

Comparison of the Performance of Four Measure-Correlate-Predict Algorithms

Anthony L. Rogers, Ph. D., Senior Research Fellow*
(Corresponding author)

John W. Rogers, M.S., Senior Statistician**

James F. Manwell, Ph. D., Assistant Research Professor*

*Renewable Energy Research Laboratory

Dept. of Mechanical and Industrial Engineering

University of Massachusetts, Amherst, MA 01003

Telephone: 413-545-4866, fax. 413-545-1027, Email: arogers@ecs.umass.edu

**Westat, 1650 Research Blvd., Rockville, MD 20850

Telephone: 301-294-2804, Email: rogersjl@westat.com

Abstract

Measure-correlate-predict (MCP) algorithms are used to predict the wind resource at target sites for wind power development. This paper describes some of the MCP approaches found in the literature and then compares the performance of four of them, using a common set of data from a variety of sites (complex terrain, coastal, offshore). The algorithms that are compared include a linear regression model, a model using distributions of ratios of the wind speeds at the two sites, a vector regression method, and a method based on the ratio of the standard deviations of the two data sets. The MCP algorithms are compared using a set of performance metrics that are consistent with the ultimate goals of the MCP process. The six different metrics characterize the estimation of 1) the correct mean wind speed, 2) the correct wind speed distribution and 3) the correct annual energy production at the target site, assuming a sample wind turbine power curve, and 4) the correct wind direction distribution. The results indicate that the method using the ratio of the standard deviations of the two data sets and the model that uses the distribution of ratios of the wind speeds at the two sites work the best. The linear regression model and the vector regression model give biased estimates of a number of the metrics, due to the characteristics of linear regression.

Keywords

Wind Resource, Wind Resource Estimation, Wind Speed Distribution, Measure-Correlate-Predict, MCP.

Introduction

Measure-correlate-predict (MCP) algorithms are used to predict the wind resource at target sites for wind power development. MCP methods model the relationship between wind data (speed and direction) measured at the target site, usually over a period of up to a year, and concurrent data at a nearby reference site. The model is then used with long-term data from the reference site to predict the long-term wind speed and direction distributions at the target site. Defining a relationship between the sites is complicated by stochastic variations in wind speed and direction over time and distance, the effects of terrain on the flow, time of flight delays, large-scale and small-scale weather patterns, local obstructions and atmospheric stability. As a result, appropriate models and periods of concurrent data need to be chosen to ensure confidence in the results. The goal of the MCP method is usually a characterization of wind speed distributions as a function of wind direction at the target site in order to be able to determine the annual energy capture of a wind farm located at the target site. Although prediction of wind speed usually includes consideration of wind direction at the reference site, wind direction is usually modeled independently of the wind speed. This paper focuses on the prediction of wind speed at the target site and discusses wind direction primarily related to the prediction of wind speed.

Over the last 15 years well over a half a dozen variations on the MCP technique have been proposed, in part, to address some of the specific concerns mentioned above. These MCP algorithms differ in terms of overall approach, model definition, use of direction sectors, length of data used for the documented validation effort, data used for validation effort, criteria used for evaluating the required length of concurrent data and criteria used for evaluating the effectiveness of the approach. This paper starts with a brief description of some of approaches found in the literature.

This paper then compares the performance of three of them and one proposed by the authors, using a common set of data from a variety of sites (complex terrain, coastal, offshore). The algorithms that are compared include a linear regression model, a model using distributions of ratios of the wind speeds at the two sites, a vector regression method, and a method based on the ratio of the standard deviations of the two data sets.

The MCP algorithms are compared using a set of performance measures that are consistent with the ultimate goals of the MCP process. The six different metrics characterize the estimation of 1) the correct mean wind speed, 2) the correct wind speed distribution and 3) the correct annual energy production at the target site, assuming a sample wind turbine power curve, and 4) the correct wind direction distribution. The metrics are determined from multiple estimates based on different periods of concurrent data. The mean and standard deviation of those estimates are used to characterize the results.

MCP Algorithms

In general, MCP algorithms use models to predict wind speed at the target site from wind speed and, possibly, other conditions at the reference site. The data may also be grouped or binned by other factors, such as the wind direction at the reference site. In this case, separate parameters are fit for each bin. Different MCP algorithms use different methods to fit the parameters. The fitted model is then applied to long-term data at the reference site to predict long-term data at the target site. If some bins have long-term reference-site data but no data from the concurrent period, assumptions must be used to make predictions. Some of the approaches proposed so far are summarized in Table 1 and described below.

Derrick [1, 2] used linear regression to characterize the relationship between the reference and target site wind speeds:

$$\hat{y} = mx + b$$

where \hat{y} is the predicted wind speed at the target site, x is the observed wind speed at the reference site, m and b are the slope and offset determined from linear regression. Separate parameters were calculated for data in twelve 30-degree direction bins defined by wind direction at the reference site. He also pointed out that linear regression could be used for models of the form:

$$\hat{y} = ax^b$$

In this case, the linear regression would be performed after taking the logarithm of both sides of the equation. He discussed the use of the variances and covariances calculated from the linear regression to determine confidence intervals for the predicted mean wind speed assuming that the assumptions of the linear regression approach apply. Derrick also concluded, using data from the UK, that the majority of the scatter in the data was caused by the passage of weather systems with significant isobaric curvature. Finally, based on the available data, Derrick concluded that at least 8 months of data was needed to minimize uncertainties in the results. In [2] Derrick considered the use of polynomials to characterize differences in wind direction at the two sites. Due to scatter

in the data, the predicted wind direction distribution only roughly approximated the known distribution. Finally, to improve energy capture estimates, Derrick suggested not using data at low wind speeds (<4 m/s) in the determination of the correlation, as wind vane behavior is erratic and wind turbines do not operate at low wind speeds. Not using low wind speed data resulted in incorrect estimates of overall mean wind speed.

