind speed through a cubic function.

When both the observed and forecasted wind speeds are in Zone 1, 3, or 4, either no power output occurs or the maximum power output occurs. For example, if both the forecasted and the observed wind speeds are in Zone 3, then the power output is the same regardless of whether the wind speed forecast is close to the observed speed or not. No penalty would be assessed in terms of power for any differences in the observed and forecast speeds. When both the predicted and observed wind speeds are in Zone 2, small differences in forecasting wind speed will result in greater differences in forecasting wind power. As a result, discrepancies between the observed and forecasted wind speeds should receive greater penalties in this region.

We define $g(\cdot)$ to be the nondecreasing function that maps speed to power. The power curve is not a nondecreasing function, but only four of the 5136 wind speeds in the testing dataset are greater than the cut-out speed, so we ignore these cases. Precise power output data is not available since the power generated at the wind farm near Vansycle is proprietary information. Instead, an estimate of the true power is obtained with $g(V_{t+2})$ that will be compared to the forecasted power output based on the forecast wind speed, $g(\hat{V}_{t+2})$. Thus, a loss function that is of the Generalized Piecewise Linear form is defined to be

$$L(V_{t+2}, \hat{V}_{t+2}) = \begin{cases} \gamma(g(V_{t+2}) - g(\hat{V}_{t+2})), & \hat{V}_{t+2} \leq V_{t+2} \\ (1-\gamma)(g(\hat{V}_{t+2}) - g(V_{t+2})), & \hat{V}_{t+2} > V_{t+2} \end{cases}, \tag{3.11}$$

where γ is a weight between 0 and 1 and allows underestimates to be penalized differently than overestimates.

Empirical data from the Dutch electricity market in 2002 suggests that $\gamma = 0.73$, penalizing underestimates more strongly than overestimates (Pinson et al., 2007), which may at first seem counterintuitive. However when viewed from a holistic system perspective, an underestimate of wind power will cause the system operator to order too much electricity from traditional sources to meet the demand. In this case, the system operator now has a surplus of electricity, and down-regulation (when generation must be reduced) tends to be more expensive than up-regulation (when generation must be increased). The Power Curve Error, PCE, averages the penalties in Equation (3.11) over all forecasts and will be directly related to the energy produced by a wind farm (Madsen et al., 2005).

The optimal forecast that minimizes a particular loss function is given by

$$\hat{V}_{t+2} = \text{arg min}_{v_{t+2}} \; E_F \left[L(v_{t+2}, V_{t+2}) \right],$$

where F is the predictive distribution. In the simple cases where the loss function is squared error or absolute error, the optimal forecast is the mean or the median, respectively. For the error in Equation (3.11), the γth quantile minimizes PCE (Gneiting, 2008). Thus, the mean, median, and γth quantile of each model's predictive distribution will be used to compare the forecasts. The mean of the truncated normal distribution in Equation (3.1) and the median and γth quantile from Equation (3.2) are extracted from the RSTD and TDD models. The mean, median, and γth quantile are computed numerically from the simulated predictive distribution generated from the BST model.

3.4 Model Robustness

3.4.1 Comparing Model Performance on Testing Data

A simple baseline forecast is the persistence model. The persistence forecast for the average wind speed at Vansycle two hours ahead is simply the current wind speed at Vansycle. The mean of the predictive distributions of RSTD, TDD, and BST is used to compute the RMSE, the median is used in the MAE, and the γth quantile is used for the PCE. A measure called the continuous ranked probability score (CRPS), which essentially measures the spread of the predictive distribution subject to calibration is also computed. The CRPS can be computed explicitly for the predictive truncated normal distribution as given in Gneiting et al. (2006), and the CRPS value for the BST model is computed using the approximation in Equation (3) from Grimit, Gneiting, Berrocal, and Johnson (2006). Table 2 lists the results on the training data. The model with the lowest of each value in each column is bolded.

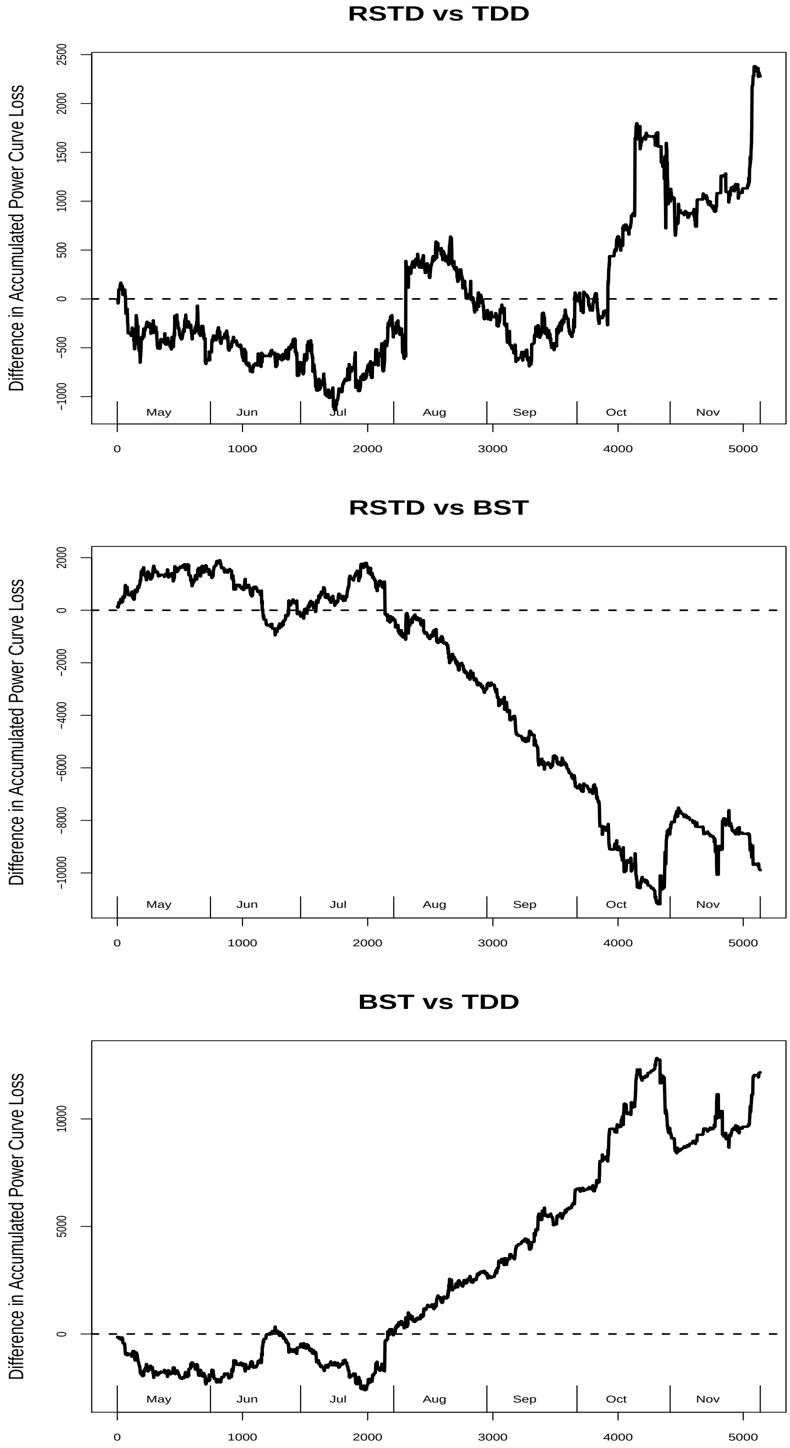
Table 2: Root mean squared error (RMSE), mean absolute error (MAE), power curve error (PCE), and continuous ranked probability score (CRPS) for 2-hour point forecasts of hourly average wind speed at Vansycle in May through November 2003, in m/s. CRPS is not given for the persistence model. The "Overall" column gives the measure over all forecasts from May through November.

Measure	Forecast	May	Jun	Jul	Aug	Sep	Oct	Nov	Overall
RMSE	Persistence	2.14	1.97	2.37	2.27	2.17	2.38	2.11	2.21
	RSTD	1.73	**1.56**	1.69	**1.78**	1.77	2.07	1.87	1.79
	TDD	1.74	**1.56**	1.68	**1.78**	**1.75**	**2.03**	**1.86**	**1.78**
	BST	**1.69**	1.59	**1.64**	1.81	1.85	2.09	2.00	1.82
MAE	Persistence	1.60	1.45	1.74	1.68	1.59	1.68	1.51	1.61
	RSTD	1.31	1.19	1.32	**1.31**	1.36	**1.48**	**1.38**	**1.34**
	TDD	1.34	**1.18**	1.31	1.33	**1.33**	**1.48**	**1.38**	**1.34**
	BST	**1.26**	1.19	**1.27**	1.37	1.42	1.51	1.50	1.36
PCE	Persistence	99.33	72.85	114.59	94.33	75.48	92.19	59.22	87.10
	RSTD	69.45	**48.19**	73.21	63.39	56.31	71.62	48.89	61.73
	TDD	70.17	48.42	**72.70**	**63.14**	**56.13**	**70.24**	**47.13**	**61.28**
	BST	**67.51**	50.46	73.42	66.90	61.57	73.83	50.98	63.65
CRPS	RSTD	0.95	**0.85**	0.94	**0.95**	0.97	1.08	**1.00**	**0.96**
	TDD	0.97	**0.85**	0.93	0.96	**0.95**	**1.07**	**1.00**	**0.96**
	BST	**0.92**	0.86	**0.91**	0.98	1.01	1.10	1.08	0.98

Overall, the TDD model has the smallest value or one of the smallest values for RMSE, MAE, PCE, and CRPS. It has the advantage over the RSTD model of being more general but retains the RSTD's predictive ability. In terms of PCE, the TDD model has the lowest values through the majority of the months and does better or best in terms of the other measures through the fall months. The BST model does not do as well as TDD and RSTD in any of the overall measures, but it does have the smallest RMSE, MAE, and CRPS in May and July and the smallest PCE in May. The BST model, like any robust fitting technique, fits to the majority of the data in the fitting window and is insensitive to unusually large or small values (Azzalini and Genton, 2008). Thus, its forecast for unusually high wind speeds tends to be poor. The wind speeds in May and July have the smallest standard deviations of any of the months, so the BST model does well during these months.

The differences among the models may seem small, but small differences are still important from a practical perspective. To test if these differences are significant, the large sample test introduced by Diebold and Mariano (1995) for comparing the forecast accuracy of competing models can be applied to check for significant differences between functions of the errors of two models. We test the null hypothesis that there is no significant difference between the overall MSE, MAE, or PCE of two models. With 5136 two-hour ahead hourly forecasts, the p-value to test for significant differences between the MSE of the RSTD and TDD models is 0.3337, and the p-value for the test of significant differences between their MAE's is 0.8713. Thus, we do not have evidence that the TDD model is significantly different in terms of squared or absolute errors. Both the TDD and RSTD models are significantly better than the BST model in terms of MSE and MAE. The p-value to test for a significant difference between the PCE of the TDD and RSTD models is 0.8457 and between the RSTD and BST models is 0.4375, neither of which is strongly significant.

A better sense of the difference between the two models in terms of wind power over the testing set is given in Figure 10. For each observation, the difference in accumulated



RSTD vs TDD

RSTD vs BST

BST vs TDD

Figure 10: These graphs plot the difference in accumulated PCE (in kW) penalties between the RSTD and TDD models (top), the RSTD and BST models (middle), and the BST and TDD models (bottom). An upward (downward) trend means that the second model is performing better (worse).

PCE penalties between the RSTD and the TDD models (top), between the RSTD and the BST models (middle), and between the BST and the TDD models (bottom) is plotted for all predictions made up until that observation. It should be noted that what is plotted is not PCE but the sum of the differences in penalties assigned by the PCE function for each prediction and has not been averaged. A similar graphical approach is taken in de Luna and Genton (2005) and serves to compare the cumulative forecasting ability of two models over a given time period and the gains or losses that would result. Based on this, the RSTD model makes steady improvements over the TDD model from May to the middle of July, from the middle of August to mid-September, and then for the first few days in November. However, the TDD model makes large gains in the beginning of August, middle of October, and end of November that leave it with a better accumulated PCE at the end of the testing period. When comparing the RSTD and TDD models with the BST model in the bottom two panels, except for the short periods in May and July, the RSTD and TDD models dominate the BST model in terms of PCE.

In all three models, the parameter estimates change with each new forecast, but to give a sense of their values, the averages over all forecasts for μ_{t+2} and σ_{t+2} in the RSTD model are 7.02 and 1.70, respectively. The average parameter estimates in the TDD model are 7.00 and 1.74, which are quite similar to the RSTD values. In the BST model, the average estimated the skewness parameter $\boldsymbol{\alpha}$ is $(-0.17, 0.01)'$, an indication that there is very little skewness in the distribution of the x and y components. The most interesting parameter in the BST model is the degrees of freedom, ν, which averages 5.26 and is always between 3.69 and 7.66, indicating that the distribution has very heavy tails.

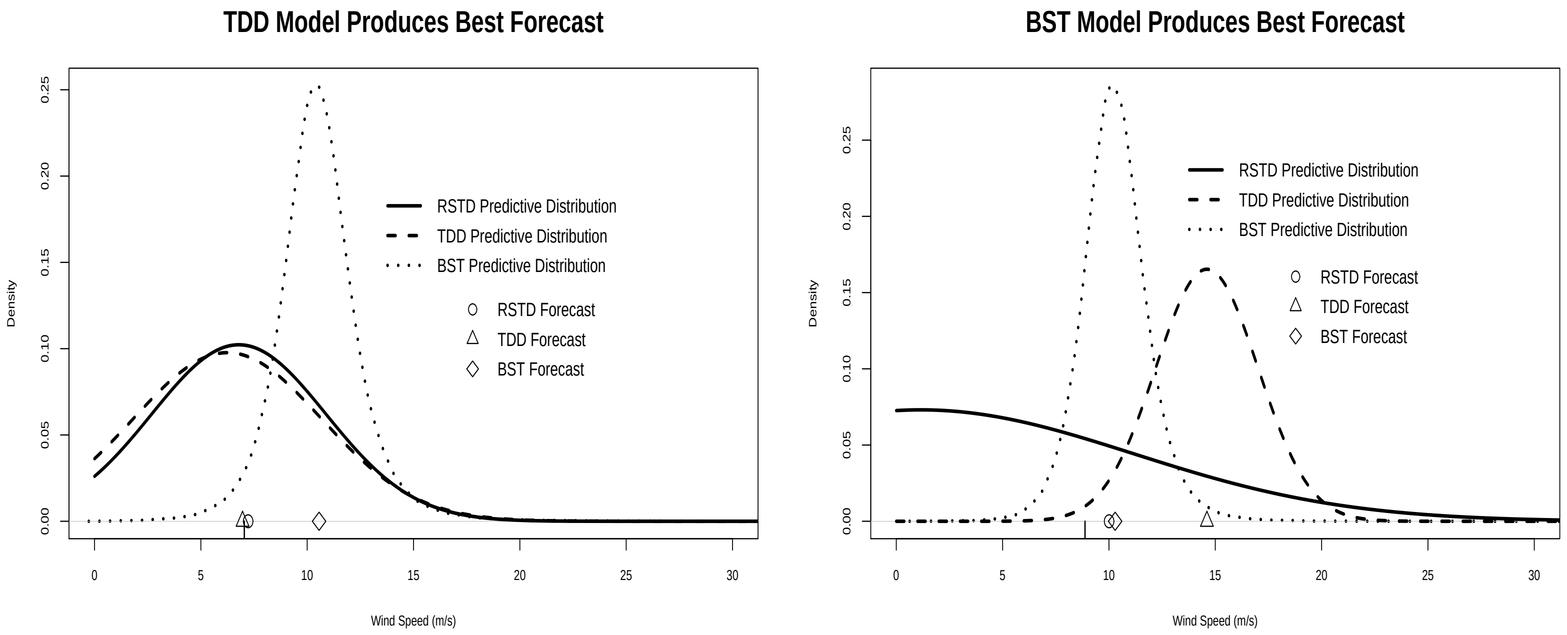


Figure 11: Comparing the predictive distributions for the models when the TDD model produces the best forecast (top panel) and when the BST model produces the best forecast (bottom panel). The small vertical line on the x-axis of each plot represents the observed wind speed.

The predictive distributions of the three models can look quite different, depending upon the forecast, as shown in Figure 11. The top panel shows the predictive distribution of all 3 models when the TDD model produces the best forecast. The RSTD distribution is very similar, but the BST model is centered incorrectly and is more concentrated. However, when the TDD model produces a poor forecast in the bottom panel of Figure 11, it also can be centered incorrectly. The RSTD model, in this case, produces a good forecast, but the predictive distribution is very widely spread. The BST model is not only centered closer the forecast, but it is also very tightly distributed. Over all forecasts, the 90% predictive intervals based on the upper 95% and lower 5% quantiles of these distributions have mean width 5.44, 5.52, and 5.96 for the RSTD, TDD, and BST models, respectively, with empirical coverages of 89.43%, 89.99%, and 91.59%. The TDD model has slightly wider intervals than the RSTD model and also slightly better empirical coverage. The BST has the widest intervals, and the coverage is a bit higher than the stated level.

3.4.2 Alternate Regime Selection

Some justification for using Goodnoe Hills as the site where the regimes are determined for RSTD is given in Gneiting et al. (2006), but Kennewick does not seem to have been considered as a potential site for the regimes to switch. We refit the RSTD model using Kennewick to determine the regimes. First, an easterly/westerly set of regimes is tested and then also a northerly/southerly set of regimes since Kennewick's main mode is nearer $\pi/2$ than it is to π, see Figure 5. The TDD and BST models do not need to be refit. Table 3 shows the results for the RSTD model for both the east/west regimes and the north/south regimes. The TDD and BST model results and the original RSTD model outcomes are also displayed for comparison.

First of note is that using an east/west set of regimes switching at Kennewick does deteriorate the RSTD predictions as compared to using Goodnoe Hills as the regime indicator. However, what is remarkable is that the north/south regimes at Kennewick can produce very good results, some values of RMSE, MAE, and PCE being smaller than those for the original RSTD model. The north/south regime is still not overall smaller than the TDD in PCE, but in three months it does produce the best PCE values. This illustrates the fact that unless all possible regimes and stations are tested, it may be impossible to empirically choose the site and regimes that yield the best predictions. If more stations with wind speed and direction data become available, this would only complicate the selection of a site at which to determine the regimes. In fact, the regimes may not depend on a single site only but on a possibly nonlinear combination of several sites. It seems reasonable to avoid such a selection when possible.

3.4.3 Predictions at Kennewick and Goodnoe Hills

To test the mobility of these models, the variables are reselected to make predictions at the other two locations in the dataset, Kennewick and Goodnoe Hills. When predicting at Kennewick and Goodnoe Hills, the best choice of regimes for the RSTD model may change, but the model is applied "blindly" in the sense that we want to see how portable it is to a new location. Variables are reselected for the RSTD predictive mean functions, but the easterly/westerly regimes that switch at Goodnoe Hills are held fixed.

Table 3: RSTD model outcomes when easterly/westerly and northerly/southerly regimes are defined by the wind direction at Kennewick. The original RSTD (with the regimes determined by the direction at Goodnoe Hills), the TDD, and the BST model results are also given.

Measure	Forecast	May	Jun	Jul	Aug	Sep	Oct	Nov	Overall
RMSE	RSTD-KW-EW	1.77	**1.56**	1.75	1.83	1.79	2.07	1.89	1.82
	RSTD-KW-NS	1.75	**1.56**	1.69	**1.77**	**1.74**	2.04	1.88	**1.78**
	RSTD-GH-EW	1.73	**1.56**	1.69	1.78	1.77	2.07	1.87	1.79
	TDD	1.74	**1.56**	1.68	1.78	1.75	**2.03**	**1.86**	**1.78**
	BST	**1.69**	1.59	**1.64**	1.81	1.85	2.09	2.00	1.82
MAE	RSTD-KW-EW	1.36	1.19	1.36	1.37	1.37	1.52	1.42	1.37
	RSTD-KW-NS	1.34	**1.18**	1.32	1.33	1.34	1.50	**1.38**	**1.34**
	RSTD-GH-EW	1.31	1.19	1.32	**1.31**	1.36	**1.48**	**1.38**	**1.34**
	TDD	1.34	**1.18**	1.31	1.33	**1.33**	**1.48**	**1.38**	**1.34**
	BST	**1.26**	1.19	**1.27**	1.37	1.42	1.51	1.50	1.36
PCE	RSTD-KW-EW	70.91	49.27	76.82	65.05	57.21	71.90	48.59	62.98
	RSTD-KW-NS	70.73	**47.30**	73.46	64.04	**54.76**	**68.90**	49.19	61.35
	RSTD-GH-EW	69.45	48.19	73.21	63.39	56.31	71.62	48.89	61.73
	TDD	70.17	48.42	**72.70**	**63.14**	56.13	70.24	**47.13**	**61.28**
	BST	**67.51**	50.46	73.42	66.90	61.57	73.83	50.98	63.65

Table 4: RSTD, TDD, and BST model outcomes for predictions made at Kennewick.