Nielsen et al. [4] proposed a two-dimensional linear regression fit of the form:

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

in which (x_1, x_2) and (\hat{y}_1, \hat{y}_2) are the observed velocities at the reference site and the predicted velocities at the target site in orthogonal coordinates. The a_{ij} and b_{ij} coefficients are determined using linear regression, separately for twelve wind direction sectors. In this manner, direction differences at the two sites can be accommodated. The final predicted velocities at the target site are:

$$\hat{y} = \sqrt{\hat{y}_1^2 + \hat{y}_2^2}$$

The authors consider the effects of averaging period on the results and compare the proposed approach to simple linear regression, as used by Derrick. They demonstrate that the cross-correlation coefficients between the known and predicted wind speeds improve as the averaging period of the data is increased and suggest, based on the scatter of the cross-correlation phases, that predicting wind direction may be difficult. The results indicate that instantaneous wind directions are not well predicted by the two-dimensional linear regression method and that instantaneous speeds are predicted better by simple linear regression.

Riedel et al. [5] propose a model of the form

$$\hat{y} = ax + \frac{M \sin(e)}{v_{\max}^2} x^2$$

in which the a and e coefficients are chosen to minimize the difference between the predicted and known speed distributions as measured by a Chi-square test. The form of the coefficient of x^2 is chosen to limit the magnitude of the coefficient of the quadratic term, with respect to that of the linear term. M is chosen to be 10 m/s at a v_{\max} of 30 m/s. The model is fit within dynamically chosen wind direction sectors. The wind direction is modeled using a 10-coefficient Chebyshev polynomial with the coefficients chosen to minimize the Chi-square test for the predicted direction distribution. The authors note problems with the numerical methods. The limited results suggest that the wind speed algorithm may not perform any better than other approaches, but that the wind direction model works well when there are no very distinct wind directions. Short periods of concurrent data work less well, especially comparing offshore and inland sites. To improve the results, the authors suggest that differences in stability at the two sites need to be considered. These stability differences are a function of time of year, thus, breaking the data into separate months might improve the results.

Landberg and Mortenson [6] compared MCP using simple linear regression with predictions from an early version of WASP, a computer code for predicting wind speeds using a Jackson-Hunt type flow model with a look-up table [7]. Their results show that concurrent data lengths should be at least 9 months long, with little improvement using longer data lengths. They also show that excluding data below 3 m/s from the linear regression may result in biased estimates of the mean wind speeds.

Woods and Watson [8] focus on wind direction issues. They bin the data by direction sectors at both the reference and target site, developing matrices of frequency of occurrence as a function of

the sectors at each site. They then use linear regression to relate wind speeds in each bin and the matrix of frequencies of occurrence to determine the final wind speed and direction distributions. Results show improved wind direction predictions between sites with significantly different wind directions. This method predicts mean wind speeds for each direction sector at the target site, rather than the distribution of wind speeds.

Vermeulen et al. [9] use a matrix approach similar to that of Woods and Watson, relating wind speeds using

$$\hat{y} = ax$$

Joensen et al. [3] consider correlations between data from significantly different heights and suggest two possible models that incorporate differences in stability:

$$\hat{y} = ax + b + cx^2 \quad \text{or} \quad \hat{y} = ax + b + cx\Delta T$$

where ΔT is the temperature gradient at two heights at the reference site. The three coefficients are determined as a function of direction by using local regression (estimating a polynomial approximation of the coefficient functions in a grid spanning the wind direction). For the estimation, a neighborhood about each direction that includes 40% of the data is used. In this neighborhood, weights are assigned to the data as a function of distance from the wind direction of interest. In addition, the space in which the fit is preformed is rotated in an attempt to correct for possible biases by using:

$$\begin{aligned} y' &= \cos(\alpha)y - \sin(\alpha)x \\ x' &= \sin(\alpha)y + \cos(\alpha)x \end{aligned}$$

where x' and y' are the variables in rotated space. The rotation angle, α , is determined by maximizing a function of the variance of the rotated target-site values. The authors test the method on a set of data and demonstrate that it produces wind speed distributions close to the measured ones.

Mortimer [10] has proposed a binning method in which concurrent data at the two sites are binned by wind direction sector and wind speed at the reference site. Within each bin the ratios of the concurrent target and reference wind speeds are calculated. Two matrices are produced: a matrix of the average of the ratios and a matrix of standard deviation of the ratios in each bin. The prediction equation takes the form:

$$\hat{y} = (r + e)x$$

where \hat{y} is the predicted wind speed at the target site, x is the observed wind speed at the reference site, r is the average ratio for the appropriate direction sector and speed bin and e is a random variable with a triangular distribution with the appropriate direction sector and speed bin standard deviation. Mortimer suggests that the method may predict extreme wind speeds better than linear regression.

MCP Algorithms Chosen for Evaluation and Their Implementation

The methods of Mortimer, Derrick and Nielsen et al. and an alternative proposed by the authors have been chosen for evaluation. In all cases the data has been divided into eight direction sectors. All low wind speed data are included in the analysis in order to evaluate, among other things, the success of the methods for determining the mean wind speed. The specifics of the implementation of these models are described below.

As mentioned above, Mortimer [10] uses a prediction equation which takes the form:

$$\hat{y} = (r + e)x$$

where r is the average ratio for the appropriate direction sector and speed bin and e is a random variable with a triangular distribution with the appropriate direction sector and speed bin standard deviation. As implemented for this paper, when determining the ratios in the concurrent data, if the reference wind speed is less than 1 m/s, the ratio for that data point is assumed to be 1. Also, if there is not enough data to determine the average ratio for any speed and direction bin, the ratio is assumed to be the ratio of the mean of all of the wind speeds for that direction sector and the standard deviation is assumed to be zero. When predicting wind speeds, if the method gives negative predicted wind speeds, these are set to zero. The target site wind direction is assumed to be the reference direction. In the following comparisons, this method will be referred to as the “Mortimer” method.