Kennewick	Forecast	May	Jun	Jul	Aug	Sep	Oct	Nov	Overall
RMSE	RSTD	2.34	1.96	2.09	2.17	2.13	2.36	2.34	2.21
	TDD	**2.32**	**1.94**	**2.08**	**2.15**	2.11	2.36	2.30	**2.19**
	BST	2.37	2.03	2.18	2.23	**2.05**	**2.28**	**2.23**	2.20
MAE	RSTD	1.82	1.44	1.60	**1.58**	1.60	1.77	1.66	1.64
	TDD	**1.79**	**1.43**	**1.59**	1.60	1.59	1.76	1.63	1.63
	BST	1.80	1.45	1.64	1.61	**1.51**	**1.72**	**1.54**	**1.61**
PCE	RSTD	87.45	65.18	**82.96**	83.78	67.53	74.52	78.60	77.24
	TDD	**85.63**	**64.93**	83.81	**83.19**	66.91	**71.31**	80.51	**76.69**
	BST	92.35	70.32	84.49	86.58	**66.47**	72.34	**73.91**	78.18

Table 5: RSTD, TDD, and BST model outcomes for predictions made at Goodnoe Hills.

Goodnoe Hills	Forecast	May	Jun	Jul	Aug	Sep	Oct	Nov	Overall
RMSE	RSTD	**1.69**	**1.51**	**1.38**	**1.55**	**1.68**	**1.87**	1.75	**1.64**
	TDD	**1.69**	1.55	1.40	**1.55**	**1.68**	**1.87**	**1.73**	1.65
	BST	1.76	1.64	1.43	1.56	1.70	1.98	1.78	1.70
MAE	RSTD	**1.31**	**1.16**	**1.06**	**1.18**	**1.25**	**1.37**	1.31	**1.23**
	TDD	**1.31**	1.19	1.08	1.20	1.26	**1.37**	**1.28**	1.24
	BST	1.38	1.29	1.09	1.19	1.27	1.45	1.34	1.28
PCE	RSTD	**81.69**	**61.18**	**67.36**	68.96	63.66	70.83	**56.78**	**67.30**
	TDD	82.54	64.33	68.52	69.17	64.67	**69.31**	56.90	68.01
	BST	86.46	67.47	68.99	**68.30**	**63.24**	76.11	61.13	70.33

The results in Tables 4 and 5 show that the TDD and BST models have smaller summary measures than the RSTD model at Kennewick, but RSTD is difficult to beat at Goodnoe Hills. The TDD model has a significantly lower RMSE at Kennewick than the RSTD model does (p-value = 0.0052), and both TDD and BST have the smallest RMSE, MAE, or PCE in various months. Predicting at Kennewick is more difficult due to the more highly variable wind speeds observed there, which is also reflected in Kennewick's larger PCE values. Goodnoe Hills is the one location situated directly in the Columbia River Gorge, so the regime-switching model best captures the wind flow pattern. Goodnoe Hills also has the fewest unusually large wind speeds, which is evidenced by the lower RMSE and MAE values. In this situation, RSTD has the lowest overall PCE, but it is not significantly different from that of TDD (p-value $= 0.7550$).

3.4.4 Finer Scale Data

One final experiment on the models returns us to the full dataset with wind speed and direction measured every ten minutes. This finer scale of data exhibits more variability and is not as predictable as the hourly averaged wind speed. Two approaches are tested in which models are rebuilt both on the full dataset and on the ten-minute observations that occur on the hour. For models built on all ten-minute observations, a twelve-step forecast horizon is needed to arrive at the two-hour prediction. Predictions are made for $5136 \times 6 = 30,816$ time-steps. The predictions made on the hour are reserved to compare with the model built from the ten-minute observations that occur on the hour. In that model, a two-step forecast is the two-hour forecast, and only $5,136$ predictions are made.

Table 6: RSTD, TDD, and BST model outcomes for the two types of models built on the ten-minute data.

All Ten-Min	Forecast	May	Jun	Jul	Aug	Sep	Oct	Nov	Overall
RMSE	RSTD	1.95	1.77	1.90	1.99	1.96	2.23	**2.13**	2.00
	TDD	1.90	**1.72**	1.84	**1.98**	**1.93**	**2.22**	**2.13**	**1.97**
	BST	**1.85**	1.73	**1.76**	2.00	2.02	2.32	2.35	2.02
MAE	RSTD	1.48	1.37	1.50	1.52	1.50	**1.62**	**1.57**	1.51
	TDD	1.45	1.32	1.44	**1.49**	**1.47**	1.63	1.60	**1.49**
	BST	**1.39**	**1.31**	**1.36**	1.51	1.55	1.68	1.77	1.51
PCE	RSTD	79.42	56.21	83.57	**69.30**	60.85	75.39	52.38	68.32
	TDD	78.10	**54.86**	79.34	69.64	**59.53**	**75.26**	**51.60**	**67.07**
	BST	**74.16**	55.61	**75.74**	74.47	65.86	79.73	63.07	69.92

Hourly Ten-Min	Forecast	May	Jun	Jul	Aug	Sep	Oct	Nov	Overall
RMSE	RSTD	1.92	1.77	1.90	1.99	1.96	2.22	2.14	1.99
	TDD	1.90	**1.73**	1.84	1.98	**1.93**	**2.21**	**2.13**	**1.97**
	BST	**1.86**	1.74	**1.76**	**1.97**	2.01	2.27	2.30	2.00
MAE	RSTD	1.46	1.37	1.49	1.51	1.50	**1.61**	**1.59**	1.51
	TDD	1.44	1.33	1.44	**1.48**	**1.47**	1.62	1.61	**1.48**
	BST	**1.39**	**1.32**	**1.36**	**1.48**	1.53	1.64	1.74	1.49
PCE	RSTD	78.80	56.24	81.84	70.37	61.11	75.45	**52.39**	68.19
	TDD	77.65	**54.83**	78.50	**70.31**	**60.07**	**74.05**	53.11	**67.08**
	BST	**75.89**	55.23	**74.50**	72.56	65.07	78.49	59.50	68.87

Questions of interest in these models include whether using the full set of ten-minute observations will improve the two-hour forecast and whether the models will have similar results to those in Table 2. In Table 6, it is shown that models built with all of the ten-minute observations have very little predictive improvement compared to the models using only the ten-minute observations on the hour. However, the TDD model appears stronger relative to the RSTD model than it does in Table 2. In fact, it is significantly better than the RSTD model in terms of MSE for both the full set of observations and the ten-minute observations on the hour (p-values 0.0031 and 0.0000, respectively) and also in terms of MAE (p-values 0.0059 and 0.0088).

3.4.5 Underestimation Penalty

The weight that is given in Section 3.3 for the Power Curve Error, $\gamma = 0.73$, deserves some attention. The purpose of this weight is to penalize underestimation more strongly than overestimation of wind power. However, it is not a fixed value. In the Dutch market over the course of the year, the value of γ ranges from 0.51 to 0.98 through the 4 quarters of the year, and it varies from 0.14 to 0.96 over the 12 months of the year (Pinson et al., 2007). Markets with different sets of rules can also affect the value. In addition, a single wind farm usually does not produce enough energy to affect electricity prices, but the larger the penetration of wind energy, the more significantly γ would be affected.

We have used $\gamma = 0.73$ as an example up to this point, but in Table 7, we show the value of PCE for the three models based on hourly data when $\gamma = 0.73$ is replaced with a range of values. We want to determine if the results from PCE are influenced by the value of γ, and in each case, the optimal γth forecast is used in the computation of PCE. With the smallest and largest values of γ, no one model has a consistently smallest PCE over the months. When $\gamma = 0.10$, BST has more small monthly values of PCE than the other models, and when $\gamma = 0.90$, the TDD model appears to be favored.

Table 7: RSTD, TDD, and BST model PCE results for varying penalties on underestimation versus overestimation. A value of γ less (more) than 0.50 penalizes overestimates more (less) heavily than underestimation.

γ	Forecast	May	Jun	Jul	Aug	Sep	Oct	Nov	Overall
	RSTD	4.44	**5.05**	4.20	**3.92**	3.69	11.06	**6.02**	5.49
0.01	TDD	**4.40**	5.13	**4.15**	4.04	3.86	**10.59**	6.14	**5.48**
	BST	4.76	5.72	4.47	4.15	**3.47**	13.02	6.72	6.05
	RSTD	**32.80**	**25.16**	33.55	27.55	25.03	42.10	26.10	30.40
0.10	TDD	33.07	25.49	32.94	27.67	25.17	41.24	**25.85**	30.27
	BST	33.28	26.88	**32.48**	**26.45**	**23.99**	**40.31**	26.90	**30.10**
	RSTD	77.28	**57.24**	84.96	69.28	63.21	81.65	55.23	70.00
0.50	TDD	78.60	57.48	83.77	**68.54**	**62.69**	80.99	**53.00**	**69.46**
	BST	**75.24**	59.56	**82.60**	70.58	65.29	**80.64**	57.64	70.35
	RSTD	**41.36**	**26.61**	40.55	36.38	33.90	48.24	35.37	37.56
0.90	TDD	42.12	27.14	**40.29**	**36.16**	**33.26**	**45.87**	**31.26**	**36.67**
	BST	43.83	30.84	41.16	40.80	38.52	50.10	35.19	40.14
	RSTD	7.76	4.65	**5.95**	7.79	**7.01**	29.28	42.14	14.90
0.99	TDD	**7.65**	**4.61**	6.44	**6.97**	7.48	26.63	**41.56**	**14.43**
	BST	8.89	6.82	5.97	10.85	8.17	**23.01**	43.39	15.24

When $\gamma = 0.50$ and there are no penalties for overproducing or underproducing, both BST and TDD have the smallest PCE for each of 3 months. This experiment just serves to demonstrate that no one model is routinely favored over the others for every possible value of γ, so PCE should be used only after a relatively stable estimate of γ for a given market can be determined.

3.5 Conclusion

The importance of conserving natural resources and exploiting the clean electricity provided by wind energy will only continue to grow in the future. One goal of this paper has been to present model-building strategies for short-term wind speed predictions when both the wind speed and direction information is available over space and time. Wind farms with different terrain and different numbers of nearby meteorological stations can use the TDD or BST modeling approaches to fit similar predictive mean functions, whereas the RSTD model is limited to few locations and known physics. Additionally, speed and direction are often converted to the Cartesian coordinate system, but models like TDD demonstrate the benefit of treating wind direction as a circular variable instead. To conclude, the TDD model produces forecasts that are as good as the RSTD model for this dataset while maintaining more generality. The BST model does not perform as well in terms of PCE on this data, but it does have the added feature of producing a wind direction forecast, which the other two models cannot do.

In comparing models, the power curve error assigns a greater penalty to wind speeds predicted to be in the region where power is roughly proportional to the cube of speed and also penalizes underestimates more strongly than overestimates. Attributing loss in this way directly exploits the nonlinear relationship between power and speed and puts wind power into the larger context of the entire utility system. PCE can easily be adapted for different turbines and different markets and can be averaged over several wind farms to get a more stable estimate. Finally, it may not be reasonable to assume that an error made at a low power has the same economic cost as the same error made at a higher power. An investigation into the effect that the magnitude of wind power for a given error has on the associated loss would need to be conducted.

The work done here could be extended in several ways. Future tests of these models should incorporate year-round observations so that model performance can be assessed in every season. Including additional covariate information, such as equatorial Pacific Ocean sea surface temperatures that affect storm frequency, numerical weather prediction model output, or pressure differences east and west of Vansycle, should also improve predictions. The optimality of the forecasts can continue to be evaluated with tests such as those introduced in Patton and Timmermann (2007).

While the focus in this work has been on point forecasts, having uncertainty estimates of the forecasts that include uncertainty about the parameter estimates and variable selection would also be of interest. Either model-free bootstrapping techniques (Alonso, Peña, and Romo, 2006) or using a fully Bayesian analysis (Wikle, Milliff, Nychka, and Berliner, 2001) could be interesting approaches to obtain such intervals. Finally, wind farms with dominant weather patterns that differ from those of the Pacific Northwest and with varying numbers and locations of off-site observations would be interesting applications for the TDD and BST models. The TDD and BST models' predictions for this data are promising that these flexible models could work well with new datasets.

Note: All circular plots were plotted using the `circular` package in `R` by Lund and Agostinelli (2006).

CHAPTER IV

SPATIAL FORECAST ACCURACY TEST

4.1 Introduction

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

Comparing the accuracy of forecasts is common in time series analysis. Beginning with the seminal work of Diebold and Mariano (1995), a test of the null hypothesis of equal forecast accuracy between two competing models was introduced. 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 correlation in competing forecast errors. Many extensions and improvements to this test have been made (West, 1996; McCracken, 2004; Harvey, Leybourne, and Newbold, 1997; Giacomini and White, 2006), and we develop a similar type of hypothesis test for spatial data that incorporates unique features of spatial data not encountered in time series.

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 (Wang, Anderson, Entekhabi, Huang, Su, Kaufmann, Potter, and Myneni, 2007; Snell, Gopal, and Kaufmann, 2000). Some have simply noted that no such test is available that incorporates the spatial correlation across forecasts (Longhi and Nijkamp, 2007). Many give point estimates of forecast 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 (Atger, 2003; Gong, Barnston, and Ward, 2003; Willis, 2002).

Currently, forecasting wind speeds for wind power generation is a particularly important area of application in which such a forecast accuracy test would be beneficial (Genton and Hering, 2007; Willis, 2002). 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 the state. The United States possesses vast regions in which many wind farms have been built, such as the western region of Texas. For a given point in time, spatial forecasts at these wind farms help utility managers plan for the transmission, purchase, and distribution of electricity. Forecasts made by competing models can be evaluated with a unique loss function that incorporates the nonlinear relationship between wind speed and wind power and a penalty for over or underestimation of wind speed (Hering and Genton, 2009). The forecast accuracy test that is described in this work would be instrumental in determining if on average a difference in the loss produced by competing forecasts is significant.

The extension of the DM test we describe here is appropriate for testing the null hypothesis that on average there is no significant difference between two sets of spatial forecasts. It does not require the forecast errors to be Gaussian or zero-mean, and it allows for both spatial correlation within the forecast errors and contemporaneous correlation between the forecast errors. Contemporaneous correlation is an important element to consider since many models share sources of information, thereby making simultaneously good or bad forecasts 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, 2008).

To the best of our knowledge, no other method exists that tests the same hypothesis as this proposal. One approach that is similar in nature is based on improving the power of the false discovery rate methodology by performing a wavelet decomposition of the spatial field (Shen et al., 2002; Sedur, Maxim, and Whitcher, 2005). This methodology, hereafter SHC, tests for a difference in spatial signal at every location in the domain as opposed to ours that tests for a difference in spatial signal on average across all locations in the domain. The wavelet-based approach will determine not only if a significant difference between two spatial signals exists but also where in the domain the difference occurs. The drawback to a wavelet approach is that 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; Matsuo, Nychka, and Paul, 2006; Shi and Cressie, 2007). Nonstandard grid sizes need to be padded with zeroes or a combination of multiscale wavelets may be used (Deckmyn and Berre, 2005). This method is developed for data assumed to be Gaussian and does not perform well under various loss functions that change the distribution of the data.

Our test procedure 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 SHC method. In Section 2 the background of forecast accuracy tests in time series are reviewed, and these ideas are extended to the spatial setting in Section 3. Section 4 summarizes the Shen et al. (2002) wavelet methodology, which will be used for comparative purposes. Size and power properties are demonstrated with Monte Carlo experiments in Section 5. Section 6 provides an applied example of the test to daily average wind speeds in Oklahoma, and we conclude with some discussion in Section 7.

4.2 History of the Test in Time Series

4.2.1 The Asymptotic DM 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 $e_{it} = y_{it} - \hat{y}_{it}$. 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

$$H_0 : E[g(e_{1t})] = E[g(e_{2t})] \quad \text{or} \quad H_0 : E[d_t] = 0,$$

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, $\bar{d} = \frac{1}{T}\sum_{t=1}^{T}[g(e_{1t}) - g(e_{2t})]$ is such that

$$\sqrt{T}(\bar{d} - \mu) \rightarrow N(0, 2\pi s_d(0))$$

in distribution as T goes to infinity. Here, μ is the population mean loss differential, and $s_d(0)$ is the spectral density of the loss differential at frequency 0. It is defined to be

$$s_d(0) = \frac{1}{2\pi}\sum_{\tau=-\infty}^{\infty}\gamma_d(\tau)$$

for $\gamma_d(\tau) = E[(d_t - \mu)(d_{t-\tau} - \mu)]$ the autocovariance of the loss differential at lag τ.

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

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

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,

$$2\pi \hat{s}_d(0) = \sum_{\tau=-(T-1)}^{(T-1)} 1\left(\frac{\tau}{S(T)}\right)\hat{\gamma}_d(\tau), \tag{4.1}$$

where

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

and $1(\tau/S(T))$ is the lag window, and $S(T)$ is the truncation lag.

The choice of lag window and truncation lag are motivated by the result that the optimal k-step forecast errors are at most $(k-1)$ dependent, which can be checked empirically. This suggests the uniform, or rectangular, lag window defined by

$$1\left(\frac{\tau}{S(T)}\right) = \begin{cases} 1 & \text{when } \left|\frac{\tau}{S(T)}\right| \le 1, \\ 0 & \text{otherwise .} \end{cases} \tag{4.2}$$

This uniform window assigns unit weight to all included autocovariances, and only $(k-1)$ sample autocovariances are used in the estimation of $s_d(0)$ because all others are set to zero.

Diebold and Mariano (1995) discuss their choice of lag window. They say that the Dirichlet spectral window associated with the rectangular lag window dips below 0 at certain locations, so the resulting estimator of the spectral density is not guaranteed to be positive semidefinite. However, because the Dirichlet kernel assigns a large positive weight near the origin, the estimate of $s_d(0)$ is unlikely to be negative. In practice, they treat a negative estimate of $s_d(0)$ as an automatic rejection of the null hypothesis.

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 (4.1), we estimate all of the 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, we only retain the empirical autocovariances computed up to half of the maximum lag. We fit an exponential covariogram of the form

$$C(\tau) = \sigma^2 \exp^{-3\tau/\theta}$$

using ordinary least squares or weighted least squares to the empirical autocovariances. We use $\hat{C}(\tau)$ to estimate the values of $\gamma_d(\tau)$ in

$$2\pi \hat{f}_d(0) = \hat{\gamma}_d(0) + 2\sum_{\tau=1}^{T-1} \hat{\gamma}_d(\tau)$$

for the parametrically estimated test statistic

$$S_p = \frac{\bar{d}}{\sqrt{\frac{2\pi \hat{f}_d(0)}{T}}}.$$

We compare this method to that described by Diebold and Mariano (1995) by simulating forecast errors as they describe and applying the quadratic loss. They represent the contemporaneous correlation with ρ and the (moving average) MA(1) parameter with θ. Using $\alpha = 0.10$ as they do in their work, the observed size of the test is dramatically improved with S_p. Table 8 shows that in samples of size 8, the size is reduced between 52.4% and 58.1%. The difference in sizes is evident from samples of size 8 through samples of size 64. In fact, empirical sizes reach the desired level at $n = 128$ for the DM test but at $n = 32$ for the parametric test.

4.2.2 Extensions of the DM Test

Many authors have worked to improve the DM test in the years since it was published. One of the first responses was a paper by West (1996) which criticized the DM test for failing to address the fact that the forecasts can depend upon estimated regression parameters. Diebold and Mariano (1995) do not mention this fact, so their test implicitly assumes that

Table 8: Empirical size of quadratic loss function for time series data simulated according to the parameters and description in Diebold and Mariano (1995). The results for their asymptotic test statistic, S_1 and for the parametrically estimated S_p, which uses an OLS exponential covariogram estimate of variance of $\bar{D}$ are given. 2,500 datasets are simulated for each combination of parameters, and $\alpha = 0.10$.

		DM Test			S_p Test		
T	ρ	$\theta = 0$	$\theta = 0.5$	$\theta = 0.9$	$\theta = 0$	$\theta = 0.5$	$\theta = 0.9$
8	0.0	32.52	31.28	28.96	15.20	13.12	12.92
8	0.5	31.36	30.00	27.16	14.12	12.20	12.92
8	0.9	32.52	28.76	29.24	15.00	11.92	14.20
16	0.0	20.48	20.00	17.60	11.08	11.56	11.96
16	0.5	21.96	19.36	17.80	13.32	10.36	12.40
16	0.9	20.48	19.72	18.12	10.40	10.96	10.76
32	0.0	15.12	13.56	13.12	9.72	9.56	10.92
32	0.5	16.92	14.48	14.72	11.76	10.40	11.92
32	0.9	14.56	13.48	12.88	10.36	9.80	10.00
64	0.0	12.04	11.44	10.92	9.72	9.64	10.04
64	0.5	12.32	11.96	11.56	9.80	10.12	10.32
64	0.9	12.88	11.68	11.04	10.24	10.00	9.76
128	0.0	11.52	10.32	10.00	10.04	9.32	9.48
128	0.5	11.44	10.28	10.40	9.44	9.60	9.32
128	0.9	12.04	11.48	9.52	10.36	10.12	8.84
256	0.0	10.84	10.44	10.64	9.72	10.08	10.16
256	0.5	9.96	10.24	10.64	9.00	9.80	10.28
256	0.9	10.00	9.56	10.88	9.20	8.88	10.52
512	0.0	9.72	10.84	10.84	9.16	10.64	10.48
512	0.5	10.96	9.88	10.12	10.36	9.60	9.60
512	0.9	11.64	9.80	10.20	11.12	9.44	9.40

Standard errors of values in the table are between 0.6% and 1.0%.

the regression parameters are known. The adjustment in the error due to estimating regression parameters depends upon several factors such as what moment is being estimated in the loss function, the regression technique, the fraction of the total sample used for out-of-sample estimation of the loss, and the probabilistic environment. West assumes that the loss function is twice differentiable in a neighborhood of the parameter vector and contends that although this excludes mean absolute error (MAE), many important loss functions are still included. McCracken (2004) allows the loss function to be nondifferentiable.

Several situations arise in which the additional variance of the loss differential that is due to estimating parameters is asymptotically irrelevant. For example, if the number of observations used to estimate the unknown parameters is large relative to the number of forecasts made, then the parameters can be treated as if they are known. Also, when the predictors are uncorrelated with the prediction error, such as with MSE or comparing non-nested models, then uncertainty due to the parameters is not important. In these cases, West's test reduces to the DM test.

An adjustment for the bias in the variance of the mean differenced series of the DM test was proposed by (Harvey et al., 1997). This improves the size of the test in small to moderately sized samples. They also propose a companion test to the forecast accuracy test called the forecast "encompassing" test (Harvey, Leybourne, and Newbold, 1998). They make a combined forecast by taking weighted averages of individual forecasts, and if the optimal combined forecast places all of the weight on one individual forecast, then that individual forecast is said to encompass the others. A robust version of the DM test is suggested by Dell'Aquila and Ronchetti (2004) who show that in small samples distributional deviations can have a large impact on the size of the DM test. Their method also identifies points that have a large influence on the size and power of the test.

Finally, Giacomini and White (2006) unite much of the preceding theory with their test of predictive ability that allows the forecasting models to be possibly misspecified, accounts for parameter estimation in the models, and tests conditional (which forecast is best for a particular horizon) versus unconditional (which forecast is best on average) forecasting objectives. Their treatment unifies theory on nested and non-nested models. Their test applies to multistep point, interval, probability, or density forecast evaluations. They also propose a two-step decision rule to select the best forecast for a future date of interest.

Many of these extensions would be interesting to study in the spatial context as well. However, given that many difficulties are already presented by spatial data that are not encountered in time series, we will assume that parameters in the forecasting models can be estimated well by the data. We also assume that the sample size is adequately large and that influential outliers are not present in the data.

4.3 Spatial Forecast Accuracy Test

We propose a test for spatial forecast accuracy following the form of the DM test for time series data. Consider a spatial process $\{Z(\mathbf{s}) : \mathbf{s} \in \mathcal{D} \subset \mathbb{R}^2\}$ that has been observed at n locations. The observed value is denoted $Z(\mathbf{s}_i)$, for $i = 1, 2, \ldots, n$. 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 forecast based on models built from the $(1 - \phi)n$ observations. Let L represent the number of randomly chosen locations to forecast, thus $L = \phi \cdot n$. Two sets of spatial forecasts are made, denoted by $\{\hat{Z}_1(\mathbf{s}_i)\}_{i=1}^{L}$ and $\{\hat{Z}_2(\mathbf{s}_i)\}_{i=1}^{L}$. The associated forecast errors are $\{e_1(\mathbf{s}_i)\}_{i=1}^{L}$ and $\{e_2(\mathbf{s}_i)\}_{i=1}^{L}$. Many times, it will be a direct realization of the forecast error, $g\left(e_j(\mathbf{s}_i)\right)$ for $j = 1, 2$.

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

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

where $\bar{Z}(\mathbf{s})$ is the mean of the L observed values, $\bar{Z}_j(\mathbf{s})$ is the mean of the L forecasts from model j, $s_{Z(\mathbf{s})}$ is the standard deviation of the observed values, and $s_{\hat{Z}_j(\mathbf{s})}$ is the standard deviation of the forecasts. In this way, the correlation skill of forecasts produced by competing models can be tested with this method.

The spatial process of interest takes the following form:

$$D(\mathbf{s}) = g\left(e_1(\mathbf{s})\right) - g\left(e_2(\mathbf{s})\right) = f(\mathbf{s}) + \delta(\mathbf{s}), \quad \mathbf{s} \in \mathcal{D}, \tag{4.3}$$

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

$$H_0 : \frac{1}{|\mathcal{D}|}\int_{\mathcal{D}} E[D(\mathbf{s})]\; d\mathbf{s} = 0, \tag{4.4}$$

where $|\mathcal{D}|$ is the area of the domain. 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 $H_0 : \mu = 0$ and reduces to a spatial version of the DM test. However, in many cases it is unlikely that the mean is constant across all locations. In this case, the null hypothesis tests that the average of the 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 forecast accuracy 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 forecast accuracy 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 spatial covariance that approaches zero as the lag distance increases (Park, Kim, Park, and Hwang, 2008), the sample mean loss differential, $\bar{D} = \frac{1}{L}\sum_{i=1}^{L} D(\mathbf{s}_i)$ is asymptotically normal,

$$\frac{\bar{D}-\mu}{\sqrt{\text{Var}\left(\bar{D}\right)}} \rightarrow N(0,1),$$

where

$$\text{Var}\left[\bar{D}\right] = \text{Var}\left[\frac{1}{L}\sum_{i=1}^{L} D(\mathbf{s}_i)\right] = \frac{1}{L^2}\sum_{i=1}^{L}\sum_{j=1}^{L} C(h_{ij}). \tag{4.5}$$

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 $\mathbf{s}_j$. All forms of the test statistics we employ to test the hypothesis in (4.4) are based on some version of Equation (4.5) 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 the trend is constant across space, i.e., $f(\mathbf{s}) = \mu$ for μ some constant. The typical empirical estimate of $C(h_{ij})$ is

$$\hat{C}(h_{ij}) = \frac{1}{|N(h_{ij})|}\sum_{N(h_{ij})}\left(D(\mathbf{s}_i)-\bar{D}\right)\left(D(\mathbf{s}_j)-\bar{D}\right), \tag{4.6}$$

where $N(h_{ij})$ is the set of all pairs of locations that are distance h_{ij} apart. Whereas in both the time series setting (Brockwell and Davis, 1991) and the space-time setting (Nychka et al., 2002), such an estimator would have a valid positive definite form, in the purely spatial setting, it does not. In addition, we have the following fact that follows from a similar outcome in time series (Percival, 1993).

Proposition 1. *The sum*

$$V\hat{a}r\left[\bar{D}\right] = V\hat{a}r\left[\frac{1}{L}\sum_{i=1}^{L} D(\mathbf{s}_i)\right] = \frac{1}{L^2}\sum_{i=1}^{L}\sum_{j=1}^{L}\hat{C}(h_{ij}) = 0,$$

where $\hat{C}(h_{ij})$ is given in Equation (4.6).

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

1. 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.

2. 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 (4.5) is truncated at $k-1$ when making k-step forecasts. In the spatial setting, the distance between a location to forecast and an observed location is not constant. Therefore, we turn to parametric estimates of the spatial covariance in which a positive definite form will be guaranteed, and only empirical estimates of $C(h_{ij})$ up to half of the maximum lag are used in forming the parametric estimate, which is a common rule of thumb. An alternative to estimating the covariogram would be to estimate the semivariogram, $\gamma(h_{ij})$, taking advantage of the relationship $\gamma(h_{ij}) = C(0) - C(h_{ij})$. Then, replace $\hat{C}(h_{ij})$ with $\hat{\gamma}(\infty) - \hat{\gamma}(h_{ij})$ in Equation (4.5) where

$$\hat{\gamma}(h_{ij}) = \frac{1}{|N(h_{ij})|}\sum_{N(h_{ij})}(D(\mathbf{s}_i) - D(\mathbf{s}_j))^2. \tag{4.7}$$

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

$$W(\boldsymbol{\theta}) = \sum_{i=1}^{p}|N(h_i)|\left(\frac{\hat{\gamma}(h_i)}{\gamma(h_i|\boldsymbol{\theta})} - 1\right)^2 \tag{4.8}$$

for the semivariogram (Cressie, 1985) and

$$W(\boldsymbol{\theta}) = \sum_{i=1}^{p} |N(h_i)| \left(\frac{\hat{r}(h_i) - r(h_i|\boldsymbol{\theta})}{1 - r(h_i|\boldsymbol{\theta})} \right)^2 \tag{4.9}$$

for the correlogram (Gneiting, 2002). In Equations (4.8) and (4.9), the functions $\gamma(h_i|\boldsymbol{\theta})$ and $r(h_i|\boldsymbol{\theta})$ are the parametric forms of the semivariogram and correlogram, respectively, with parameters $\boldsymbol{\theta}$. The maximum lag to which to sum, in each equation is p, defined as half of the maximum lag. Using maximum likelihood to estimate the parameters in the covariogram or semivariogram is another 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 forecast error can change the distribution of the data even when the forecast errors are Gaussian. Therefore, assumptions about the distribution of the data are avoided when fitting the covariogram and semivariogram with weighted least squares.

Thus, we propose the following two potential test statistics for testing the hypothesis of equal forecast accuracy under constant trend:

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

In application, the assumption of isotropy should be tested first (Li, Genton, and Sherman, 2007), 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. Simple extensions can be made when the data $D(\mathbf{s})$ is irregularly spaced by smoothing the observations within a specified tolerance region (Cressie, 1993).

As mentioned previously, a non-constant 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) do not need to estimate the trend of their loss differential series since all forecasts are for the same forecast horizon. However, with any set of spatial forecasts, the trend can be a concern since the forecasts are all made at varying lag distances. 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 (4.6) and (4.7) 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 covariogram or semivariogram estimated from detrended data by $\hat{C}^r(h_{ij}|\hat{\boldsymbol{\theta}})$ and $\hat{\gamma}^r(h_{ij}|\hat{\boldsymbol{\theta}})$, respectively. Then, the test statistics become

$$S_C^r = \frac{\bar{D}}{\sqrt{\frac{1}{L^2}\sum_{i=1}^{L}\sum_{j=1}^{L}\hat{C}^r(h_{ij}|\hat{\boldsymbol{\theta}})}} \quad \text{and} \quad S_V^r = \frac{\bar{D}}{\sqrt{\frac{1}{L^2}\sum_{i=1}^{L}\sum_{j=1}^{L}(\hat{\gamma}^r(\infty|\hat{\boldsymbol{\theta}}) - \hat{\gamma}^r(h_{ij}|\hat{\boldsymbol{\theta}}))}}.$$

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

$$\hat{D}_b(x_i, y_i) = \frac{\sum_{i=1}^{L} K\left(\frac{x-x_i}{b}\right) K\left(\frac{y-y_i}{b}\right) D(x_i, y_i)}{\sum_{i=1}^{L} K\left(\frac{x-x_i}{b}\right) K\left(\frac{y-y_i}{b}\right)}, \tag{4.10}$$

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

$$\text{CV}(b) = \frac{1}{L}\sum_{i=1}^{L}\left(D(\mathbf{s}_i) - \hat{D}_b^{(-i)}(\mathbf{s}_i)\right)^2, \tag{4.11}$$

where $\hat{D}_b^{(-i)}(\mathbf{s}_i)$ is the estimate of $D(\mathbf{s}_i)$ with the $\mathbf{s}_i$ location removed, is too small when the data is 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 forecast accuracy 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. Similar to the adjustment for time series data (Hart 1996), the adjustment for spatial data is

$$b_a = \left[\sum_{i=1}^{L} \sum_{j=1}^{L} C(h_{ij})/C(0) \right]^{1/5} b_0, \tag{4.12}$$

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

$$C(h_{ij}) = \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases}$$

into Equation (4.12) 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_C^r or S_V^r.

In summary, the steps in performing the spatial forecast accuracy test (SFAT) are as follows:

1. Evaluate the loss at each location for each set of forecasts, and form the differenced field, $D(\mathbf{s})$.

2. Estimate the trend of the differenced field.

3. Compute the test statistic S_C^r or S_V^r.

4. Find the p-value, for example $2(1 - (\Phi(|S_V^r|)))$, where Φ is the cumulative distribution function of a $N(0, 1)$ distribution.

When the trend appears to be constant, then Step 2 can be skipped, and S_C^r and S_V^r can be replaced with S_C and S_V in Step 3.

4.4 An Alternative Approach

Shen et al. (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 difference between 2 spatial signals at every location in the domain. To reduce the number of hypotheses that must be tested, the model is represented in the wavelet domain. Then, the method tests not only if a statistically significant signal is present but also estimates the location and magnitude of such a signal. It must be emphasized that this method is intended to be used on a complete grid of data, such as fMRI or climate model output data. It was not introduced in the forecasting context, and it cannot be applied as generally as the spatial forecast accuracy test can be. However, for comparative purposes, the basic outline is given, and a simulation experiment comparing SHC to the spatial forecast accuracy test will be performed and reported in Section 4.5.4.

The goal of the SHC method is to nonparametrically test the hypothesis that a spatial signal is present or not based on a single image. The problem is stated as $H_0 : f = f_0$ versus $H_a : f \neq f_0$ where $f(\cdot)$ is the deterministic mean function of a spatial Gaussian process $\{Z(\mathbf{s}) : \mathbf{s} \in \mathcal{D} \subset \mathbb{R}^d\}$ generated by $Z(\mathbf{s}) = f(\mathbf{s}) + \delta(\mathbf{s})$. Here, $\delta(\cdot)$ is a mean-0 stationary Gaussian process with an unknown stationary covariance function $C(\mathbf{h}) =$ cov$(\delta(\mathbf{s}), \delta(\mathbf{s} + \mathbf{h}))$. To increase the power of testing whether $f(\cdot)$ is f_0 (and also where and by how much $f(\cdot)$ differs from f_0), a parsimonious representation of $f(\cdot)$ is given with a small number of wavelet coefficients.

Inference is made on $f(\cdot)$ based on observations $\{(\mathbf{s}_i, Z(\mathbf{s}_i)) : i = 1, \ldots, n\}$. An observation $Z(\mathbf{s}_i)$ can be written as $Z(\mathbf{s}_i) = f(\mathbf{s}_i) + \delta(\mathbf{s}_i)$ where $\{\delta(\mathbf{s}_i)\}$ follows a multivariate normal distribution with $n \times n$ covariance matrix V whose (j, k)th element is $C(\mathbf{s}_j - \mathbf{s}_k)$.

The d-dimensional discrete wavelet transform (DWT) is applied to this representation to obtain a representation of wavelet coefficients $\boldsymbol{\nu} = \boldsymbol{\nu}^0 + \boldsymbol{\epsilon}$. Here, $\boldsymbol{\nu}^0 = (\nu_1^0, \ldots, \nu_n^0)'$ is the vector of wavelet coefficients of $\{f(\mathbf{s}_i)\}$. Now, $\boldsymbol{\epsilon}$ is a random component distributed according to $N_n(\mathbf{0}, V^*)$ with an almost diagonal matrix V^*, thereby nearly decorrelating $\{\delta(\mathbf{s}_i)\}$. Then, testing $H_0 : f(\cdot) = 0$ versus $H_a : f(\cdot) \neq 0$ is equivalent to simultaneously testing n simple versus composite hypotheses–$H_{0i} : \nu_i^0 = 0$ versus $H_{ai} : \nu_i^0 \neq 0$ for $i = 1, \ldots, n$.

Part of the procedure is based on False Discovery Rates (FDR). Let R be the number of rejected null hypotheses. Of these R hypotheses, V are erroneously rejected, and $R - V$ are correctly rejected. Define $Q = V/R$ if $R > 0$ and $Q = 0$ if $R = 0$. Then, the FDR is defined as $E(Q)$, the expected proportion of erroneously rejected null hypotheses. In a family of L hypothesis tests to be performed, the *standard FDR procedure* computes the p-value, p_i for each set of hypotheses, $i = 1, \ldots, L$. Then, compute the order statistics of the p-values, $p_{(1)} \leq \cdots \leq p_{(L)}$ corresponding to the hypotheses $H_{0(1)}, \ldots, H_{0(L)}$. Denote by K the largest i for which $p_{(i)} \leq (i/L)\alpha$. If such a $K \geq 1$ exists, then reject all $H_{0(i)}; \;\; i = 1, \ldots, K$. If such a K does not exist, then reject none of the $H_{0i}; \;\; i = 1, \ldots, L$.

In the spatial setting, a 2-dimensional DWT is used on the data. A simultaneous test that each of the wavelet coefficients is 0 or not could be done directly with the standard FDR procedure described above. However, this does not take advantage of the “spatial” structure of the wavelet coefficients that is likely present under the alternative hypotheses. They gain more power by observing that wavelet coefficients of a signal within each scale and across different scales are related. The “large” wavelet coefficients of a pure signal typically cluster both within each scale and across different scales, whereas the corresponding wavelet coefficients of either white noise or correlated noise are approximately uncorrelated. This spatial structure allows the test to predict whether a wavelet coefficient of the signal is 0 or not from observing its neighbors. This can be used to identify individual

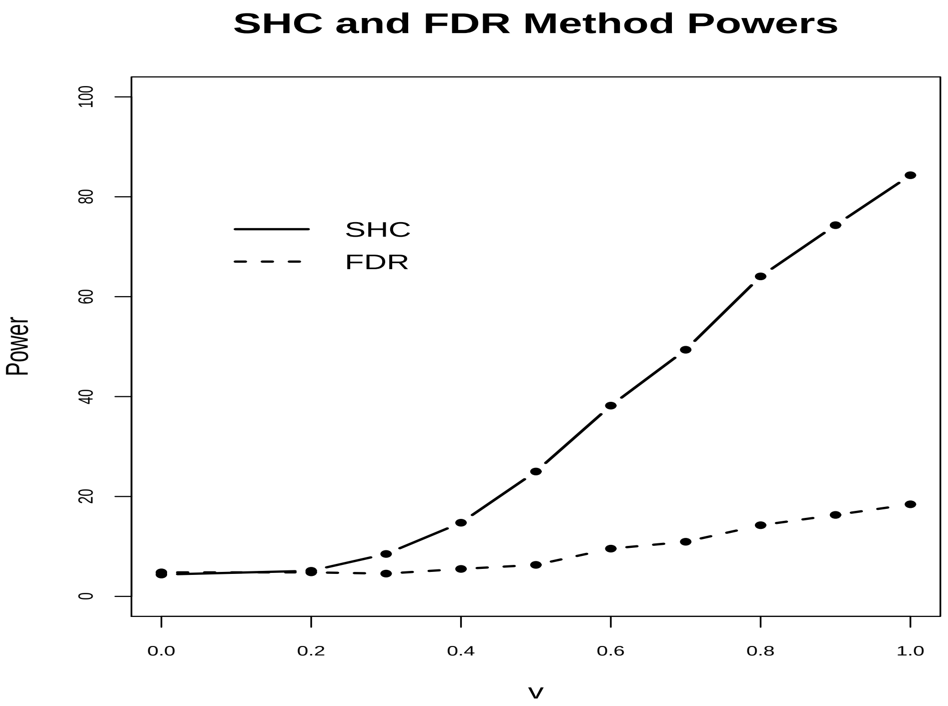


Figure 12: Reproduction of powers from SHC and FDR methods described in Shen et al. (2002). The same number of replicates in the simulation, 1600, are generated for the circle of radius 10 with mean in the interior of the circle varying across the following values of v, 0, 0.1, 0.2, . . ., 0.9, and 1.

hypotheses that should be removed before applying the FDR procedure. This is named the *enhanced* FDR procedure, which we refer to as SHC.