Derrick [1, 2] uses linear regression to characterize the relationship between the reference and target site for each direction bin. The data are binned by the reference site wind direction and the regression coefficients are determined. When using those coefficients, if negative wind speeds are predicted, then they are set to zero. The target site wind direction is assumed to be the reference direction. In the analysis that follows, for simplicity this method is referred to as the “Linear Regression” method. Derrick presents formulas for confidence intervals for the results. These formulas are not addressed in this paper.

Finally, as mentioned above, Nielsen et al. [4] have proposed a two-dimensional linear regression fit. The final predicted velocities at the target site are:

$$\hat{y} = \sqrt{\hat{y}_1^2 + \hat{y}_2^2}$$

The predicted wind speeds cannot be negative using this method. The method directly predicts target site wind direction. These predictions are compared to the assumption that target site wind direction is the same as the reference direction. This method is referred to as the “Vector” method in the results section.

Alternate MCP Algorithm

A number of the methods mentioned above use linear regression. Linear regression is used for many problems. It has the advantage that it can be easily implemented using available software and, if the linear regression assumptions apply, provides measures of precision, including confidence intervals. As applied to the MCP problem, given the wind speed at the reference site, x , linear regression provides a predicted wind speed, \hat{y} , at the target site. Using regression, the overall mean of the observed values, y , at the target site should be very close to the overall mean of the predicted values, \hat{y} . (They would be the same if the fit were calculated using the complete target site data set.) On the other hand, the variance of the predicted wind speed about the mean will be smaller than the variance of the observed wind speeds by a factor equal to r -square from the regression fit [11]. This can result in biased predictions of wind speed distributions.

Another model has been considered. The variance of the predicted target site wind speeds, $\sigma^2(\hat{y})$, for a linear model of the form $\hat{y} = mx + b$ is: $\sigma^2(\hat{y}) = \sigma^2(mx + b) = m^2\sigma^2(x)$. Setting m^2 equal to $\sigma^2(y)/\sigma^2(x)$ ensures that $\sigma^2(\hat{y}) = \sigma^2(y)$. Thus, the linear model for which the predicted values are expected to have the same overall mean and variance as the observed values is:

$$\hat{y} = (\mu_y - (\sigma_y/\sigma_x)\mu_x) + (\sigma_y/\sigma_x)x$$

where μ_x , μ_y , σ_x and σ_y are the mean and standard deviations of the two concurrent data sets. This alternate model has been tested in addition to the other three and is referred to as the “Variance Ratio” method for simplicity. Again, if negative wind speeds are predicted, then they are set to zero. The target site wind direction is assumed to be the reference direction.

Data Sets

The four MCP algorithms have been compared using a variety of data sets. The data sets include hourly averages from eight sites: three in offshore or coastal areas near Massachusetts, on the eastern coast of the US, three sites in rolling terrain in the upper Midwest plains of the US and three on ridges in complex terrain in Western Massachusetts. The site characteristics are shown in Table 2. Pairs of sites that were used are shown in Table 3. Table 3 shows the maximum value of the cross correlation function, using the data from the two sites, and the time lag of the maximum cross correlation. All of the paired data sets had very similar reference and target site wind directions except the Century Tower-Brodie Mountain set which had significant differences in wind direction between the two sites and the Brodie Mountain - Burnt Hill data which had a relatively constant 20-degree difference in wind direction.

Some of the data sets have missing data due to icing and sensor or logger failure. These data are flagged. In the analysis, any sets of data (two wind speeds and two directions with a common time stamp) that have a flagged value for any one of the individual values are excluded.

Metrics Used

Six metrics are used to evaluate the usefulness of each of the four MCP methods. The six different metrics, designated m_1 to m_6 , characterize the correct estimation of 1) mean wind speed, 2) wind speed distribution and 3) annual energy production, assuming a sample wind turbine power curve, and 4) wind direction distribution.

Mean long-term wind speed is often used to characterize a potential wind power site and is an important measure of wind power potential. To facilitate comparisons among data sets, a normalized mean wind speed (predicted long-term target site wind speed divided by the observed long-term target site wind speed), m_1 , is calculated, using the following formula, where y and \hat{y} are the measured and predicted hourly averaged wind speeds at the target site and N is the total number of paired data points used in the analysis:

$$m_1 = \frac{\frac{1}{N} \sum_{i=1}^N \hat{y}_i}{\frac{1}{N} \sum_{i=1}^N y_i}$$

Wind speed distributions are necessary to correctly determine the wind farm energy capture and fatigue loading. Wind speed distributions are characterized by three measures. The first and second are the estimated Weibull shape and scale parameters, k and c . To facilitate comparisons among data sets, normalized values of k and c (the ratio of $k_{\hat{y}}$ (or $c_{\hat{y}}$) for the predicted wind speeds to k_y (or c_y) for the observed target wind speeds), m_2 and m_3 , are calculated:

$$m_2 = \frac{k_{\hat{y}}}{k_y}$$

$$m_3 = \frac{c_{\hat{y}}}{c_y}$$

The final metric used to characterize wind speed distributions is a Chi-squared goodness of fit measure. The Chi-square goodness of fit statistic provides a measure of how well two distributions agree with each other. For the wind speed data, the chi-squared statistic is calculated by first dividing the data into bins, $i = 1$ to M . For the observed target wind speeds, y , and predicted target site wind speeds, \hat{y} , the number of observations in each bin ($n_{y,i}$ and $n_{\hat{y},i}$ respectively) are counted and used to calculate the Chi-square statistic. If the observations were

independent, the chi-square statistic would have a probability interpretation. However, because the observations are serially correlated and therefore not independent, the chi-square statistic is used only as a relative measure of goodness of fit. The magnitude of the chi-square statistic depends on the number of bins, or degrees of freedom. The number of bins that were used was the same for all data sets. As a result, the degrees of freedom is not a factor when using the chi-square statistic as a relative measure of goodness of fit. If the observed and predicted values are independent and have the same distribution then the chi-square statistic does not depend on the number of observations. However, in this application the magnitude of the chi-square statistic does depend on the number of observations. To facilitate comparisons across data sets, the chi-square statistic was scaled by dividing by the number of observations used to calculate the statistic.¹ The normalized chi-square statistic, m_4 , is calculated as:

$$m_4 = \sum_{i=1}^M \frac{(n_{y,i} - n_{\hat{y},i})^2}{n_{y,i}N}$$

A value of zero would indicate perfect correspondence between the observed and predicted distributions. For comparing wind speed, the following wind speed bins were used: less than 3 m/s, 3 to less than 4, 4 to less than 5, ..., 11 to less than 12, and 12 m/s or greater. The bins were defined so that each bin had at least 50 observations.

Wind direction is important for energy capture calculations in wind farms where wake losses and complex terrain affect total energy capture. This paper does not directly address prediction of wind direction. However, wind directions are predicted by the vector method as part of predicting the wind speed. The wind directions from the vector method were compared to the simple assumption (often used in practice) that the wind directions at the target site are the same as at the reference site. For these models, the wind direction distributions are also characterized using a Chi-squared goodness of fit measure, using the normalized frequency in each of the wind direction sectors. For comparing wind direction distributions, 8 bins of 45 degrees each were used. For the observed target wind speeds, y , and predicted target site wind speeds, \hat{y} , the number of observations in each direction bin ($d_{y,i}$ and $d_{\hat{y},i}$ respectively) are counted; and the normalized direction distribution chi-square statistic, m_5 , is calculated as:

$$m_5 = \sum_{i=1}^8 \frac{(d_{y,i} - d_{\hat{y},i})^2}{d_{y,i}N}$$

When assuming that the predicted target wind direction is the same as that of the reference site, the Chi-squared goodness of fit measures the degree of the correspondence between the target and reference wind directions.

Finally, wind speed distributions and directional effects do not totally determine energy capture. Many wind turbines have constant power output over a range of wind speeds. Thus, energy production calculations, assuming an appropriate power curve, help to indicate a method's success in predicting expected energy capture. The actual metric used is capacity factor. The wind turbine power curve is assumed to have a cut-in wind speed of 4 m/s, a cubic power curve up to the rated wind speed of 11 m/s and rated power of 2 MW and then a flat power curve to 25 m/s. Furthermore, it is assumed that the hub height is adjusted so that the mean hub height wind speed is 8 m/s. Thus, all of the data are scaled by a multiplicative factor such that the resulting mean is 8 m/s. To facilitate comparison between paired data sets, a normalized capacity factor, m_6 , is calculated. To determine the value of this metric, the ratio of the capacity factor for the predicted wind, $c_{p,\hat{y}}$, is divided by the capacity factor for the observed target site wind speed, $c_{p,y}$:

¹ This is equivalent to calculating the chi-square statistic using percentages instead of counts.

$$m_6 = \frac{c_{p,\hat{y}}}{c_{p,y}}$$

Calculations

Each of the four MCP algorithms is examined using the eight paired sets of long-term data. The analysis consists of 1) an investigation to consider the effect of concurrent data length on the variability of the results with each algorithm and 2) comparisons of the performance of each algorithm, using multiple independent sets of concurrent data of a fixed period.

To analyze the effect of concurrent data length, multiple experiments were performed with a variety of presumed lengths of concurrent data to analyze the effect of concurrent data length on the results. The mean and standard deviations of the metrics as a function of concurrent data length are used to evaluate appropriate lengths of concurrent data. For these calculations, concurrent data from non-overlapping sub-sets of the data are used to calculate model parameters. The data are divided into 8 direction sectors of 45 degrees width and model parameters are calculated for each sector. For each sub-set, those parameters are used to predict the long-term target site wind data from the long-term reference site wind data and the metrics discussed above are calculated. Across all subsets, the average and standard deviation of the metrics are calculated. The standard deviation of the metrics from non-overlapping sub-sets provides a measure of precision that is relatively unaffected by the serial correlation of the data.²

Values of the metrics are determined for a variety of lengths of sub-sets of the data. If there are a total of N data points for two paired sites, then for each choice of concurrent data length, n_j , there are p_j possible subsets of concurrent data, where:

$$p_j = \text{Int}\left(\frac{N}{n_j}\right)$$

Seventeen values of concurrent data length are used, ranging from $n_j = 2000$ to 10000 hours, in steps of 500. The mean, $\bar{m}_{j,k,l}$, and standard deviation, $\sigma_{m_{j,k,l}}$, of each metric was determined for each data set and length of concurrent data:

$$\bar{m}_{j,k,l} = \frac{1}{p_j} \sum_i^{p_j} m_{i,j,k,l}$$

$$\sigma_{m_{j,k,l}} = \sqrt{\frac{\sum_i^{p_j} m_{i,j,k,l}^2 - p_j (\bar{m}_{j,k,l})^2}{p_j - 1}}$$

where i is the index for each separate section of concurrent data, j is the index for each choice of length of concurrent data, k is the index for each data set and l is the index for each of the six metrics.

To compare the performance of each of the methods, the statistics for each of the normalized metrics across all of the data sets were determined for a concurrent data length of 7000 hours.

² The results from adjacent subsets of the data have a small correlation from the observations that are closest in time, and are thus not strictly independent. The assumption of independence also assumes that seasonal patterns affect each subset equally. This last assumption is less likely for subsets of less than one year.

Thus, the mean, $M_{j,l}$, and standard deviation, $\sigma_{\bar{m}_{j,l}}$, of each metric, $\bar{m}_{j,k,l}$, across the eight data sets is:

$$M_{j,l} = \frac{1}{8} \sum_k \bar{m}_{j,k,l}$$

$$\sigma_{\bar{m}_{j,l}} = \sqrt{\frac{\sum_k \bar{m}_{j,k,l}^2 - 8(\bar{M}_{j,l})^2}{7}}$$

where j has been fixed at the index for a concurrent data length of 7000 hours. There were 2 to 5 non-overlapping subsets of data set at a concurrent data length of 7000 hours. The results are presented in Table 4.