A detailed formula is given for determining which wavelet coefficients are neighbors, and a system with $b = 11$ neighbors is adopted. Then, a method for determining both which and how many wavelet coefficients to retain to be tested with the standard FDR procedure is described. The null hypotheses of the eliminated tests are accepted. The magnitude of f where $f \neq 0$ is estimated by performing the inverse DWT on the retained coefficients. Further details can be found in the original paper.

To illustrate, we have reproduced the size and powers of both the SHC and the FDR methods in which a 64×64 grid of data is generated with a circle of radius 10 in the center with varying values of a non-zero mean, v. In other words, for $\mathbf{s} = (x, y)$

$$f(\mathbf{s}) = \begin{cases} v, & (x - 32.5)^2 + (y - 32.5)^2 \leq 10^2, \\ 0, & \text{otherwise}, \end{cases}$$

and $\delta(\mathbf{s})$ is standard Gaussian white noise with $C(0) = 1$ and zero elsewhere. As is evident in Figure 12, both methods have the same level, but the SHC method is far more powerful in rejecting false null hypotheses.

4.5 Monte Carlo Simulation Study

In this section, the finite sample size and power properties of the test statistics S_C and S_V for $D(\mathbf{s})$ with constant trend and of S_V^r with spatially varying trend in simulated datasets are presented. The data simulation method is first described, and then results under the null and nonzero constant trend are presented. Three different types of alternatives with spatially varying mean are also simulated, and the effect of estimating the trend both under the null and for these alternatives is explored. Finally, the power of these test statistics in comparison with the Shen et al. (2002) wavelet method when a complete dyadic grid of forecasts is available is presented for various values of constant trend and for the circular pattern tested in their simulations.

4.5.1 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 outline is to generate two sets of forecast 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. 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 (Gelfand, Schmidt, Banerjee, and Sirmans, 2004). This model allows each set of forecast errors to have its own spatial correlation.

The general cross-covariance matrix is

$$C_X(h_{ij}) = \sum_{k=1}^{2} r_k(h_{ij})\mathbf{a}_k\mathbf{a}_k^T,$$

where $r_k(h_{ij}) = \exp\{-3h_{ij}/\theta_k\}$ is the stationary exponential correlation function for the kth process, and $\mathbf{a}_k^T$ is a column of A. In the bivariate case, A is defined based on T where T is defined as

$$T = \begin{bmatrix} \sigma_1^2 & \rho \\ \rho & \sigma_2^2 \end{bmatrix}.$$

Subsequently, A is

$$A = \begin{bmatrix} \sqrt{t_{11}} & 0 \\ t_{12}/\sqrt{t_{11}} & \sqrt{t_{22} - t_{12}^2/t_{11}} \end{bmatrix} = \begin{bmatrix} \sigma_1 & 0 \\ \rho/\sigma_1 & \sqrt{\sigma_2^2 - \rho^2/\sigma_1^2} \end{bmatrix}.$$

Here, σ_1^2 and σ_2^2 are the variability of the first and second set of forecast errors, respectively, and both are set to 1. The contemporaneous correlation between the forecast errors is denoted by ρ. We then generate the bivariate random field from a Gaussian distribution with mean zero and $2n \times 2n$ (where $n = r \times c$) variance-covariance matrix of the forecast errors

$$C_e(h_{ij}) = \begin{bmatrix} \sigma_1^2 e^{-3(h_{ij})/\theta_1} & \rho e^{-3(h_{ij})/\theta_1} \\ \rho e^{-3(h_{ij})/\theta_1} & \sigma_2^2 e^{-3(h_{ij})/\theta_2} + \left(\frac{\rho^2}{\sigma_1^2}\right)\left(e^{-3(h_{ij})/\theta_1} - e^{-3(h_{ij})/\theta_2}\right) \end{bmatrix}.$$

Note that the spatial range of the first set of forecast 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 forecast errors equal to θ_2.

We generate grids of sizes 5×5, 8×8, 10×10, 16×16, 20×20, and 25×25. With a forecasting fraction of $\phi = 0.40$, the number of randomly selected locations for each grid size is $L = 10$, 25, 40, 102, 160, and 250. We consider values of the contemporaneous correlation parameter, ρ, to be 0, 0.5, and 0.9. The spatial correlation parameters vary among $\theta_1 = \theta_2 = 3$; $\theta_1 = \theta_2 = 6$; $\theta_1 = \theta_2 = 9$; and $\theta_1 = 3,\ \theta_2 = 9$. The variance of each process is set to 1 by dividing each simulated set of forecast 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. Three loss functions are evaluated, the quadratic loss, $g_1(e(\mathbf{s})) = (e(\mathbf{s}))^2$, the absolute loss, $g_2(e(\mathbf{s})) = |e(\mathbf{s})|$, and the simple loss, $g_3(e(\mathbf{s})) = e(\mathbf{s})$. Unless otherwise stated, for each combination of parameters, 2500 simulated datasets are generated.

4.5.2 Constant Trend

In this section, both the size and power properties of the spatial forecast accuracy 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 true covariance of $\bar{D}$ under the simple loss function is known and can be derived from $C_e(h_{ij})$. The test statistic with this simulated or true variance is denoted $S_T = \frac{\bar{D}}{\sqrt{\hat{\sigma}_D^2}}$. Tables 9, 10, and 11 give the size results (i.e., $f(\mathbf{s}) = 0$) for test statistics S_T, S_C, and S_V, respectively.

Table 9: Empirical size of loss functions under the true or simulated true estimate of variance of $\bar{D}$ for the spatial accuracy test. All tests are reported at the 5% level, and 2,500 Monte Carlo replications are performed.

			Quadratic Loss				Absolute Loss				Simple Loss			
Grid	L	ρ	$\theta_{i,j} = 3$	$\theta_{i,j} = 6$	$\theta_{i,j} = 9$	$\theta_{i,j} = 3, 9$	$\theta_{i,j} = 3$	$\theta_{i,j} = 6$	$\theta_{i,j} = 9$	$\theta_{i,j} = 3, 9$	$\theta_{i,j} = 3$	$\theta_{i,j} = 6$	$\theta_{i,j} = 9$	$\theta_{i,j} = 3, 9$
5	10	0.0	6.60	6.64	5.92	5.60	5.88	6.72	5.72	5.20	5.00	5.40	5.12	4.60
5	10	0.5	6.20	6.32	6.56	5.12	6.00	5.88	6.12	5.28	4.72	4.68	5.08	5.04
5	10	0.9	5.52	4.96	5.40	5.64	4.92	5.56	5.52	5.56	4.64	4.96	5.08	5.40
8	25	0.0	4.76	5.72	6.60	4.80	4.72	5.44	6.56	4.96	4.96	6.12	5.08	4.80
8	25	0.5	5.08	5.20	5.48	5.16	4.44	4.72	5.48	4.92	4.76	4.76	4.60	5.76
8	25	0.9	5.12	6.28	5.84	5.32	5.08	5.40	5.48	5.36	4.76	5.08	4.76	4.76
10	40	0.0	5.32	5.40	5.92	5.28	5.40	4.80	5.88	5.12	4.92	5.04	5.08	4.48
10	40	0.5	5.12	5.56	5.16	5.52	5.80	5.60	4.76	5.48	4.76	5.08	5.24	5.32
10	40	0.9	5.36	5.52	6.80	5.56	5.24	4.64	7.08	4.72	5.00	5.64	4.84	4.88
16	102	0.0	5.20	5.36	4.80	4.76	5.40	4.92	4.12	5.08	5.32	4.60	4.44	4.48
16	102	0.5	4.92	6.08	5.68	4.96	4.52	5.52	5.80	4.88	4.60	4.92	4.92	4.64
16	102	0.9	4.24	6.00	5.20	5.44	4.48	5.28	5.20	5.52	5.44	4.80	4.32	4.60
20	160	0.0	4.64	5.64	4.76	5.52	5.08	5.60	4.48	5.68	5.08	5.88	5.48	4.68
20	160	0.5	4.76	5.32	6.60	5.36	4.36	4.76	6.12	5.08	4.08	4.80	4.56	5.28
20	160	0.9	4.28	5.96	6.04	5.00	5.00	4.68	5.68	5.04	5.44	4.68	4.80	5.04
25	250	0.0	4.72	5.28	4.88	4.08	4.56	4.96	4.76	4.36	4.36	5.00	4.96	4.88
25	250	0.5	5.32	5.32	5.44	4.72	5.36	5.24	5.24	4.96	4.68	4.72	4.60	5.40
25	250	0.9	5.00	5.08	5.24	5.84	4.84	5.60	4.88	5.24	4.84	5.48	5.20	4.76

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

Table 10: Empirical size of loss functions under WLS covariogram estimate of variance of $\bar{D}$ for the spatial accuracy test. All tests are reported at the 5% level, and 2,500 Monte Carlo replications are performed.

			Quadratic Loss				Absolute Loss				Simple Loss			
Grid	L	ρ	$\theta_{i,j} = 3$	$\theta_{i,j} = 6$	$\theta_{i,j} = 9$	$\theta_{i,j} = 3, 9$	$\theta_{i,j} = 3$	$\theta_{i,j} = 6$	$\theta_{i,j} = 9$	$\theta_{i,j} = 3, 9$	$\theta_{i,j} = 3$	$\theta_{i,j} = 6$	$\theta_{i,j} = 9$	$\theta_{i,j} = 3, 9$
5	10	0.0	10.56	17.16	24.32	17.16	12.08	19.00	25.76	18.60	22.20	40.28	50.56	37.96
5	10	0.5	10.92	17.28	24.80	15.36	12.84	20.00	26.24	17.16	22.68	39.92	48.80	45.00
5	10	0.9	9.28	17.16	22.48	14.88	11.28	17.44	25.00	14.72	21.96	38.60	48.56	48.00
8	25	0.0	7.16	14.04	19.32	14.72	7.80	14.64	21.16	16.24	17.00	31.72	39.36	33.28
8	25	0.5	7.36	12.96	18.04	12.92	7.36	13.44	20.60	14.24	17.92	30.64	41.44	37.40
8	25	0.9	7.48	13.40	18.60	11.32	7.76	13.28	18.92	11.12	16.72	30.48	40.76	39.48
10	40	0.0	7.72	11.08	16.00	14.12	7.84	12.20	17.76	14.92	14.36	27.16	35.20	26.64
10	40	0.5	7.60	11.28	15.28	13.16	7.16	11.64	16.48	13.88	14.36	25.68	35.48	32.08
10	40	0.9	6.60	10.92	16.76	9.24	7.24	11.08	16.80	8.84	15.92	27.44	33.36	34.40
16	102	0.0	5.32	8.28	9.20	12.20	6.20	8.44	10.08	11.68	10.28	15.56	22.48	21.28
16	102	0.5	5.92	9.20	11.16	11.00	5.92	9.20	12.28	11.36	9.32	17.20	24.76	23.36
16	102	0.9	5.56	8.20	10.92	8.48	6.24	8.88	11.88	8.32	10.44	16.32	23.24	23.96
20	160	0.0	5.24	7.40	7.84	10.40	5.64	8.04	8.28	10.44	8.52	15.20	19.52	18.00
20	160	0.5	4.64	6.80	9.40	9.92	5.04	6.96	10.96	9.68	8.08	14.28	19.84	21.12
20	160	0.9	6.44	6.08	8.52	7.56	6.52	6.84	9.40	8.04	9.40	13.60	19.96	19.08
25	250	0.0	5.28	5.76	7.28	9.88	5.40	6.36	8.36	9.44	7.04	11.00	15.72	15.60
25	250	0.5	5.76	6.12	7.84	9.92	5.80	6.84	8.72	9.88	6.92	12.20	15.80	15.88
25	250	0.9	6.04	6.92	7.52	6.24	5.88	7.96	8.68	6.80	6.52	11.44	15.92	16.24

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

Table 11: Empirical size of loss functions under WLS semivariogram estimate of variance of $\bar{D}$ for the spatial accuracy test. All tests are reported at the 5% level, and 2,500 Monte Carlo replications are performed.

			Quadratic Loss				Absolute Loss				Simple Loss			
Grid	L	ρ	$\theta_{i,j} = 3$	$\theta_{i,j} = 6$	$\theta_{i,j} = 9$	$\theta_{i,j} = 3, 9$	$\theta_{i,j} = 3$	$\theta_{i,j} = 6$	$\theta_{i,j} = 9$	$\theta_{i,j} = 3, 9$	$\theta_{i,j} = 3$	$\theta_{i,j} = 6$	$\theta_{i,j} = 9$	$\theta_{i,j} = 3, 9$
5	10	0.0	5.64	9.76	11.68	8.96	7.00	11.36	13.84	10.44	12.68	22.44	26.68	20.52
5	10	0.5	5.60	8.96	13.24	7.80	6.84	11.12	14.60	9.44	12.92	23.12	26.20	24.32
5	10	0.9	5.08	8.68	11.96	7.88	6.12	10.20	13.24	7.36	11.40	22.44	25.16	26.12
8	25	0.0	4.32	6.80	7.40	8.08	5.12	7.52	9.28	9.16	8.40	13.84	16.28	17.64
8	25	0.5	4.28	6.52	7.60	7.64	4.40	7.16	8.68	8.56	9.44	14.52	17.44	17.48
8	25	0.9	4.40	5.92	7.88	5.96	4.56	6.52	10.40	6.12	9.08	13.60	17.52	17.16
10	40	0.0	5.12	6.32	7.28	7.84	5.48	6.72	7.64	9.48	8.36	11.96	14.96	14.32
10	40	0.5	4.64	5.52	6.48	7.96	4.80	7.08	7.88	8.96	8.56	11.80	15.12	16.60
10	40	0.9	4.40	6.64	7.64	5.64	4.96	6.64	8.24	5.40	8.84	12.68	15.36	14.48
16	102	0.0	4.44	5.92	5.68	8.00	5.32	6.32	5.84	8.04	6.84	8.92	10.16	13.32
16	102	0.5	4.24	5.32	5.72	8.48	4.72	6.04	6.88	8.88	6.08	9.48	11.56	13.16
16	102	0.9	4.64	4.68	5.24	6.32	5.20	6.16	8.36	6.48	7.32	8.24	11.08	11.32
20	160	0.0	4.36	5.68	4.52	8.88	4.80	5.68	5.36	8.72	6.36	9.16	9.68	12.28
20	160	0.5	4.08	5.32	6.28	8.32	4.36	5.32	7.64	8.20	6.00	8.08	10.24	13.36
20	160	0.9	5.04	4.44	5.48	5.64	5.56	4.92	7.08	6.92	6.76	7.88	10.80	9.96
25	250	0.0	5.12	4.60	4.96	7.96	5.08	5.52	6.00	8.08	5.96	7.32	8.44	12.08
25	250	0.5	4.96	5.20	5.16	8.76	5.04	5.76	6.24	8.48	5.92	8.00	8.96	10.84
25	250	0.9	5.16	5.48	4.92	5.40	5.32	6.84	7.04	5.96	5.32	7.56	8.32	9.48

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

When the true 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, then the spatial forecast accuracy test is correctly sized. In comparing test statistics S_C and S_V in Tables 10 and 11, estimation of the semivariogram produces a more accurate estimate of the variance of $\bar{D}$, resulting in empirical sizes that are much closer to α. In simulations, the estimated parametric covariogram on average underestimated the true covariogram across all lag distances. The estimated parametric semivariogram did not suffer from such a problem, so all results subsequent to these will be based on S_V. Upon examining Table 11, several points become clear.

- The contemporaneous correlation appears to have little influence on the size of the test.
- The size 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.
- When the spatial ranges of the errors differ, the size is larger than when the spatial correlation is the same for both sets of forecast errors.
- As the sample size increases, the size of the test improves.
- The size is also influenced by the type of loss function that is used. The simple loss performs much worse than the quadratic or absolute losses.

The effect of the quadratic loss on the spatial correlation can be explained theoretically.

Proposition 2. *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 Corr$(X^2, Y^2) = \rho^2$.

In other words, when two Gaussian random variables whose correlation to each other is ρ are squared, the correlation between the squared variables is ρ^2. The derivation of this result is given in the Appendix. Thus, the positive spatial correlation is reduced under the quadratic loss. A similar effect occurs for a general loss $g(\cdot)$, as stated in this proposition.

Proposition 3. *Let* $\mathbf{Z} = (X, Y)^T$ *be a bivariate random vector with mean* $\boldsymbol{\mu} = (\mu_x, \mu_y)^T$ *and covariance matrix,* Σ. *When the first and second derivatives of* $g(\cdot)$ *exist,*

$$\textit{Corr}(g(X), g(Y)) \approx \frac{g'(0)^2\rho\sigma_x\sigma_y + 2g''(0)^2\rho^2\sigma_x^2\sigma_y^2}{\sqrt{g'(0)^2\sigma_x^2 + 2g''(0)^2\sigma_x^4}\sqrt{g'(0)^2\sigma_y^2 + 2g''(0)^2\sigma_y^4}},$$

which reduces to ρ^2 *when* $g'(0) = 0$.

The proof is given in the Appendix. The absolute loss is not twice differentiable, but its form is still very similar to the quadratic loss' form, and the spatial correlation will also be reduced by it.

The power of the test using the test statistic S_V is given in Figure 13 for all combinations of ρ and θ_i and θ_j in grid sizes 10×10, 16×16, and 20×20. The mean, $f(\mathbf{s}) = \mu$, is allowed to vary from 0 to 7 in increments of 0.5. From these figures, we see that

- For a given value of μ, the power increases with an increase in sample size.
- The power reaches (or nearly reaches) 100% when $\mu = 4, 2, 1$ for grid sizes 10, 16, and 20, respectively.
- 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.

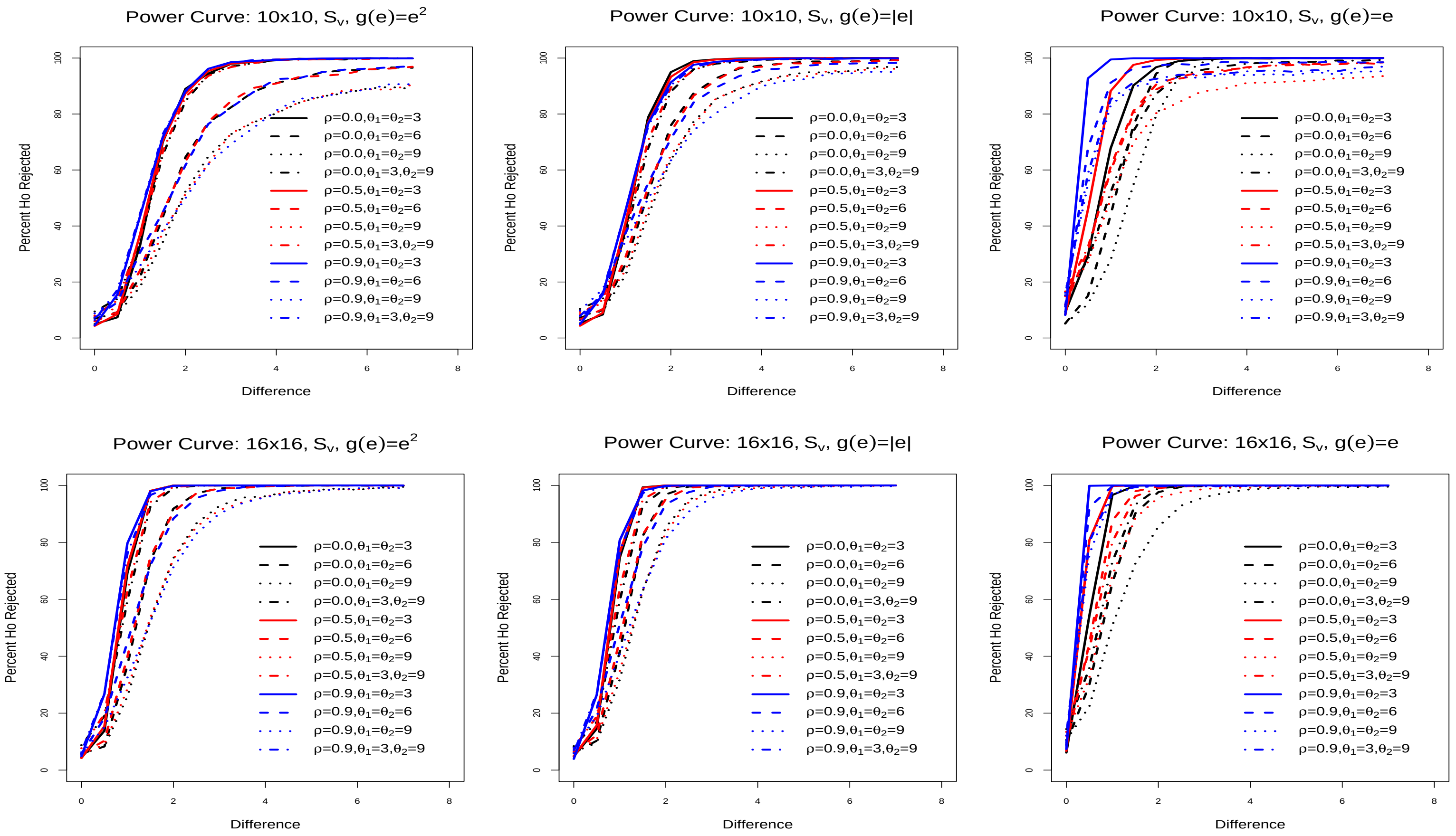


Figure 13: Power curves for 10×10 grid (top row) and 16×16 grid (bottom row). First column is quadratic loss, second column is absolute loss, and last column is the simple loss.