Results - Effect of Data Length

The length of the concurrent data set determines the standard deviation of the metrics. In general, the longer the concurrent data length is, the smaller the standard deviation of the metric is. The general trend that is observed with all of the data sets and all of the metrics is illustrated in Figure 1 for mean wind speed using the BUZM3-44013 paired data set. Beyond concurrent data lengths of 6000 to 7000 hours (8 to 9.5 months), the standard deviation of the metrics does not decrease significantly. It would be expected that the standard deviation should decrease as the square root of the concurrent data length, if the MCP model that is used correctly models the relationship between the two data sets. However, the effect of un-modeled seasonal and other characteristics affect the results. In the example in figure 1, the Variance Ratio method appears to have a lower standard deviation, but the standard deviations of each of the methods are generally about the same magnitude when predicting mean wind speed.

Results - Example Using Logan - 44013 Data

Figures 2 to 5 and Figures 7 and 8 show the values of the metrics as a function of concurrent data length for the Logan-44013 data set. The graphs show the values of the various normalized metrics, $\bar{m}_{j,k,l}$. The other data sets showed similar behavior.

Results - Mean Long-term Wind Speed

The average of the predicted mean long-term wind speeds over all of the data sets was within 0.4% of the true wind speed for all methods but vector regression, as seen in Figure 2 and Table 4. The Vector method approach consistently underestimated the mean wind speed. The degree of underestimation depended on the cross correlation between the two data sets. The standard deviation of the means of the three other methods were all less than 0.7% of the true mean, indicating that the mean for each of the different data sets was very close to the true mean. Thus the Linear Regression, Mortimer and Variance Ratio methods all provided unbiased estimates³ of the mean wind speed, with very low standard deviations. The vector method provides biased estimates of the mean wind speed with a very large standard deviation.

Results - Wind Speed Distributions

Predicted wind speed distributions were compared using both Weibull parameters and Chi-squared goodness of fit measures. The Weibull parameters are a two-parameter set of measures that describe the shape of the wind speed distribution. The Weibull shape factor, k , determines the shape of the distribution. The Weibull scale factor, c , depends on k and the mean wind speed.

³ For this study, estimates for which the mean is within 1.8 standard deviations of 1.0 are considered to be unbiased. Estimates within between 1.8 and 2.2 standard deviations of 1.0 are considered to be possibly biased and estimates that are greater than 2.2 standard deviations away from 1.0 are considered to be biased estimates.

However, for typical wind speed distributions, c is primarily a function of the mean wind speed. The Chi-squared goodness of fit measures the degree of agreement between the predicted and true wind speed distributions over the whole distribution.

Only the Variance Ratio method appears to give an unbiased estimate of k , with a very low standard deviation over all of the data sets (see Figure 3 and Table 4). The Mortimer method gives a possibly biased estimate for k , with a very low standard deviation. The Linear Regression method provides unbiased estimates of k but with a very large standard deviation. For example, the mean k value with the Linear Regression method was 1.37 times the correct k parameter for the data sets. Finally, the Vector method also provides unbiased estimates of k , with a large standard deviation.

The Weibull scale parameter results reflect the mean wind speed results, as shown in Figure 4 and Table 4. The Linear Regression, Mortimer and Variance Ratio methods give unbiased estimate of c with very low standard deviations over all of the data sets. Again, the Vector method significantly underestimates c , with a standard deviation on the order of 10 times that of the results of the other methods.

The Chi-squared goodness of fit measures for the Mortimer and Variance Ratio methods provide very low values with low standard deviations (see Figure 5 and Table 4). This indicates that all of the speed distributions determined using these methods are very similar to the true ones. The mean of the Chi-squared goodness of fit measures for the Linear Regression and Vector methods are much higher than those of the Mortimer and Variance Ratio methods and have very large standard deviations, indicating that these methods do not predict the correct speed distributions. These results are consistent with the Weibull parameter results. A low Chi-squared goodness of fit measure requires both correct k and c values.

An example of one of the set of predicted probability distributions for the BUZM3-44013 data set and a data segment length of 7000 hours, using the Linear Regression method, is illustrated in Figure 6. For this case, the true Weibull k was 1.95 and predicted Weibull k is 2.62. The actual Weibull c was 6.62 and the predicted c was 6.70. Thus, the normalized mean Weibull k is 1.34 and the normalized mean Weibull c is 1.01 for this particular 7000-hour segment. The mean Chi-squared goodness of fit measure for this sample was 0.132. For predicting the distribution, regression does not do well because the variance of the predicted values is less than that of the observed values.

Results - Capacity factor

Figure 7 and Table 4 show the capacity factor results. The capacity factor (CF) results reflect the consequences of the ability of each method to predict the correct mean and wind speed distribution, although CF is less sensitive to an incorrect wind speed distribution above rated wind speed, where the power curve is flat. The Linear Regression, Mortimer and Variance Ratio methods all provide unbiased estimates of the CF, with low standard deviations. The means of the Mortimer and Variance Ratio methods are within 0.6% of the correct mean CF but the standard deviation using the Mortimer method is 1.2% of the mean and that of the Variance Ratio is 2.9% of the mean. The Linear Regression method tends to slightly underestimate the CF (by about 4%), with a standard deviation of 2.7% of the correct mean. This is probably due to the overestimation of the Weibull shape factor, which would result in more wind speeds near the mean wind speed. The mean of the CF estimates using the Vector method is 76% of the correct mean, but the standard deviation of the estimates is 13% of the mean. Thus the Vector method yields possibly biased estimates of CF, with very a large standard deviation.

Comments on Linear Regression

Standard linear regression will always give predictions with smaller variance than that of the observations. If, in addition, there are errors in the x values, then the predicted slope will have a negative bias and the offset will have a positive bias. A lower slope is associated with lower variance of the predicted values. One could attempt to determine an error corrected slope. The problem with fitting an error corrected slope is determining the error variances for x and y such that the combined error explains all the error around the linear relationship. The error variance includes not just measurement error but also the stochastic differences between the two sites. The variance ratio model is equivalent to using an error corrected slope assuming the error variances have the same ratio as the variances of the data. These error assumptions seem at least reasonable and result in better predictions than standard linear regression.

Linear regression parameter estimates can be sensitive to outliers. One could consider robust estimation techniques but these are unlikely to improve predictions since there are not really any outliers in the data. The problems with the data include missing values, particularly due to ice, what to do with small wind speeds that are hard to measure, and measurements that are discrete rather than continuous. None of these problems is solved by robust methods.

Comments on the Vector Method

Only the Vector method attempts to predict the wind direction distribution at the target site as part of predicting the wind speed. As implemented here, all of the other methods assume the wind direction at the target site is the same as that of the reference site. Thus the Chi-squared goodness of fit measures for the Linear Regression, Mortimer and Variance Ratio methods are just a measure of the difference in the direction distributions between the two sites (see Figure 8). In all cases, the Vector method provided a better prediction of the target site wind directions than using the reference site wind directions. The most extreme direction differences were found in the Century Tower - Brodie Mountain data. A plot of the respective direction values is shown in Figure 9. The Chi-squared goodness of fit measure for the Vector method (Figure 10) is significantly below that of the other methods. The superior performance of the Vector method is not as apparent in the other paired sites. This is because of the large amount of skewed winds at the reference site with respect to the target site in this data set.

The Vector method consistently under-predicts the wind speed due to the smaller variance of the predicted data compared to the observed data. This is a general systematic property of this approach. Writing the wind speed components (either predicted or observed) as a difference from the mean, $y_1 = \bar{y}_1 + e_1$ and $y_2 = \bar{y}_2 + e_2$, the wind speed is:

$$y = \sqrt{(\bar{y}_1 + e_1)^2 + (\bar{y}_2 + e_2)^2}$$

If the variance of e_1 and e_2 are zero (all observations are (\bar{y}_1, \bar{y}_2)), the mean wind speed is:

$$\bar{y} = \sqrt{\bar{y}_1^2 + \bar{y}_2^2}$$

As the variances of e_1 and e_2 increase, \bar{y} increases. Since the predicted values have smaller variance than the observed values, \bar{y} is lower for the predicted values than for the observed values. The magnitude of the bias depends on many factors (primarily the coefficient of variation of the wind components and how well the model fits the data for the two components).

Results and Conclusions

A number of conclusions can be reached from these investigations:

- As found by other researchers, the most useful data length is about 9 months or more, with little improvement in the standard deviation of estimates after that period.
- The MCP algorithms investigated here did not include seasonal terms. It is possible that the inclusion of seasonal terms might change the conclusions about useful data lengths and/or result in smaller standard deviations.
- Only the Variance Ratio method seems to give consistently reliable predictions of all of the metrics. The Variance Ratio method works remarkably very well given that it only uses a two parameter model for each direction sector.
- The Mortimer method gives unbiased predictions of all of the metrics except for estimates of the Weibull k parameter. In spite of this, the wind speed distributions, as indicated by the Chi-square measure, are as good as those of any of the other approaches, indicating that the Mortimer method, also, produces reliable results. Another choice of default value, when there is not enough binned data to determine a ratio for any given bin and sector, might improve the results for k .
- The method referred to as the Linear Regression method suffers from the characteristics of linear regression where the variance of the predicted wind speed is less than that of the observed wind speed, resulting in unbiased estimations of the mean wind speed but incorrect wind speed distributions.
- The Vector method compounds the bias associated with linear regressions, but does predict the wind direction distribution relatively well, even at a site with significantly skewed directions with respect to the reference site.
- Although the Vector method appears to be reasonable for modeling wind direction effects, more work is needed to determine the best method to predict wind direction.
- Finally, while the distance between the paired sites does not seem to affect the conclusions presented here, a thorough investigation of the consequences of the distance between sites on MCP estimates has not been undertaken.

Acknowledgements

This work has been conducted with the support from the Massachusetts Division of Energy Resources and the Massachusetts Technology Collaborative.

References

1. Derrick A., Development of the measure-correlate-predict strategy for site assessment, *Proc. BWEA*, 1992.
2. Derrick A., Development of the Measure-Correlate-Predict Strategy for Site Assessment, *Proc. EWEC*, 1993.
3. Joenson A., Landberg L., Madsen, H., A New Measure-Correlate-Predict Approach for Resource Assessment, *Proc. EWEC*, 1999
4. Nielsen M., Landberg L., Mortensen N. G., Barthelmie, R. J., Joensen A., Application of Measure-Correlate-Predict Approach for Wind Resource Measurement, *Proc. EWEA*, 2001.
5. Riedel V., Strack M., Robust approximation of functional relationships between meteorological data: Alternative measure-correlate-predict algorithms, *Proc. EWEA*, 2001
6. Landberg L, Mortenson NG, A comparison of physical and statistical methods for estimating the wind resource at a site, *Proc. BWEA*, 1993

7. Mortensen, N. G., Landberg, L. Troen, I., Petersen, E. L., Wind Atlas Analysis and Application Program. Vol. 2:Users Guide. Riso-I-666(ENN)(v.2). Riso National Laboratory, Roskilde, Denmark. 133 pp.
8. Woods J. C. and Watson S. J., A new matrix method of predicting long-term wind roses with MCP, *Journal of Wind Engineering and Industrial Aerodynamics*, Vol 66, n. 2, Feb 1997, pp 85-94.
9. Vermuelen P. E. J., Marijanyan A., Abrahamyan A., den Boon J. H., Application of Matarix MCP Analysis in Mountainous Armenia, *Proc. EWEA*, 2001.
10. Mortimer A. A., A new correlation/prediction method for potential wind farm sites, Mortimer, *Proc. BWEA*, 1994
11. Draper N. R., Smith H., *Applied Regression Analysis*, John Wiley and Sons, Inc. New York, 1966.