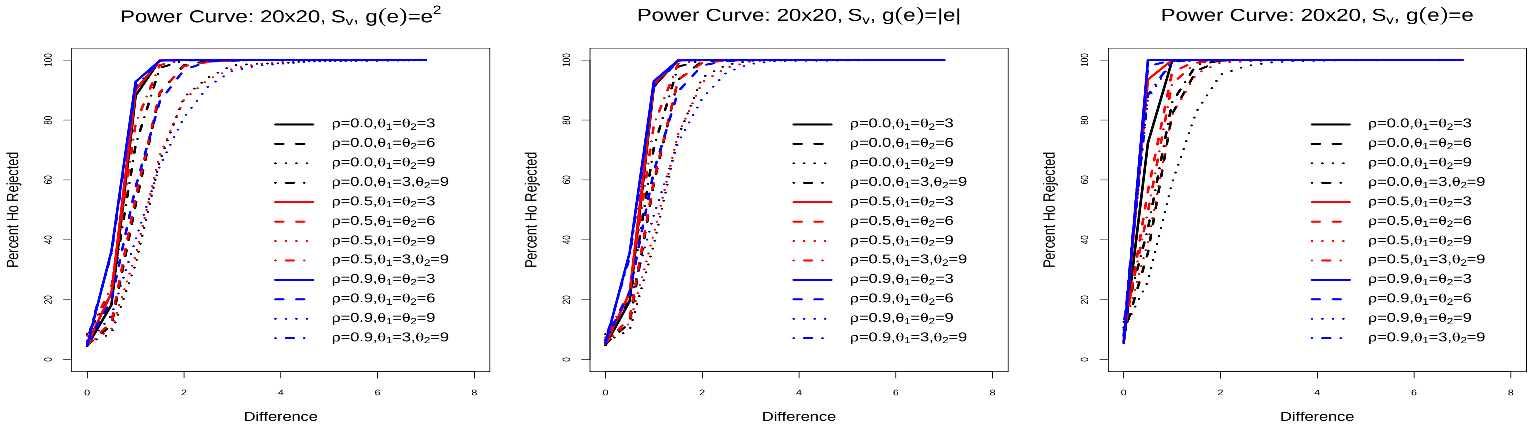


Figure 13: Continued. Power curves for 20×20 grid. Left plot is quadratic loss, middle plot is absolute loss, and right plot is the simple loss.

4.5.3 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. Under the null hypothesis, it should be noted that estimating the trend with a function that is linear in the coordinates will simply return a value close to the mean of the field, and results similar to those in Table 11 would be expected. Recall that the bandwidth for nonparametric estimation must be adjusted to account for the spatial correlation in $\delta(\mathbf{s})$. The results in Table 12 illustrate the effect on the size of the test when the bandwidth is not adjusted for the spatial correlation. 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 13 shows how much improvement is gained using the test statistic S_V^r when the bandwidth is adjusted, even with a rudimentary iterative method. The size of the test still becomes worse as the spatial correlation increases, but it does well, even in small samples, when the spatial correlation is low, and the size improves as the sample size increases. Similar to the outcome when $f(\mathbf{s})$ is constant, the size of the test when using S_C^r is still too large (results not shown).

Table 12: Empirical size of loss functions under the weighted semivariogram estimate of variance of $\bar{D}$ for the spatial accuracy test and unadjusted bandwidth used in the nonparametric trend estimation. All tests are reported at the 5% level, and 2,500 Monte Carlo replications are performed.

			Quadratic Loss				Absolute Loss				Simple Loss			
Grid	L	ρ	$\theta_{i,j} = 3$	$\theta_{i,j} = 6$	$\theta_{i,j} = 9$	$\theta_{i,j} = 3, 9$	$\theta_{i,j} = 3$	$\theta_{i,j} = 6$	$\theta_{i,j} = 9$	$\theta_{i,j} = 3, 9$	$\theta_{i,j} = 3$	$\theta_{i,j} = 6$	$\theta_{i,j} = 9$	$\theta_{i,j} = 3, 9$
5	10	0.0	16.04	30.84	39.16	27.28	17.12	31.36	40.00	29.40	33.44	57.88	66.52	51.64
5	10	0.5	16.04	29.84	40.72	25.16	16.84	29.32	40.32	26.32	32.88	55.24	65.44	58.32
5	10	0.9	15.72	32.08	39.08	24.80	16.84	31.76	37.68	24.52	34.84	56.40	64.68	65.04
8	25	0.0	10.84	27.64	46.44	27.92	10.84	26.52	44.44	27.00	33.16	63.16	75.20	58.48
8	25	0.5	11.16	30.04	46.84	25.88	11.12	29.92	44.32	24.36	33.96	65.20	75.48	68.92
8	25	0.9	11.84	29.28	45.44	20.56	11.40	27.04	41.24	19.16	33.56	64.20	74.40	73.76
10	40	0.0	10.00	30.64	50.16	29.88	9.96	29.68	47.32	28.96	33.12	67.04	80.40	63.44
10	40	0.5	10.12	30.68	48.80	27.00	10.32	28.04	46.56	25.04	33.36	69.92	79.08	74.56
10	40	0.9	10.12	29.72	50.36	22.52	9.44	25.76	45.44	19.00	34.00	68.24	80.72	76.48
16	102	0.0	8.28	33.68	55.68	32.32	8.08	29.52	51.72	30.72	34.92	74.32	83.04	70.48
16	102	0.5	7.20	32.88	57.44	28.52	8.04	28.56	53.20	25.92	35.44	76.44	84.12	79.16
16	102	0.9	8.44	34.04	55.64	23.08	7.08	27.28	47.68	18.96	35.68	73.60	84.84	83.36
20	160	0.0	8.40	35.08	57.80	33.64	7.72	32.24	54.80	32.04	37.44	74.96	83.20	71.60
20	160	0.5	7.08	34.72	59.84	31.40	6.96	30.04	54.72	27.80	37.68	75.92	83.92	80.32
20	160	0.9	8.00	35.20	59.48	21.84	6.84	26.80	49.64	17.52	37.76	75.08	84.84	85.28

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

Table 13: Empirical size of loss functions under the weighted semivariogram estimate of variance of $\bar{D}$ for the spatial accuracy test and iteratively estimated adjusted bandwidth used in the nonparametric trend estimation. All tests are reported at the 5% level, and 2,500 Monte Carlo replications are performed.

			Quadratic Loss				Absolute Loss				Simple Loss			
Grid	L	ρ	$\theta_{i,j} = 3$	$\theta_{i,j} = 6$	$\theta_{i,j} = 9$	$\theta_{i,j} = 3, 9$	$\theta_{i,j} = 3$	$\theta_{i,j} = 6$	$\theta_{i,j} = 9$	$\theta_{i,j} = 3, 9$	$\theta_{i,j} = 3$	$\theta_{i,j} = 6$	$\theta_{i,j} = 9$	$\theta_{i,j} = 3, 9$
5	10	0.0	10.90	21.08	26.32	19.00	12.16	21.56	26.64	20.80	23.76	41.20	48.20	38.00
5	10	0.5	11.16	20.00	28.48	17.64	12.84	20.60	27.60	18.44	22.88	40.52	46.52	43.40
5	10	0.9	11.04	22.36	26.92	17.52	12.16	22.56	25.60	17.12	24.72	40.52	45.80	46.92
8	25	0.0	6.28	14.56	23.60	15.56	6.80	13.68	22.84	14.92	17.12	33.12	41.44	32.80
8	25	0.5	6.48	15.72	23.60	14.08	6.88	15.04	22.76	13.36	18.20	34.36	41.88	39.12
8	25	0.9	7.32	15.20	23.08	11.00	6.76	13.36	21.68	10.52	17.40	34.80	40.92	39.52
10	40	0.0	6.40	13.48	22.44	15.96	6.40	12.80	21.52	15.48	15.00	32.36	42.48	35.72
10	40	0.5	5.84	14.04	21.52	14.40	6.32	13.12	20.68	13.28	15.16	34.64	43.80	41.24
10	40	0.9	6.32	13.60	23.36	10.76	5.88	12.00	20.96	9.80	15.12	33.08	43.72	41.68
16	102	0.0	6.00	12.20	22.08	16.28	5.76	11.56	19.72	14.60	13.52	32.88	43.76	36.48
16	102	0.5	5.00	13.44	22.20	12.92	5.28	11.24	21.28	12.52	12.92	34.00	42.88	41.72
16	102	0.9	5.64	12.84	21.72	9.72	5.00	10.76	17.76	8.68	13.72	34.12	41.64	43.28
20	160	0.0	5.32	11.44	20.04	14.08	5.92	10.84	20.00	14.28	13.24	31.60	40.68	36.16
20	160	0.5	5.20	12.32	21.56	12.92	5.28	11.36	20.64	12.72	13.36	31.16	41.44	39.84
20	160	0.9	5.40	12.24	20.96	10.00	5.40	10.76	18.60	9.48	13.12	30.76	42.52	43.24

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

Many types of spatially varying means for the alternative hypothesis could be imagined. For a 16×16 size grid with $\rho = 0.50$ and $\theta_1 = \theta_2 = 6$, three different types of trends for $f(\mathbf{s})$ will be examined, a random, a split, and a linear trend. For the random trend, a set of A locations are randomly selected, and the mean function is

$$f(\mathbf{s}) = \begin{cases} v, & \mathbf{s} \in A, \\ 0, & \text{otherwise}, \end{cases}$$

for v the value of the mean at location $\mathbf{s}_i$. Without any pattern, a random trend will be difficult to estimate, but as the set A grows, it should become easier to detect a nonzero averaged difference field. See Figure 14 for an example plotted without noise where t, the intensity, is 4. The split mean function at location $\mathbf{s} = (x, y)$ is

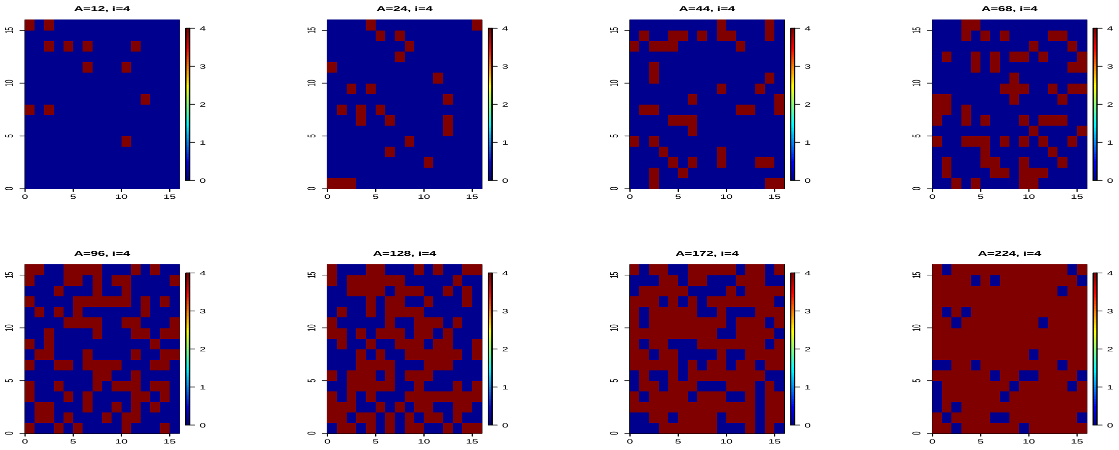


Figure 14: Examples of a trend with randomly distributed cells with nonzero means. The number of affected cells, A varies between 12, 24, 44, 68, 96, 128, 172, and 224. Shown here is intensity level 4 without noise, but the intensity is allowed to vary through the values $2, 3, 4,$ and 5.

$$f(\mathbf{s}) = \begin{cases} v, & x \leq 8, \\ -v, & x > 8, \end{cases}$$

and Figure 15 illustrates this trend. A nonparametric estimate may capture this trend effectively, but a linear fitted trend will simply estimate the overall mean. Finally, linear trends with varying coefficients, $f(\mathbf{s}) = A + Bx + Cy$, are plotted in Figure 16. Of course, fitting a linear trend should work well for such a mean function, but the nonparametric estimate should do well also since the trend is a smooth function.

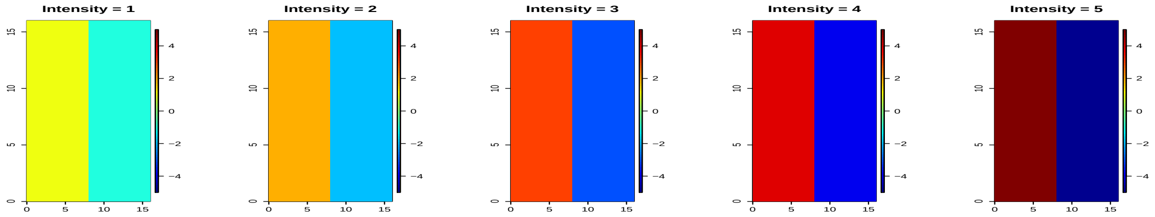


Figure 15: Example of data simulated without noise in which the mean on the left-hand side is t, and the mean on the right-hand side is $-t$.

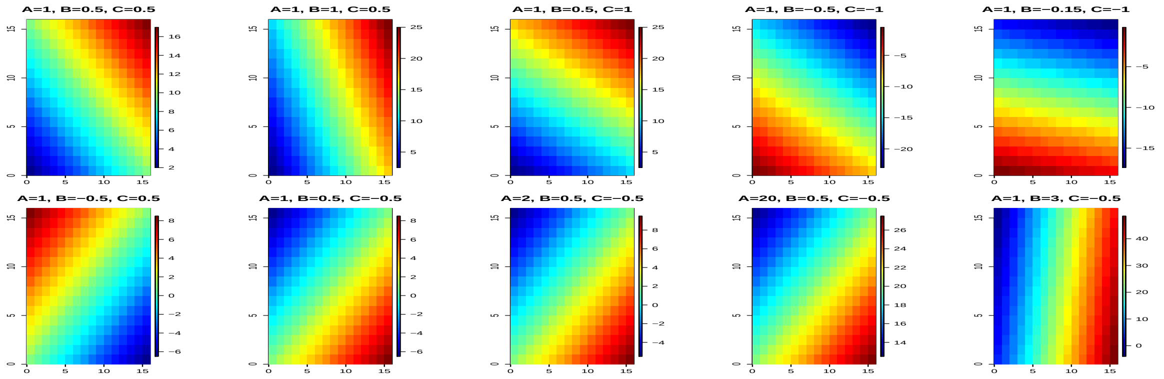


Figure 16: Linear trend patterns without noise that are used for the linear trend simulation. The pattern follows the formula $f(\mathbf{s}) = A + Bx + Cy$.

The results for the random trend are given in Table 14. For comparison, the true trend is removed as if it were known so that various methods of trend removal can be fairly evaluated. When no trend is removed, the test does not reject the null as often as it should, particularly when A is small or the intensity, v, is low. These low percentages of rejections are due to the fact that these locations with non-zero means are spuriously increasing the variance of $D(\mathbf{s})$. Removing a linear trend and an iteratively reweighted generalized least squares trend (IRWGLS) yield very similar results to having removed no trend at all. This is not surprising since the trend is not linear. Finally, the nonparametric trend with the adjusted bandwidth has powers that are closer to those obtained when the true trend is removed, particularly as L increases. But, this type of trend will not be estimated very well by the nonparametric procedures for low values of L since it does not vary smoothly.

In the split pattern (results given in Table 15), it should be noted that under the simple loss that the null hypothesis is actually true. The positive and negative values will sum to zero, so the expected outcome for the simple loss for any value of v is the size of the test. Based on results in Table 11 when $\theta_1 = \theta_2 = 6$, the empirical size of the test will be around 10%. However, in Table 15 when the true mean is removed, the sizes grow 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%. Of all the trend removal techniques tested, the size with the nonparametric trend removed is the closest to what is expected under the null. Without any trend removed, the simple loss is undersized since variability in the values of $D(\mathbf{s})$ is high relative to the value of $\bar{D}$. This is a classic example that illustrates how the forecast accuracy test is only designed to detect a difference in two sets of spatial forecasts on average and will not be successful in detecting local differences between two sets of forecasts.

Table 14: Percent of null hypotheses rejected in 2,500 simulated datasets with random trend for the quadratic, absolute, and simple losses using weighted semivariogram estimator.