Figure Captions

Figure 1. Sample standard deviation of long term mean wind speed metric as a function of concurrent data length, using the BUZM3-44013 data set. Methods shown include: Linear Regression (solid line), Vector (long dashes), Mortimer (dots) and Variance Ratio (short dashes). Beyond concurrent data lengths of 6000 to 7000 hours (8 to 9.5 months), the standard deviations of the metrics do not decrease significantly.

Figure 2. Results of mean wind speed metric, m_1 , for the four methods, including true value (thin solid line), using the Logan-44013 data set. Methods include: Linear Regression (solid line), Vector (short dashes), Mortimer (dots) and Variance Ratio (long dashes). All methods except Vector method provided unbiased estimates of the mean wind speed.

Figure 3. Results of Weibull k metric, m_2 , for the four methods, including true value (thin solid line), using the Logan-44013 data set. Methods include: Linear Regression (solid line), Vector (short dashes), Mortimer (dots) and Variance Ratio (long dashes). Only the Mortimer and Variance Ratio methods provide unbiased estimates of k with small standard deviations.

Figure 4. Results of Weibull c metric, m_3 , for the four methods, including true value (thin solid line), using the Logan-44013 data set. Methods include: Linear Regression (solid line), Vector (short dashes), Mortimer (dots) and Variance Ratio (long dashes). All methods except the Vector method provide unbiased estimates of c.

Figure 5. Results of speed Chi-square metric, m_4 , for the four methods, using the Logan-44013 data set. Methods include: Linear Regression (solid line), Vector (short dashes), Mortimer (dots) and Variance Ratio (long dashes). The Mortimer and Variance Ratio methods result in the lowest values of this metric.

Figure 6. Comparison of predicted wind speed distribution using the Linear Regression method (thin solid line) and actual wind speed distribution (dashed line), using the BUZM3-44013 data set. The predicted distribution using the Linear Regression method indicates a smaller standard deviation in values than the actual distribution.

Figure 7. Results of capacity factor metric, m_6 , for the four methods, including true value (thin solid line), using the Logan-44013 data set. Methods include: Linear Regression (solid line), Vector (short dashes), Mortimer (dots) and Variance Ratio (long dashes). The Linear Regression, Mortimer and Variance Ratio methods all provide unbiased capacity factor estimates. The Vector method provides biased results.

Figure 8. Results of direction Chi-square metric, m_5 , for the four methods, using the Logan-44013 data set. Methods include: Linear Regression, Mortimer and Variance Ratio (all the same solid line) and Vector (short dashes). Only the Vector method improves on the assumption that the wind direction distribution at the target site is that of the reference site.

Figure 9 Wind direction relationship between Century Tower and Brodie Mtn. data. The data indicate significantly skewed winds, under some circumstances, at the reference site with respect to the target site in this data set.

Figure 10. Direction Chi-square metric, m_6 , from Century Tower - Brodie Mountain data. Methods include: Linear Regression, Mortimer and Variance Ratio (all the same solid line) and Vector (short dashes). The graph illustrates the superior performance of the Vector method for predicting wind direction, compared with the assumption that the wind directions at the two sites are the same.

Tables

<i>Reference</i>	<i>Approach</i>	<i>Data processing</i>	<i>Method of Evaluation</i>
Derrick [1, 2]	Speed: linear fit Direction: None or polynomial fit	Data filtering	Fit length: annual mean, error, %, annual energy error, 15 yrs. of data Overall method: 15 yr., sector average
Nielsen et al. [4]	Linear transformation: using u, v	Band-limited correlation Different averaging periods	Averaging period: Pearson's r for each sector fit Overall method: u^3 , 7 yrs. of data
Riedel et al. [5]	Speed: Quadratic fit Direction: Chebyshev polynomials	Dynamic sector positioning Minimization of residuals and prediction errors	Fit length: median, quartiles of energy yield error (using power curve) Visual comparison of wind roses Onshore: 1 yr, multiple sites Offshore: 6 yrs
Landberg and Mortenson [6]	Speed: Linear fit Direction: None	Cross-correlation vs. distance	Fit length: scatter of predicted mean with different experiments 2 yrs. of data
Woods and Watson [8]	Speed: Linear fit Direction: Matrix bins	Matrix bin count cut-off level	Direction bins: sector means, sector counts. 100 days of data
Vermeulen et al. [9]	Speed: Linear fit Direction: Matrix bins	Consideration of criteria for using method	Direction bins: sector means, sector counts. 1 yr. of data
Joensen et al. [3]	Speed: Quadratic fit, to address atmospheric stability Direction: constant offset	Rotate axes for highest correlation coefficient Fit for direction neighborhood	Visual comparison of speed distributions, 2 yrs. of data
Mortimer [10]	Speed: Binned mean and standard deviation of ratios Direction: None		Speed distributions vs. linear models 9 mos. of data

Table 1. Overview of seven MCP approaches found in the literature, including computational approach, data processing details and the reported method of evaluation.

<i>Designation</i>	<i>Location</i>	<i>Location</i>	<i>Source</i>	<i>Latitude</i>	<i>Longitude</i>	<i>Elevation, m</i>
Logan Airport	Boston, MA	Coastal	NCDC	42°21' N	71° 0'W	3
Buoy 44013	Boston Harbor	Offshore	NOAA	42.35 N	70 41' W	0
Platform BUZM3	Buzzards Bay, MA	Offshore	NOAA	41.40 N	71 2' W	0
Petersburg	Petersburg, ND	Plains	ND State Gov't	47 59.2' N	98 0.58' W	477
Olga	Olga, ND	Plains	ND State Gov't	48 46.8' N	98 02.3' W	475
Alfred	Alfred, ND	Plains	ND State Gov't	46 35.3' N	99 0.77' W	631
Century Tower	Western MA	Ridge top	UMass	42 10' N	73 19' W	~500
Brodie Mt.	New Ashford, MA	Ridge top	UMass	42 36' N	73 16' W	~500
Burnt Hill	Heath, MA	Ridge top	UMass	42 39' N	72 48'W	~500

Sources: ND State Gov't: <http://wind.undeerc.org/scripts/wind/NDwindsites.asp>
 UMass: <http://www.ecs.umass.edu/mie/labs/rerl/research/MassData.html>
 NOAA: <http://www.ndbc.noaa.gov/index.shtml>
 NCDC: National Climactic Data Center, Cornell, NY.