	Quadratic				Absolute				Simple			
	True Trend Removed											
	$v=2$	$v=3$	$v=4$	$v=5$	$v=2$	$v=3$	$v=4$	$v=5$	$v=2$	$v=3$	$v=4$	$v=5$
$A=12$	14.72	36.96	65.28	82.64	10.72	17.40	29.12	40.80	10.72	11.84	16.84	18.00
$A=24$	32.28	72.56	93.36	98.80	21.16	44.36	65.68	81.88	14.88	21.48	31.40	40.08
$A=44$	61.36	95.12	99.44	99.96	43.20	82.36	94.44	98.88	28.20	45.76	63.84	77.80
$A=68$	82.04	99.16	100.0	100.0	70.72	97.12	99.72	99.92	47.48	73.32	89.56	95.52
$A=96$	93.52	99.88	100.0	100.0	88.60	99.36	99.96	100.0	70.44	91.92	98.24	99.60
$A=128$	97.68	99.96	100.0	100.0	96.40	99.92	100.0	100.0	86.84	98.00	99.56	99.80
$A=172$	99.12	100.0	100.0	100.0	98.68	100.0	100.0	100.0	96.36	99.68	99.84	100.0
$A=224$	99.60	100.0	100.0	100.0	99.64	100.0	100.0	100.0	99.04	99.92	100.0	99.96
	No Trend Removed											
	$v=2$	$v=3$	$v=4$	$v=5$	$v=2$	$v=3$	$v=4$	$v=5$	$v=2$	$v=3$	$v=4$	$v=5$
$A=12$	12.16	20.32	32.96	41.08	11.36	15.32	22.00	27.68	13.12	16.68	23.52	27.00
$A=24$	20.88	45.84	67.28	80.16	18.24	33.96	49.80	60.56	17.36	26.52	38.00	44.48
$A=44$	49.08	85.68	96.68	98.84	41.92	75.32	91.16	97.12	35.68	57.40	74.40	85.40
$A=68$	71.20	96.80	99.68	99.92	67.00	95.20	99.52	100.0	58.88	85.84	95.96	98.36
$A=96$	86.80	99.08	99.96	100.0	86.80	99.32	99.96	100.0	82.60	98.04	99.88	100.0
$A=128$	91.16	99.56	100.0	100.0	92.92	99.88	100.0	100.0	94.32	99.84	100.0	100.0
$A=172$	94.12	99.88	100.0	100.0	95.68	100.0	100.0	100.0	99.44	100.0	100.0	100.0
$A=224$	94.28	99.56	100.0	100.0	96.68	99.92	100.0	100.0	99.84	100.0	100.0	100.0
	Linear Trend Removed											
	$v=2$	$v=3$	$v=4$	$v=5$	$v=2$	$v=3$	$v=4$	$v=5$	$v=2$	$v=3$	$v=4$	$v=5$
$A=12$	14.08	23.68	36.72	45.48	13.36	17.96	25.00	31.12	19.08	22.40	29.04	32.24
$A=24$	23.64	49.44	71.80	83.28	20.20	36.96	53.96	63.92	24.04	32.84	44.04	49.48
$A=44$	51.60	87.76	97.00	99.16	44.64	78.08	92.28	97.52	43.12	63.88	78.48	87.84
$A=68$	74.64	97.72	99.84	99.96	70.64	96.28	99.72	100.0	66.68	89.12	97.04	99.00
$A=96$	89.48	99.40	99.96	100.0	88.56	99.56	100.0	100.0	87.04	98.72	99.96	100.0
$A=128$	94.20	99.84	100.0	100.0	95.20	99.92	100.0	100.0	96.72	99.96	100.0	100.0
$A=172$	96.16	100.0	100.0	100.0	97.24	100.0	100.0	100.0	99.68	100.0	100.0	100.0
$A=224$	96.96	99.96	100.0	100.0	98.16	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	IRWGLS Trend Removed											
	$v=2$	$v=3$	$v=4$	$v=5$	$v=2$	$v=3$	$v=4$	$v=5$	$v=2$	$v=3$	$v=4$	$v=5$
$A=12$	13.96	23.60	36.68	45.40	13.24	17.84	24.80	31.00	18.00	21.32	28.36	31.64
$A=24$	23.52	49.08	71.60	83.08	19.96	36.64	53.52	63.84	23.08	31.68	43.56	49.12
$A=44$	51.12	87.44	96.92	99.12	44.40	77.84	92.28	97.44	41.88	63.04	77.84	87.52
$A=68$	74.44	97.68	99.84	99.96	70.28	96.12	99.72	100.0	65.80	89.00	96.96	99.00
$A=96$	89.16	99.32	99.96	100.0	88.36	99.56	99.96	100.0	86.16	98.60	99.96	100.0
$A=128$	93.56	99.76	100.0	100.0	94.64	99.92	100.0	100.0	96.28	99.92	100.0	100.0
$A=172$	95.36	99.96	100.0	100.0	96.80	100.0	100.0	100.0	99.52	100.0	100.0	100.0
$A=224$	95.96	99.76	99.84	99.92	97.64	100.0	100.0	100.0	99.96	100.0	100.0	100.0
	Nonparametric Trend Removed											
	$v=2$	$v=3$	$v=4$	$v=5$	$v=2$	$v=3$	$v=4$	$v=5$	$v=2$	$v=3$	$v=4$	$v=5$
$A=12$	15.60	23.08	35.52	44.96	14.76	20.40	26.44	31.16	32.76	33.40	34.76	34.88
$A=24$	24.72	48.84	70.44	82.96	21.96	38.84	54.44	63.40	34.80	41.32	46.92	54.08
$A=44$	52.88	88.72	97.48	99.28	45.52	78.48	92.28	97.68	53.76	67.32	79.12	88.92
$A=68$	76.84	98.08	99.88	99.88	72.40	97.00	99.84	99.92	73.00	89.68	96.80	99.24
$A=96$	91.28	99.68	99.96	100.0	90.24	99.76	100.0	100.0	90.04	98.68	99.84	100.0
$A=128$	96.08	99.92	100.0	100.0	96.80	100.0	100.0	100.0	97.36	100.0	100.0	100.0
$A=172$	98.24	100.0	100.0	100.0	98.72	100.0	100.0	100.0	99.92	100.0	100.0	100.0
$A=224$	99.08	99.92	100.0	100.0	99.52	99.96	100.0	100.0	100.0	100.0	100.0	100.0

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

Under the quadratic and absolute loss functions in the split pattern, the alternative hypothesis is true. The spatial forecast accuracy test does very well for these losses but can be strongly undersized when $v = 1$. If the left/right pattern is assumed known, and the mean on each side of the domain is found and used to detrend $D(\mathbf{s})$, then the results are similar to knowing the true trend. The linear and IRWGLS trends are undersized for $v = 1$ and $v = 2$. 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.

For the linear trend, results shown in Table 16, removing the correct mean for the mean functions with coefficients given in Table 17 gives very good results. However, ignoring the trend results in the rejection of almost none of the null hypotheses. This simply illustrates how failure to remove a strong trend can negatively influence the test. After the quadratic and absolute losses are applied to the forecast errors, the trend is no longer linear, so a quadratic trend is fit for the quadratic and absolute losses, but a linear trend is still fit for the simple loss. Fitting these types of trends yields results much, much closer to removal of the true mean. The IRWGLS fit (linear for the simple loss and quadratic for the quadratic and absolute losses) is even more conservative, rejecting just a bit less frequently than when the trend is estimated with least squares. Finally, the nonparametric fitted trend does not fare as well as the linear and quadratic trends do.

Table 15: Percent of null hypotheses (under weighted semivariogram estimation) rejected in 2,500 simulated datasets for the split 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	100.0	100.0	100.0
	Absolute	72.96	99.84	100.0	100.0	100.0
No Trend	Simple	0.08	0.00	0.00	0.00	0.00
	Quadratic	42.68	95.60	99.64	99.92	100.0
	Absolute	48.12	97.52	99.92	100.0	100.0
Pattern Known	Simple	9.08	13.12	17.12	23.68	25.92
	Quadratic	76.92	99.84	100.0	100.0	100.0
	Absolute	74.92	99.84	100.0	100.0	100.0
Linear Trend	Simple	6.84	1.20	0.40	0.04	0.00
	Quadratic	52.44	98.16	99.92	100.0	100.0
	Absolute	57.08	99.20	100.0	100.0	100.0
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	100.0
Nonparametric Trend	Simple	21.24	13.44	14.60	16.60	17.64
	Quadratic	69.48	99.56	100.0	100.0	100.0
	Absolute	71.40	99.76	100.0	100.0	100.0

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

Table 16: Percent of null hypotheses (under weighted semivariogram estimation) rejected in 2,500 simulated datasets for the linear trend datasets.

		Combination of Coefficients									
	Loss	1	2	3	4	5	6	7	8	9	10
Right Mean	Simple	100	100	100	100	100	83.68	84.16	99.48	100	100
	Quadratic	100	100	100	100	100	100	100	100	100	100
	Absolute	100	100	100	100	100	100	100	100	100	100
Wrong Mean	Simple	0	0	0	0	0	0	0	0	0	0
	Quadratic	0.04	0.04	0.08	0	0	0	0	0	0.08	0.4
	Absolute	0	0	0	0	0	0.28	0.36	0.04	0.28	0
Linear Mean	Simple	100	100	100	100	100	90.60	90.64	99.96	100	100
Quadratic Mean	Quadratic	100	100	100	100	100	100	99.96	100	100	100
Quadratic Mean	Absolute	100	100	100	100	100	100	100	100	100	100
IRWGLS Mean	Simple	100	100	100	100	99.96	88.20	87.52	99.48	99.96	100
	Quadratic	99.8	99.96	99.88	99.68	99.84	98.44	97.60	98.08	99.96	99.92
	Absolute	99.96	100	100	100	100	99.04	98.76	97.40	100	97.16
Nonparametric Mean	Simple	38.52	5.20	4.20	4.92	8.16	16.72	18.16	25.44	40.44	1.28
	Quadratic	56.28	31.36	29.40	24.80	36.60	28.52	30.00	33.64	71.80	73.28
	Absolute	56.24	13.60	12.64	12.36	19.64	33.68	33.56	40.72	60.36	2.20

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

Table 17: Coefficients for the linear trend simulation results given in Table 16.

#	A	B	C
1	1	0.5	0.5
2	1	1	0.5
3	1	0.5	1
4	1	-0.5	-1
5	1	-0.15	-1
6	1	-0.5	0.5
7	1	0.5	-0.5
8	2	0.5	-0.5
9	20	0.5	-0.5
10	1	3	-0.5

4.5.4 Comparison with SHC Method

It should be noted that when $f(\mathbf{s})$ is constant, then the SHC method and the spatial forecast accuracy test are testing the same null hypothesis that the constant mean of the spatial process is zero. To compare the two methods in this setting, we need to make adjustments to the spatial forecast accuracy 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 the spatial forecast accuracy test are applied to the quadratic, absolute, and simple loss differentials. The results are given in Figure 17. It is immediately evident that when $\mu = 0$, which is the null hypothesis, that 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.

Shen et al. (2002) only test their method in simulations with normally distributed data and spatial independence. The data that is generated in this simulation is multivariate

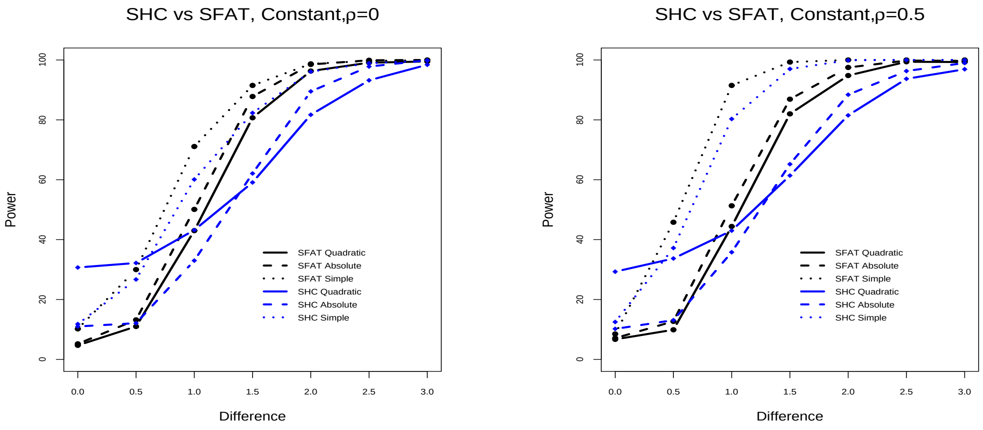


Figure 17: Comparison of power between the SHC test and the spatial forecast accuracy test (SFAT) when the trend is constant.

normal, but once the quadratic and absolute loss functions are applied to the errors, the data is no longer normal. In fact, for interesting loss functions, the resulting loss differential will rarely be normally distributed. Therefore, in the context of comparing the accuracy of spatial predictions, many modifications would need to be made for the SHC method to be generally applicable.

4.6 Oklahoma Wind Speed Example

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 forecast, 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. Two spatial models are built based on 70 locations to forecast the daily average wind speed at 46 reserved locations. Figure 18 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.

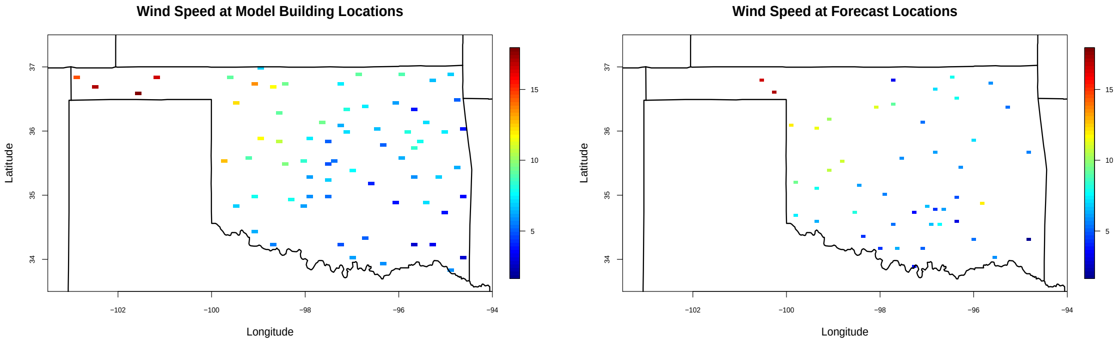


Figure 18: Average daily wind speeds in miles per hour at 116 locations in Oklahoma on September 10, 2008. On the left are the 70 locations used to build the spatial models. On the right are the 46 locations where predictions are made.

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 with an exponential covariance with a nugget. Parameters are estimated in both cases using an iteratively reweighted generalized least squares procedure described in Schabenberger and Gotway (2005). The preceding 3 years of daily average wind speed data before September 10, 2008 is used to build a time series model, T, at each of the 46 locations where a forecast is desired. At each location, a smoothed monthly mean and a smoothed monthly standard deviation is used to standardize the data. These smoothed values are obtained by regressing the monthly means on a pair of harmonics. Then, the order, p, of an AR(p) model is selected with BIC, and parameters are estimated for the selected order.

Forecasts are made at each of the 46 locations based on these three models. These forecasts are compared to the observed average wind speeds using Mean Squared Error (MSE) and Power Curve Error (PCE). The PCE was introduced by Hering and Genton

Table 18: MSE and PCE of each set of forecasts for the Oklahoma wind speed dataset.

Forecast	MSE	PCE
TS	5.29	121.81
S1	4.01	97.79
S2	2.51	72.84

(2009) as a more realistic assessment of wind speed forecasts in the context of wind power generation. It incorporates not only information about the power curve that transforms wind speed observations to wind power but also allows the user to specify an asymmetric penalty for overestimation versus underestimation. It is also an example of a loss function that cannot be written in terms of the forecast errors alone. Table 18 gives the values of MSE and PCE for each of the three models. Spatial model S2 produces forecasts with the smallest MSE and PCE, and the time series forecasts have the largest MSE and PCE.

The top left-hand plots in Figures 19 through 24 show the differences in the squared errors or power curve errors at each location comparing the time series forecasts with the S1 forecasts and with the S2 forecasts and also comparing the S1 and S2 forecasts to each other. 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 curve; the top right-hand plots in Figures 19 through 24 show 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-hand plots). The biggest difference between the forecasts of the time series and spatial models appears to occur in the northwestern region of the state, and the time series forecasts only has smaller errors in isolated regions of the state. The spatial models' forecasts differ the most in the southeastern part of the state. Finally, the empirical semivariogram of the detrended difference field is computed, and a Gaussian covariance is fitted in all six cases (bottom right-hand plots).

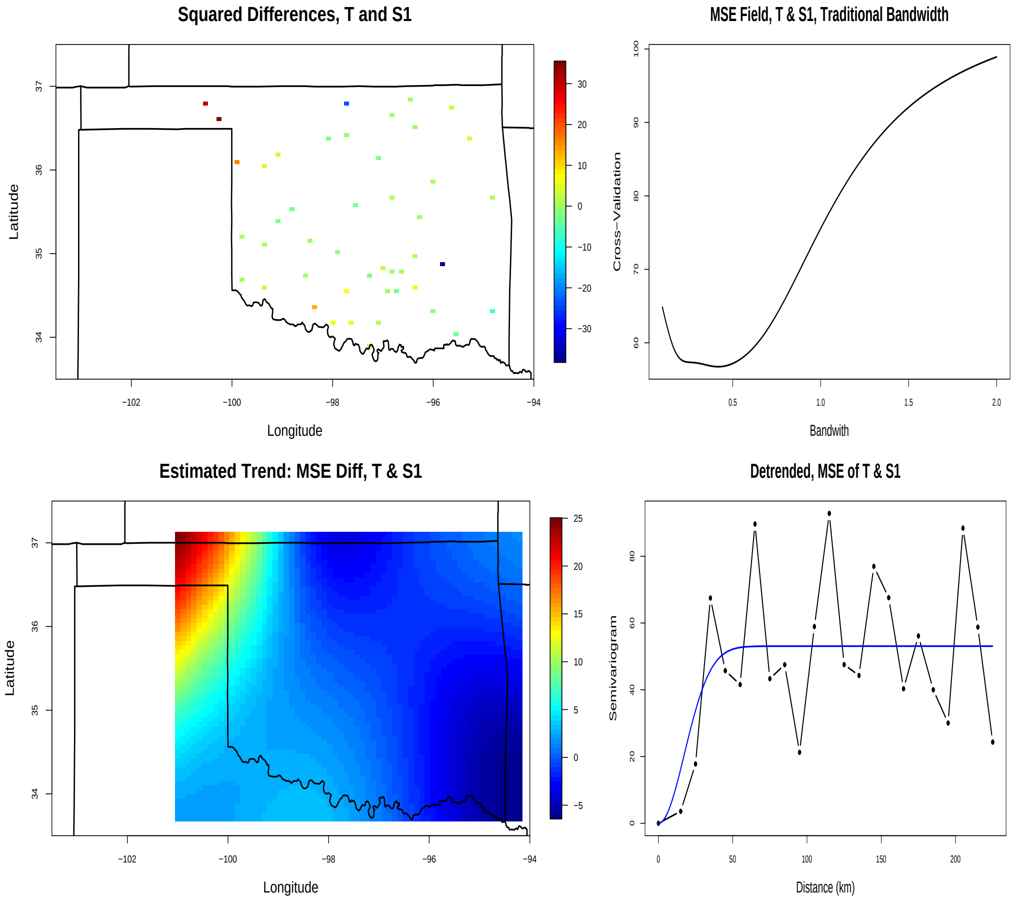


Figure 19: Plots of (upper left) quadratic errors differenced field (time series errors minus S1 errors), (upper right) traditional bandwidth selection ignoring spatial correlation, (lower left) reconstructed spatial field using nonparametric smooth and adjusted bandwidth, and (lower right) empirical semivariogram of detrended field with fitted Gaussian semivariogram overlaid.

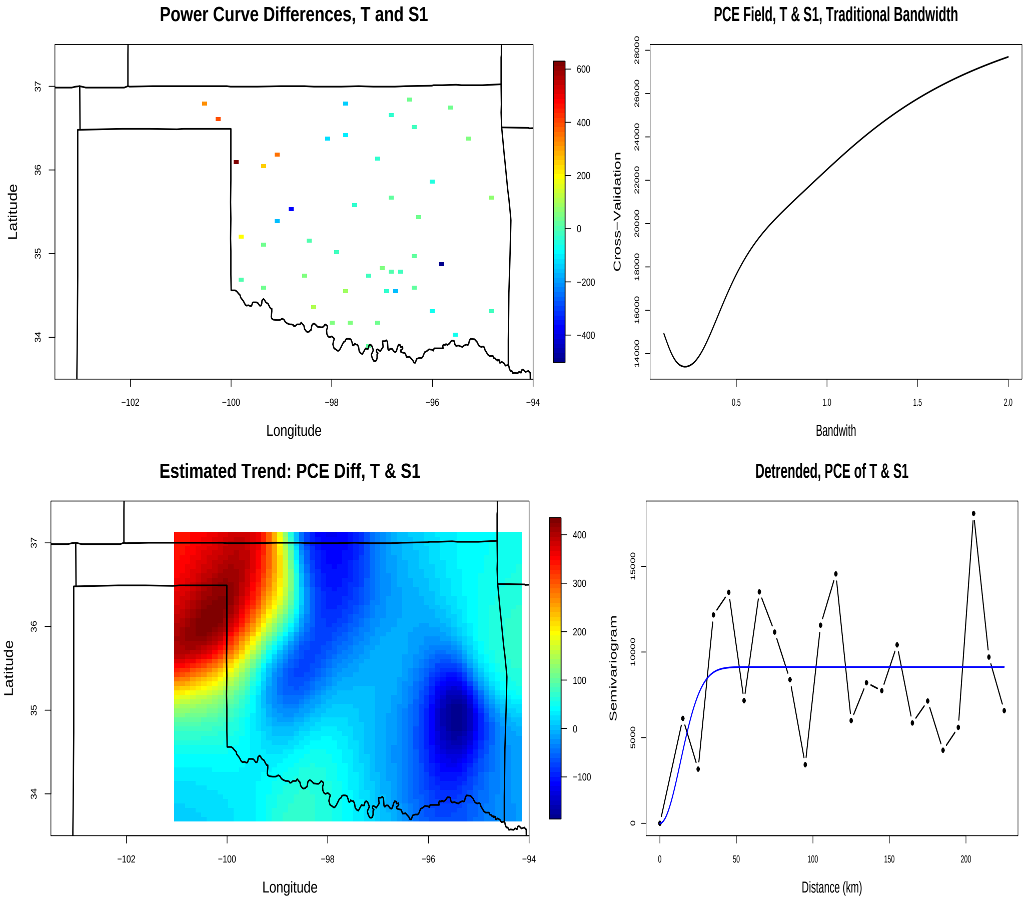


Figure 20: Plots of (upper left) power curve errors differenced field (time series errors minus S1 errors), (upper right) traditional bandwidth selection ignoring spatial correlation, (lower left) reconstructed spatial field using nonparametric smooth and adjusted bandwidth, and (lower right) empirical semivariogram of detrended field with fitted Gaussian semivariogram overlaid.

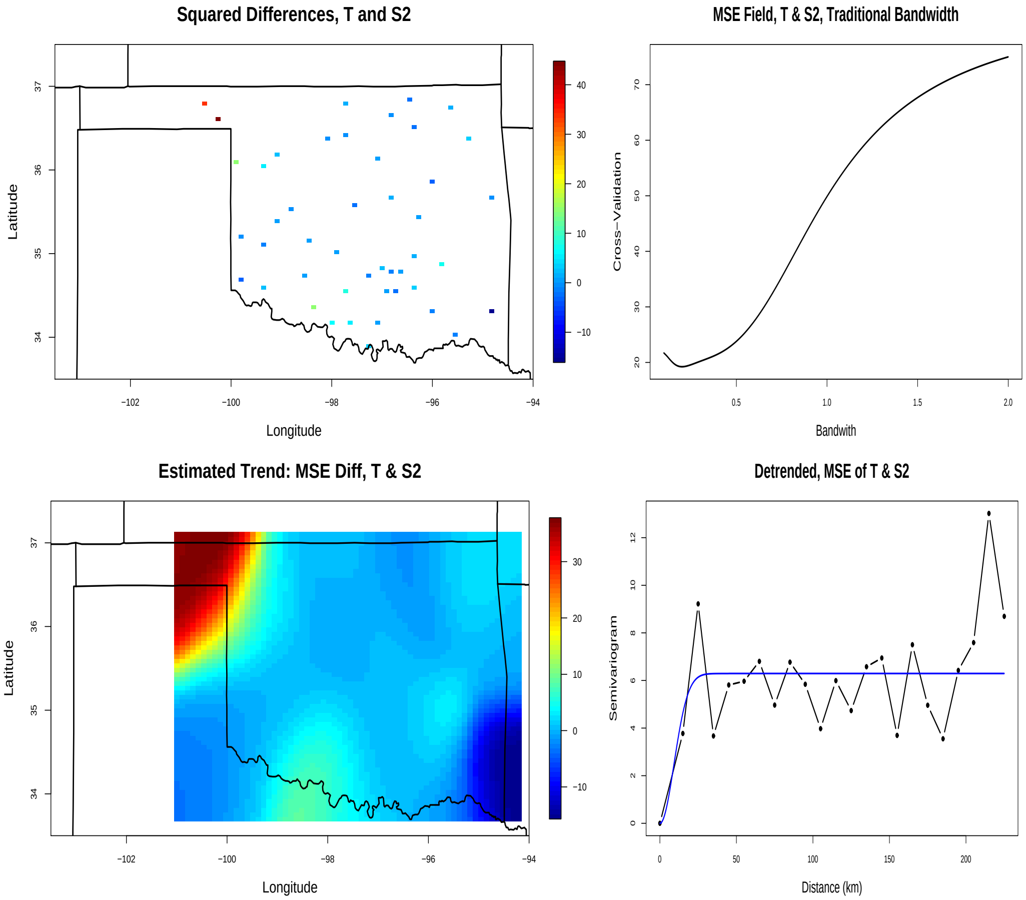


Figure 21: Plots of (upper left) quadratic errors differenced field (time series errors minus S2 errors), (upper right) traditional bandwidth selection ignoring spatial correlation, (lower left) reconstructed spatial field using nonparametric smooth and adjusted bandwidth, and (lower right) empirical semivariogram of detrended field with fitted Gaussian semivariogram overlaid.

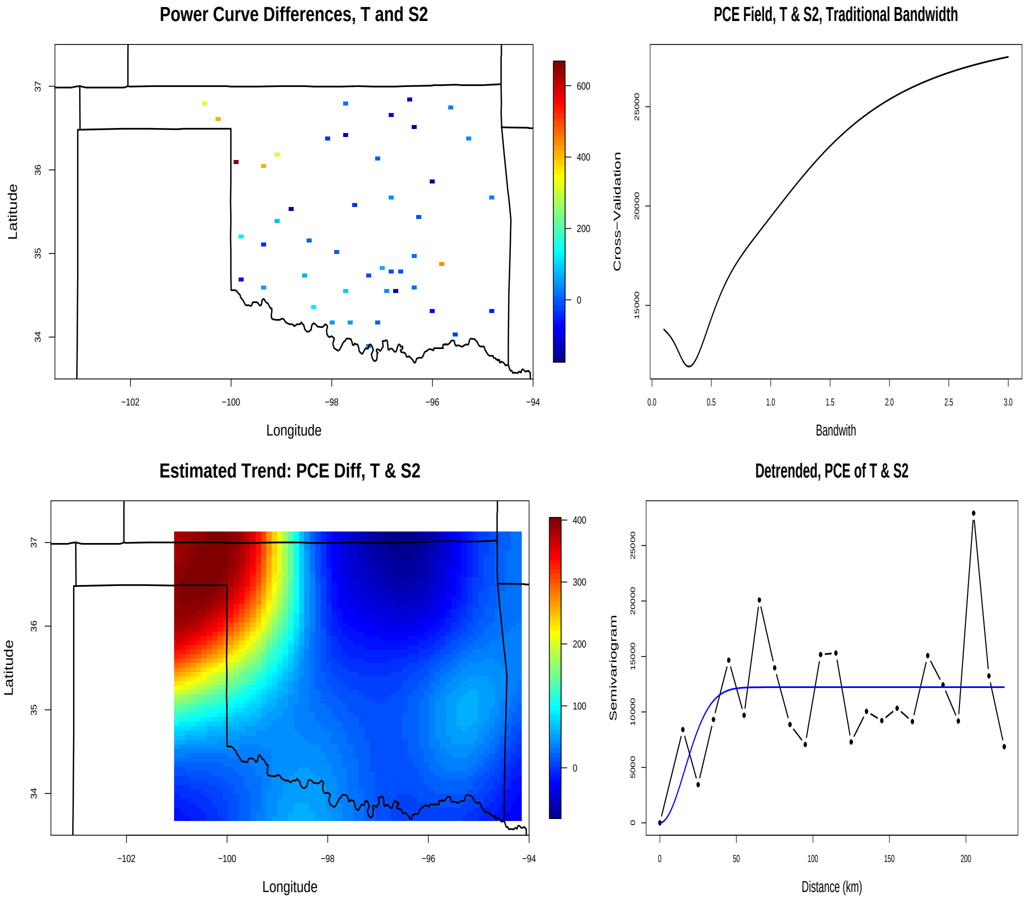


Figure 22: Plots of (upper left) power curve errors differenced field (time series errors minus S2 errors), (upper right) traditional bandwidth selection ignoring spatial correlation, (lower left) reconstructed spatial field using nonparametric smooth and adjusted bandwidth, and (lower right) empirical semivariogram of detrended field with fitted Gaussian semivariogram overlaid.

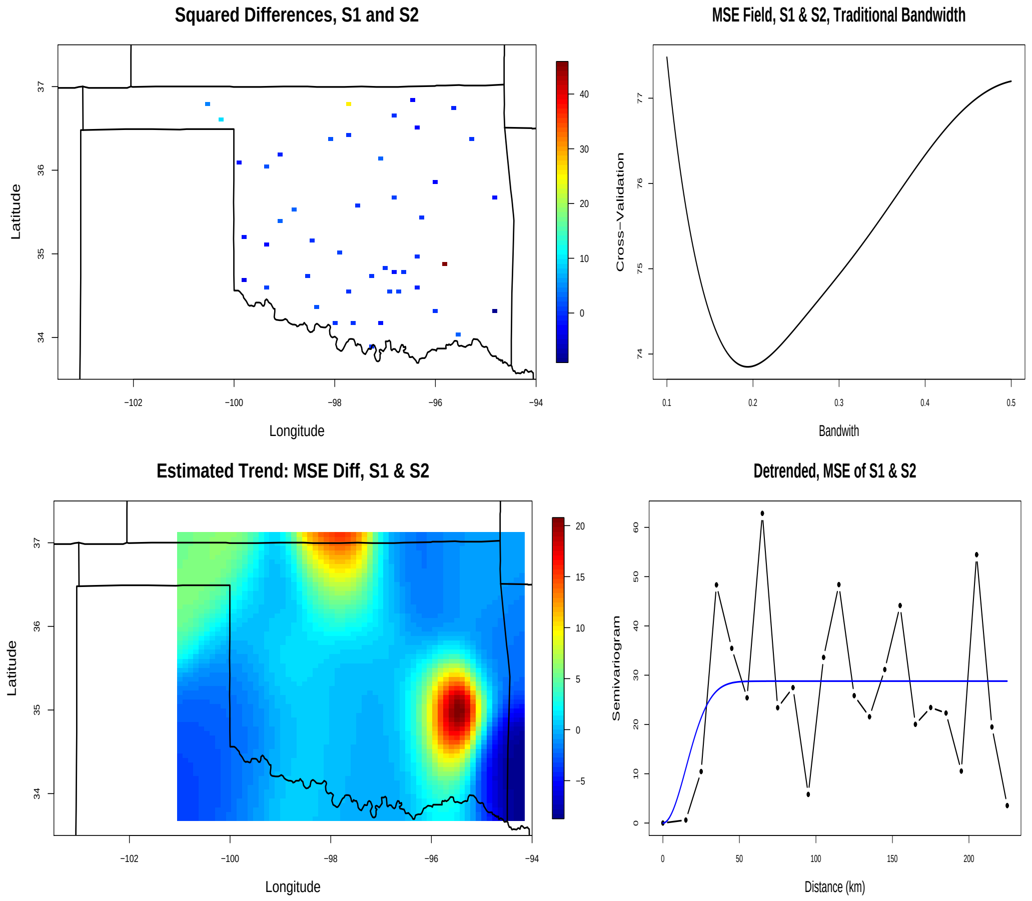


Figure 23: Plots of (upper left) quadratic errors differenced field (S1 errors minus S2 errors), (upper right) traditional bandwidth selection ignoring spatial correlation, (lower left) reconstructed spatial field using nonparametric smooth and adjusted bandwidth, and (lower right) empirical semivariogram of detrended field with fitted Gaussian semivariogram overlaid.

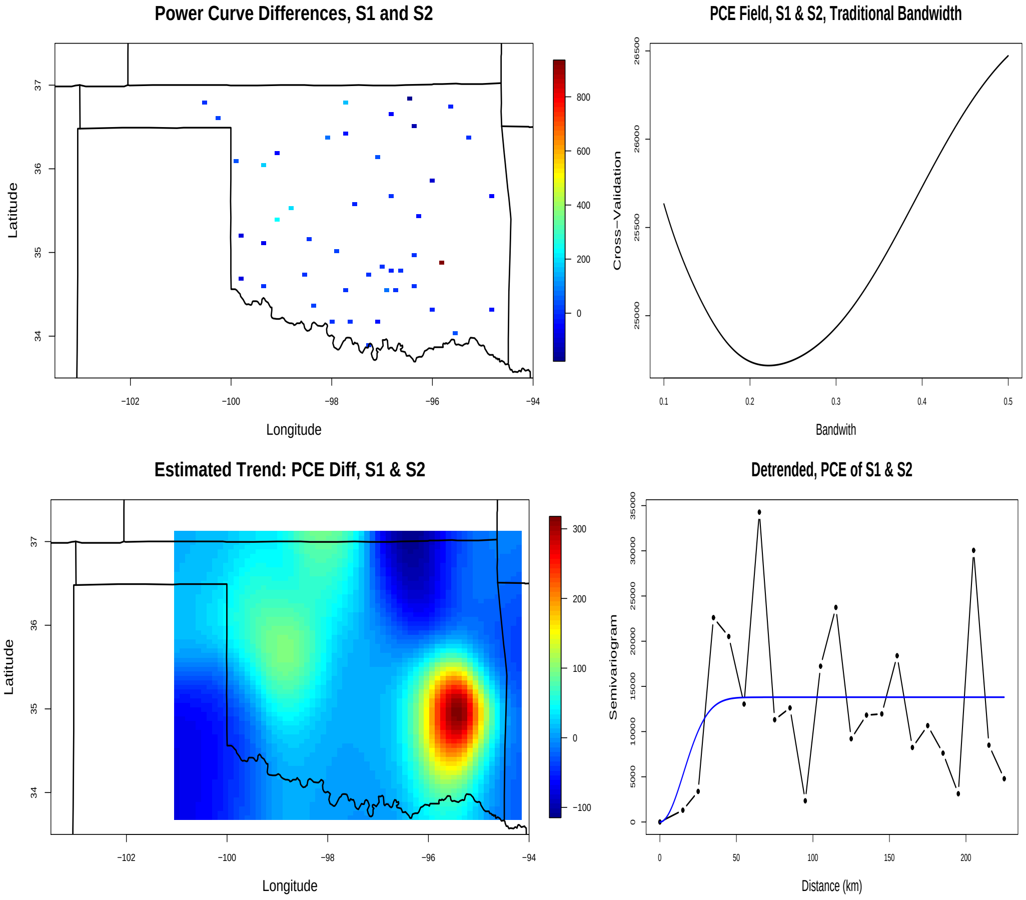


Figure 24: Plots of (upper left) power curve errors differenced field (S1 errors minus S2 errors), (upper right) traditional bandwidth selection ignoring spatial correlation, (lower left) reconstructed spatial field using nonparametric smooth and adjusted bandwidth, and (lower right) empirical semivariogram of detrended field with fitted Gaussian semivariogram overlaid.

Bin centers at 15, 25, and 35 km of the semivariograms estimated from the detrended loss differentials in Figures 19 through 24 all give some indication of spatial dependence for those distances. These values are not spurious, and the data is sufficient to reflect such estimates. Figure 25 shows a histogram of the distances from each forecast location to the closest model building location. These values range from roughly 8 km to 70 km, and seventy-five percent of the distances are less than 40 km. The average of the distances is 34.3 km.

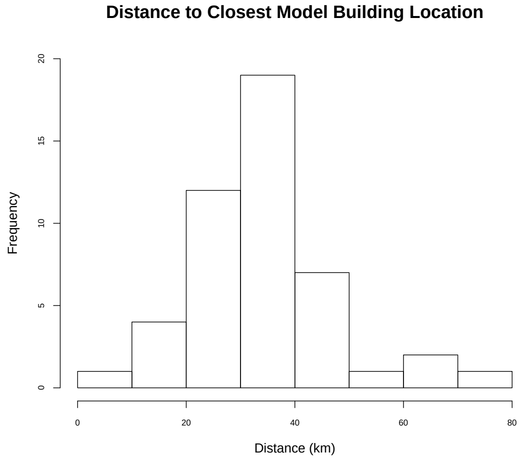


Figure 25: Histogram of the distance from each forecast location to its nearest model building neighbor's location. Seventy-five percent of these 46 distances are less than 40 kilometers.

We obtain estimates of the denominator of the forecast accuracy test, test statistics, and p-values given in Table 19. The time series forecasts and the forecasts produced by model S1 are not significantly different from each other on average in terms of MSE or PCE. Even though S1 has MSE that is 1.275 less than the time series forecasts and has PCE 24.02 less, the variability in the squared errors and the power curve errors is quite large. The S1 and S2 models also do not differ significantly from each other on average in either MSE or PCE. However, the S2 model does produce significantly better forecasts on average in terms of MSE 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 forecast.

Table 19: Comparison of the time series forecast, T, with two sets of spatial forecasts, S1 and S2, for the Oklahoma wind speed dataset.

Comparison	Loss	Numerator	Denominator	Test Statistic	p-value
T versus S1	MSE	1.275	1.214	1.05	0.2935
	PCE	24.02	14.85	1.62	0.1057
T versus S2	MSE	2.774	0.375	7.41	< 0.001
	PCE	48.97	17.99	2.72	0.0065
S1 versus S2	MSE	1.50	0.86	1.75	0.0799
	PCE	25.95	18.88	1.32	0.1864

4.7 Discussion and Conclusion

Several versions of the spatial forecast accuracy test have been proposed in this work. Test statistics under parametric estimation of the covariogram and the semivariogram for both constant and spatially varying trends have been studied. Estimating the semivariogram yields better estimates of the variability in the loss differenced field and is recommended in practical applications. When a spatially varying trend is present, the importance of estimating this trend cannot be understated. Yet overall, the spatial forecast accuracy test is simple to compute, accounts for the presence of spatial correlation amongst the errors of a given loss function and for contemporaneous correlation, and allows flexible loss functions.

A comparison in the accuracy of competing models should not be the only diagnostic check used when comparing models. Forecasts produced by one model may contain information not included in another set, so a test of forecast encompassing, such as the one by Harvey et al. (1997), or a weighted average of forecasts can be very valuable tools as well.

This work highlights promising directions for future research. Some are evident, such as a forecast accuracy test for multivariate spatial forecasts or for space-time data. In fact, many of the most interesting examples involve forecasting multiple variables over space, and the wind speed forecasting 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 propose 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 forecast accuracy test.

The SHC method is at a disadvantage since it is more complex to implement, can only be applied to full dyadic grids (without making modifications), and does not perform well with non-Gaussian and spatially correlated data. However, one advantage of the SHC method that the spatial forecast accuracy test lacks is that it is able to estimate where the significant differences occur spatially. Looking at maps of the estimated trend, $\hat{f}(\mathbf{s})$, produced when detrending the data in the spatial forecast accuracy test does give some qualitative information about where the differences may exist, but reducing the domain of interest to detect regional differences may be a better quantitative solution. Benjamini and Heller (2007) argue that in analyzing fMRI data differences in signals at the individual locations are not as important as detecting differences in clusters of voxels. Their approach is more powerful than the SHC test since they have fewer hypotheses of interest to test. This suggests that one solution to detecting regional differences in forecast accuracy could be to apply the spatial forecast accuracy test in local regions of interest instead of across the entire set of forecasts.

In summary, the spatial forecast accuracy test is a very flexible and easily applied test. Used as a tool in model evaluation, it can help researchers determine if the difference they see in the average losses of two competing models is significant or not, which gives them a more complete, informed picture of their forecasts.

CHAPTER V

CONCLUSION

Forecasting wind speeds accurately is important for wind power integration into utility systems, and having the tools to statistically evaluate forecasts is important for decision makers. The TDD and BST models have been introduced, and they are flexible and can be fitted with a sparse number of locations. Neither is limited by the selection of regimes based on the prevailing wind patterns. The TDD model's predictions perform as well as the RSTD model predictions do, and the BST model outperforms both of the other models when the variability in the wind speed is low. In addition, the BST model produces a wind direction forecast as well, which is crucial in obtaining a wind power forecast. The TDD and BST models perform similarly when built to forecast at other locations in the dataset and when the hourly data is replaced with ten-minute data.

The PCE is a sketch of a loss function that incorporates the relationship between wind speed and wind power. It can be adapted to different types of turbines, and the penalty for underestimating wind speeds can be tuned to the particular season and utility system at hand. The optimal forecast from the predictive distribution for minimizing PCE is the quantile based on the underestimation penalty, and varying this penalty can have an effect on which model's forecasts are favored. The differences between the losses observed for each model can be tested for statistical significance using the time series test by Diebold and Mariano (1995). Using this test, we observe no significant difference in MSE, MAE, or PCE between the RSTD and TDD models and no significant difference in MAE or PCE between the RSTD and BST models.

For wind speed models that produce spatial predictions, they can be compared with the spatial forecast accuracy test, which tests the null hypothesis of equal forecast accuracy averaged across all spatial locations under a given loss function. This test does not require that the data at each location be normally distributed, and it incorporates spatial correlation in the difference field as well as contemporaneous correlation between sets of forecasts. The best estimator of the variance of $\bar{D}$ uses the weighted least squares estimate of the semivariogram. The main factor slowing the convergence of the test statistic to normality is the strength of the spatial correlation. As the spatial correlation increases, the larger the empirical size of the test becomes, and the loss function can have an effect on the spatial correlation, such as the quadratic loss that shrinks the spatial correlation.

Separating the difference between the trend and the spatial covariance contributs directly to how well the variance of $\bar{D}$ is estimated. If the trend is treated as covariance, then the null hypothesis will be rejected less often than it should be, and the reverse is true if the covariance is treated as trend. If the spatial trend of the difference field is constant, then detrending the data with a linear or a constant trend does not affect the outcome of the test. However, in the presence of a spatially varying trend, removing the trend with a non-parametric regression and a bandwidth adjusted for the spatial correlation does negatively affect the empirical size of the test, although it diminishes as the sample size increases.

It must be emphasized that this test is designed to determine if the average of the difference field is zero or not. If local regions of positive and negative values in the domain still sum to zero, as in the split pattern simulation with simple loss, then the null hypothesis of the test is still true. The SHC test, on the other hand, is designed to estimate regions in the domain where the signal is nonzero. Yet when the mean is constant, the SHC and spatial forecast accuracy tests are equivalent. In this case, the SHC test is missized, especially for the quadratic loss, since applying the loss function to the forecast errors transforms the data to a non-Gaussian distribution. Even if the SHC test were correctly sized, it can only be applied to full sets of gridded data of dyadic size, while the spatial forecast accuracy test can be used under more general conditions.

The spatial forecast accuracy test is demonstrated by comparing spatial forecasts of daily average wind speed at locations in Oklahoma. The difference field is formed for each set of forecasts based on the quadratic loss and the power curve loss. A nonparametric trend is removed on the data after adjusting the bandwidth selected by cross-validation for spatial correlation. The second set of spatial forecasts are shown to be significantly different on average from the time series forecasts in both quadratic and power curve losses.

Tests of forecast accuracy provide a rich area for future research. Improvements in the spatial version for small samples and adaptations for space-time forecasts can both be studied. These tests are just one type of tool that can be used to determine with statistical confidence whether two sets of forecasts are on average significantly different or not. This information can be used to select the best models, which ultimately guides decision-making and resource allocation.

REFERENCES

Alexiadis, M. C., Dokopoulos, P. S., and Sahsamanoglou, H. S. (1999). Wind speed and power forecasting based on spatial correlation models. *IEEE Transactions on Energy Conversion* **14**, 836–842.

Alonso, A. M., Peña, D., and Romo, J. (2006). Introducing model uncertainty by moving blocks bootstrap. *Statistical Papers* **47**, 167–179.

Atger, F. (2003). Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts: consequences for calibration. *Monthly Weather Review* **131**, 1509–1523.

Azzalini, A. (2005). The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics* **32**, 159–188.

Azzalini, A. (2006). `sn`: The skew-normal and skew-t distributions. *CRAN R Contributed Packages* **version 0.4-2**, Accessed 05/27/09 http://cran.r–project.org/web/packages/sn/index.html.

Azzalini, A. and Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew-t distribution. *Journal of the Royal Statistical Society, Series B* **65**, 367–389.

Azzalini, A. and Genton, M. G. (2008). Robust likelihood methods based on the skew-t and related distributions. *International Statistical Review* **76**, 106–129.

Benjamini, Y. and Heller, R. (2007). False discovery rates for spatial signals. *Journal of the American Statistical Association* **102**, 1272–1281.

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

Brown, B. G., Katz, R. W., and Murphy, A. H. (1984). Time series models to simulate and forecast wind speed and wind power. *Journal of Climate and Applied Meteorology* **23**, 1184–1195.

Cressie, N. (1985). Fitting variogram models by weighted least squares. *Journal of the International Association for Mathematical Geology* **17**, 563–586.

Cressie, N. (1993). *Statistics for Spatial Data*. New York: John Wiley & Sons.

de Luna, X. and Genton, M. G. (2005). Predictive spatio-temporal models for spatially sparse environmental data. *Statistica Sinica* **15**, 547–568.

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.

Dell'Aquila, R. and Ronchetti, E. (2004). Robust tests of predictive accuracy. *Metron* **62**, 161–184.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economics Statistics* **13**, 253–263.

Fisher, N. I. (1993). *Statistical Analysis of Circular Data*. Cambridge, Great Britain: Cambridge University Press.

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.

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.

Genton, M. G. and Hering, A. S. (2007). Blowing in the wind. *Significance* **4**, 11–14.

Giacomini, R. and White, H. (2006). Tests of conditional predictive ability. *Econometrica* **74**, 1545–1578.

Giebel, G., Brownsword, R., and Kariniotakis, G. (2003). The state-of-the-art in short-term prediction of wind power; a literature overview. Tech. rep., ANEMOS.

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association* **97**, 590–600.

Gneiting, T. (2008). Quantiles as optimal point predictors. Tech. rep. 538, Seattle, WA, University of Washington.

Gneiting, T., Larson, K., Westrick, K., Genton, M. G., and Aldrich, E. (2006). Calibrated probabilistic forecasting at the stateline wind energy center: the regime-switching space-time method. *Journal of the American Statistical Association* **101**, 968–979.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**, 359–378.

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.

Grimit, E. P., Gneiting, T., Berrocal, V. J., and Johnson, N. A. (2006). The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Quarterly Journal of the Royal Meteorological Society* **132**, 2925–2942.

Hart, J. D. (1996). Some automated methods of smoothing time-dependent data. *Journal of Nonparametric Statistics* **6**, 115–142.

Harvey, D., Leybourne, S., and Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting* **13**, 281–291.

Harvey, D., Leybourne, S., and Newbold, P. (1998). Tests for forecast encompassing. *Journal of Business and Economic Statistics* **6**, 254–259.

Hering, A. S. and Genton, M. G. (2009). Powering up with space-time wind forecasting. *Journal of the American Statistical Association* ***Accepted***.

Kestens, E. and Teugels, J. L. (2002). Challenges in modelling stochasticity in wind. *Environmetrics* **13**, 821–830.

Klink, K. (1999). Climatological mean and interannual variance of united states surface wind speed, direction, and velocity. *International Journal of Climatology* **19**, 471–488.

Kretzschmar, R., Eckert, P., and Cattani, D. (2004). Neural network classifiers for local wind prediction. *Journal of Applied Meteorology* **43**, 727–738.

Lange, M. (2005). On the uncertainty of wind power predictions—analysis of the forecast accuracy and statistical distribution of errors. *Journal of Solar Energy Engineering* **127**, 177–184.

Lange, M. and Focken, U. (2005). *Physical Approach to Short-Term Wind Power Prediction*. Berlin: Springer-Verlag.

Larson, K. and Westrick, K. (2006). Short-term wind forecasting using off-site observations. *Wind Energy* **9**, 55–62.

Li, B., Genton, M. G., and Sherman, M. (2007). A nonparametric assessment of properties of space-time covariance functions. *Journal of the American Statistical Association* **102**, 736–744.

Longhi, S. and Nijkamp, P. (2007). Forecasting regional labor market developments under spatial autocorrelation. *International Regional Science Review* **30**, 100–119.

Lund, U. and Agostinelli, C. (2006). `circular`: Circular statistics. *CRAN R Contributed Packages* **version 0.3-6**, Accessed 05/27/09 http://cran.r–project.org/web/packages/circular/index.html.

Madsen, H., Pinson, P., Kariniotakis, G., Nielsen, H. A., and Nielson, T. (2005). Standardizing the performance evaluation of short-term wind power prediction models. *Wind Engineering* **29**, 475–489.

Mardia, K. V. and Jupp, P. E. (2000). *Directional Statistics*. London: John Wiley and Sons.

Mardia, K. V. and Marshall, R. J. (1984). Maximum likelihood estimation of models for residual spatial covariance in spatial regression. *Biometrika* **71**, 135–146.

Martin, M., Cremades, L. V., and Santabárbara, J. M. (1999). Analysis and modelling of time series of surface wind speed and direction. *International Journal of Climatology* **19**, 197–209.

Matsuo, T., Nychka, D., and Paul, D. (2006). Multiresolution (wavelet) based nonstationary covariance modeling for incomplete data: smoothed monte carlo approach. Tech. rep., Boulder, CO National Center for Atmospheric Research.

McCracken, M. W. (2004). Parameter estimation and tests of equal forecast accuracy between non-nested models. *International Journal of Forecasting* **20**, 503–514.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications* **9**, 141–142.

Nychka, D., Wikle, C., and Royle, J. A. (2002). Multiresolution models for nonstationary spatial covariance functions. *Statistical Modelling* **4**, 315–331.

Opsomer, J., Wang, Y., and Yang, Y. (2001). Nonparametric regression with correlated errors. *Statistical Science* **16**, 134–153.

Park, B. U., Kim, T. Y., Park, J., and Hwang, S. Y. (2008). Practically applicable central limit theorem for spatial statistics. *Mathematical Geosciences* ***Online***, Accessed 05/28/09, http://www.springerlink.com.lib–ezproxy.tamu.edu:2048/content/p8671g543379p335/fulltext.pdf.

Patton, A. J. and Timmermann, A. (2007). Testing forecast optimality under unknown loss. *Journal of the American Statistical Association* **102**, 1172–1184.

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.

Pinson, P., Chevallier, C., and Kariniotakis, G. N. (2007). Trading wind generation from short-term probabilistic forecasts of wind power. *IEEE Transactions on Power Systems* **22**, 1148–1156.

Potter, C. W., Gil, H. A., and McCaa, J. (2007). Wind power data for grid integration studies. Tech. rep. Tampa Bay, Florida 07GM0808, Proceedings of the IEEE/PES General Meeting.

Schabenberger, O. and Gotway, C. A. (2005). *Statistical methods for spatial data analysis*. Boca Raton, FL: Chapman & Hall.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

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.

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.

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.

Smith, J. C., Parsons, B., Acker, T., Milligan, M., Zavadil, R., Schuerger, M., and DeMeo, E. (2007). Best practices in grid integration of variable wind power: Summary of recent US case study results and mitigation measures. Tech. rep., Milan, Italy European Wind Energy Conference.

Snell, S., 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.

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.

Watson, G. S. (1964). Smooth regression analysis. *Shankya Series A* **26**, 359–372.

Weisberg, S. (2005). *Applied Linear Regression*. Hoboken, NJ: Wiley-Interscience.

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica* **64**, 1067–1084.

Wikle, C. K., Milliff, R. F., Nychka, D., and Berliner, L. M. (2001). Spatiotemporal hierarchical bayesian modeling: Tropical ocean surface winds. *Journal of the American Statistical Association* **96**, 382–397.

Willis, H. L. (2002). *Spatial Electric Load Forecasting*. New York, NY: Marcel-Dekker, Inc.

APPENDIX A

PROOFS OF PROPOSITIONS PRESENTED IN CHAPTER IV

Proof of Proposition 1:

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

$$\mathbf{S} = \begin{bmatrix} (D(\mathbf{s}_1) - \bar{D})(D(\mathbf{s}_1) - \bar{D}) & (D(\mathbf{s}_1) - \bar{D})(D(\mathbf{s}_2) - \bar{D}) & \dots & (D(\mathbf{s}_1) - \bar{D})(D(\mathbf{s}_L) - \bar{D}) \\ (D(\mathbf{s}_2) - \bar{D})(D(\mathbf{s}_1) - \bar{D}) & (D(\mathbf{s}_2) - \bar{D})(D(\mathbf{s}_2) - \bar{D}) & \dots & (D(\mathbf{s}_2) - \bar{D})(D(\mathbf{s}_L) - \bar{D}) \\ \vdots & \vdots & \ddots & \vdots \\ (D(\mathbf{s}_L) - \bar{D})(D(\mathbf{s}_1) - \bar{D}) & (D(\mathbf{s}_L) - \bar{D})(D(\mathbf{s}_2) - \bar{D}) & \dots & (D(\mathbf{s}_L) - \bar{D})(D(\mathbf{s}_L) - \bar{D}) \end{bmatrix}. \tag{A.1}$$

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. For example, the sum of the ith row is

$$\begin{aligned} \sum_{j=1}^{L}(D(\mathbf{s}_i) - \bar{D})(D(\mathbf{s}_j) - \bar{D}) &= (D(\mathbf{s}_i) - \bar{D})\sum_{j=1}^{L}(D(\mathbf{s}_j) - \bar{D}) \\ &= (D(\mathbf{s}_i) - \bar{D})\left(\sum_{j=1}^{L} D(\mathbf{s}_j) - L\bar{D}\right) \\ &= (D(\mathbf{s}_i) - \bar{D})\left(L\bar{D} - L\bar{D}\right) = 0. \end{aligned}$$

Let $\mathbf{h} = \{h_0, h_1, h_2, \dots, h_m\}$ be the ordered set of unique distances between all pairs of observations. For example, on a lattice with locations spaced one unit apart, $h_0 = 0$, $h_1 = 1$, and $h_2 = 1.41$, and h_m is the maximum distance between any two pairs of observations. Then, $N(h_i)$ is the set of all pairs of points $(\mathbf{s}_i, \mathbf{s}_i')$ distance h_i apart, and $|N(h_i)|$ is the number of pairs of points distance h_i apart. The sum of all of the elements in the matrix in Equation (A.1) is

$$\begin{aligned}
\sum_{\{i,j\}} s_{ij} &= \sum_{i=1}^{L}(D(\mathbf{s}_i)-\bar{D})^2 + \sum_{N(h_1)}(D(\mathbf{s}_1)-\bar{D})(D(\mathbf{s}_1')-\bar{D}) + \ldots \\
&\quad \ldots + \sum_{N(h_m)}(D(\mathbf{s}_m)-\bar{D})(D(\mathbf{s}_m')-\bar{D}) \\
&= L\hat{C}(0) + |N(h_1)|\hat{C}(h_1) + \ldots + |N(h_m)|\hat{C}(h_m).
\end{aligned}$$

Now, we show that Equation (4.5) with the estimated covariances substituted for $C(h_{ij})$ is equal to $\sum_{\{i,j\}} s_{ij}$.

$$\begin{aligned}
V\hat{a}r\left[\bar{D}\right] &= \frac{1}{L^2}\sum_{i=1}^{L}\sum_{j=1}^{L}\hat{C}(h_{ij}) \\
&= \frac{1}{L^2}\left[L\hat{C}(0) + |N(h_1)|\hat{C}(h_1) + |N(h_2)|\hat{C}(h_2) + \ldots + |N(h_m)|\hat{C}(h_m)\right] \\
&\quad \text{(after collecting like terms according to distance)} \\
&= \frac{1}{L^2}\left[\sum_{\{i,j\}} s_{ij}\right] = 0.
\end{aligned}$$

□

Proof of Proposition 2:

We must find Cov$(X^2, Y^2) =$ E$(X^2Y^2) -$ E(X^2)E(Y^2). With mean zero, E$(X^2) = \sigma_x^2$ and E$(Y^2) = \sigma_y^2$. The moment generating function (mgf) can be used to find E(X^2Y^2). The mgf for a multivariate normal distribution is

$$\begin{aligned}
M_{\mathbf{Z}}(\mathbf{t}) &= \exp\left(\boldsymbol{\mu}'\mathbf{t} + \frac{1}{2}\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}\right) = \exp\left(\frac{1}{2}\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}\right), \text{ for } \boldsymbol{\mu} = \mathbf{0} \\
&= \exp\left((1/2)(s_1t_1^2 + 2s_2t_1t_2 + s_3t_2^2)\right).
\end{aligned}$$

Then, E$(X^2Y^2) = \frac{\partial^4}{\partial t_1^2 \partial t_2^2} M_{\mathbf{Z}}(\mathbf{t})|_{\mathbf{t}=\mathbf{0}}$.

The partial derivative is:

$$\begin{aligned}\frac{\partial^4}{\partial t_1^2 \partial t_2^2} M_{\mathbf{Z}}(\mathbf{t}) = & \quad \exp\left((1/2)(s_1 t_1^2 + 2s_2 t_1 t_2 + s_3 t_2^2)\right)(s_1 t_1 + s_2 t_2)^2 (s_2 t_1 + s_3 t_2)^2 \\ + & \quad 2s_2 \exp\left((1/2)(s_1 t_1^2 + 2s_2 t_1 t_2 + s_3 t_2^2)\right)(s_1 t_1 + s_2 t_2)(s_2 t_1 + s_3 t_2) \\ + & \quad s_3 \exp\left((1/2)(s_1 t_1^2 + 2s_2 t_1 t_2 + s_3 t_2^2)\right)(s_1 t_1 + s_2 t_2)^2 \\ + & \quad 2s_2 \exp\left((1/2)(s_1 t_1^2 + 2s_2 t_1 t_2 + s_3 t_2^2)\right)(s_2 t_1 + s_3 t_2)(s_1 t_1 + s_2 t_2) \\ + & \quad 2s_2^2 \exp\left((1/2)(s_1 t_1^2 + 2s_2 t_1 t_2 + s_3 t_2^2)\right) \\ + & \quad s_1 \exp\left((1/2)(s_1 t_1^2 + 2s_2 t_1 t_2 + s_3 t_2^2)\right)(s_2 t + s_3 t_2)(s_2 t_1 + s_3 t_2) \\ + & \quad s_1 s_3 \exp\left((1/2)(s_1 t_1^2 + 2s_2 t_1 t_2 + s_3 t_2^2)\right).\end{aligned}$$

Evaluating at $\mathbf{t} = \mathbf{0}$ yields

$$\begin{aligned}\text{E}(X^2 Y^2) &= 2s_2^2 + s_1 s_3 = 2(\rho\sigma_x\sigma_y)^2 + \sigma_x^2\sigma_y^2 \\ &= 2\rho^2\sigma_x^2\sigma_y^2 + \sigma_x^2\sigma_y^2, \text{ so} \\ \text{Cov}(X^2, Y^2) &= 2\rho^2\sigma_x^2\sigma_y^2 + \sigma_x^2\sigma_y^2 - \sigma_x^2\sigma_y^2 \\ &= 2\rho^2\sigma_x^2\sigma_y^2.\end{aligned}$$

The correlation is

$$\begin{aligned}\text{Corr}(X^2, Y^2) &= \frac{\text{Cov}(X^2, Y^2)}{\sqrt{\text{Var}(X^2)}\sqrt{\text{Var}(Y^2)}} \\ &= \frac{2\rho^2\sigma_x^2\sigma_y^2}{\sqrt{2\sigma_x^4}\sqrt{2\sigma_y^4}} \\ &= \rho^2,\end{aligned}$$

which is obtained by computing Var(X^2) and Var(Y^2) similarly. □

Proof of Proposition 3:

For $\mathbf{Z} = (X, Y)^T \sim (\mathbf{0}, \Sigma)$, the effect of $g(\cdot)$ is similar to that of $g(x) = x^2$, where g is a function such that $g(0) = g'(0) = 0$, and $g(\cdot)$ is twice differentiable. Performing a second order Taylor expansion of $g(x)$ about 0 yields

$$\begin{aligned} g(x) &= g(\mu) + g'(\mu)(x - \mu) + g''(\mu)(x - \mu)^2 + R \\ &= g''(0)x^2 + R. \end{aligned}$$

Then,

$$\begin{aligned} &\text{E}(g(X)) \approx g''(0)\text{E}(X^2), \\ &\text{E}(g(X)g(Y)) \approx (g''(0))^2\text{E}(X^2Y^2), \text{ and} \\ &\text{Var}(g(X)) \approx (g''(0))^2\text{Var}(X^2). \end{aligned}$$

The correlation between $g(X)$ and $g(Y)$ is then

$$\begin{aligned} \text{Corr}(g(X), g(Y)) &= \frac{\text{Cov}(g(X), g(Y))}{\sqrt{\text{Var}(g(X))}\sqrt{\text{Var}(g(Y))}} \\ &\approx \frac{(g''(0))^2 \cdot (\text{E}(X^2Y^2) - \text{E}(X^2)\text{E}(Y^2))}{\sqrt{g''(0)^2\text{Var}(X^2)}\sqrt{g''(0)^2\text{Var}(Y^2)}} \\ &= \text{Corr}(X^2, Y^2), \text{ which, if } \mathbf{Z} = (X, Y)' \sim \text{N}_2(\mathbf{0}, \Sigma) \\ &= \rho^2 \end{aligned}$$

However, if the functions $g(\cdot)$ and $g'(\cdot)$ are not zero at 0, then for a second order expansion of $g(X)$ and using tactics similar to the ones above, the correlation takes the following form.

$$\text{Corr}(g(X), g(Y)) \approx \frac{g'(0)^2\rho\sigma_x\sigma_y + 2g''(0)^2\rho^2\sigma_x^2\sigma_y^2}{\sqrt{g'(0)^2\sigma_x^2 + 2g''(0)^2\sigma_x^4}\sqrt{g'(0)^2\sigma_y^2 + 2g''(0)^2\sigma_y^4}}$$

This reduces to ρ^2 when $g'(0) = 0$ and $g''(0) \neq 0$. It reduces to ρ when $g''(0) = 0$ and $g'(0) \neq 0$.

□

VITA

Amanda S. Hering was born in Lafayette, Lousiana. In August of 1995, she entered the undergraduate program in mathematics at Baylor University in Waco, Texas and graduated summa cum laude with a Bachelor of Science degree in May 1999. She continued her studies at Montana State University in Bozeman, Montana, earning a Master of Science degree in statistics in August 2002. She began pursuing a Doctor of Philosophy degree in statistics at Texas A&M University in College Station, Texas in August 2004, which was completed under the advisement of Professor Marc G. Genton in August 2009. Her mailing address is:

Department of Mathematical and Computer Sciences
Colorado School of Mines
Golden, Colorado 80401-1887