Table 2. Data site characteristics for the nine data sites used in this analysis, including location, data source and elevation.

Reference Site	Target Site	Concurrent Data Length	Cross correlation	
			Max.	Lag at Max
Logan	BUZM3	5+ years	0.65	0 hour
Logan	44013	4.5 years	0.70	0 hour
BUZM3	44013	4.5 years	0.75	0 hour
Petersburg	Alfred	2 years	0.73	1 hour
Petersburg	Olga	2 years	0.82	0 hour
Alfred	Olga	2.5+ years	0.63	1 hour
Century Tower	Brodie Mt.	2 years	0.66	0 hour
Brodie Mt.	Burnt Hill	1.7 years	0.63	1 hour

Table 3. Pairs of sites used for the analysis of MCP algorithms, including concurrent data lengths and cross correlations.

Metric		Linear Regression	Mortimer	Variance Ratio	Vector Regression
Long-term mean	Mean	0.999	0.998	1.004	0.857
	Std. Dev.	0.004	0.005	0.007	0.059
Weibull k	Mean	1.374	0.982	1.001	0.983
	Std. Dev.	0.131	0.009	0.004	0.091
Weibull c	Mean	0.991	0.998	1.003	0.860
	Std. Dev.	0.009	0.005	0.006	0.061
Speed Distribution Chi- square	Mean	1.000	0.048	0.064	1.106
	Std. Dev.	0.876	0.028	0.046	0.866
Capacity Factor	Mean	0.958	1.006	1.005	0.757
	Std. Dev.	0.027	0.012	0.029	0.129

Table 4. Summary of the analysis of results for the four MCP methods. The table shows statistics of normalized metrics across all 8 data sets, including means and standard deviations. In general Mortimer and Variance Ratio methods provide unbiased results with low standard deviations. A value of zero for the speed distribution Chi-square metric would indicate correctly predicted wind speed distributions. For all of the other metrics, a value of 1.000 would indicate correct predictions.

Figures

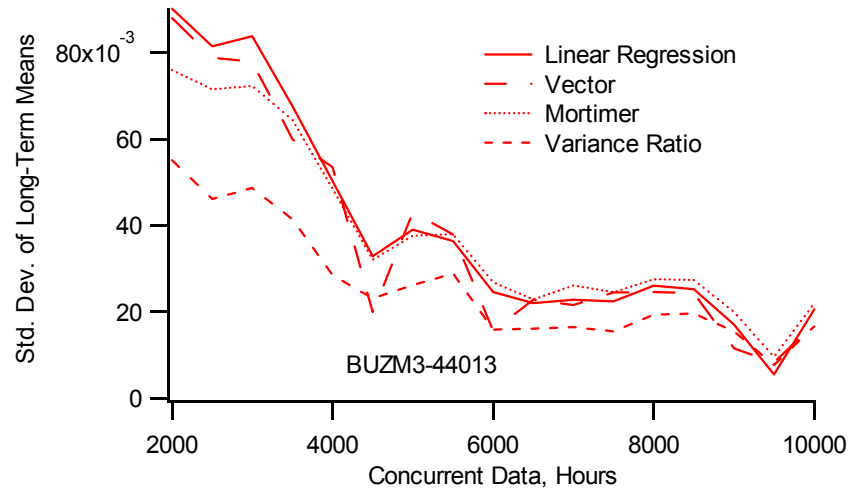


Figure 1. Sample standard deviation of long term mean wind speed metric as a function of concurrent data length, using the BUZM3-44013 data set. Methods shown include: Linear Regression (solid line), Vector (long dashes), Mortimer (dots) and Variance Ratio (short dashes). Beyond concurrent data lengths of 6000 to 7000 hours (8 to 9.5 months), the standard deviations of the metrics do not decrease significantly.

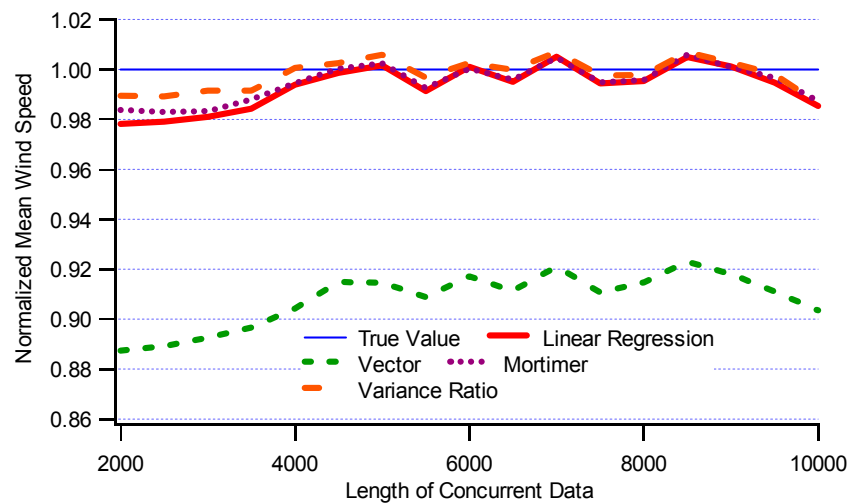


Figure 2. Results of mean wind speed metric, m_I , for the four methods, including true value (thin solid line), using the Logan-44013 data set. Methods include: Linear Regression (solid line), Vector (short dashes), Mortimer (dots) and Variance Ratio (long dashes). All methods except Vector method provided unbiased estimates of the mean wind speed.

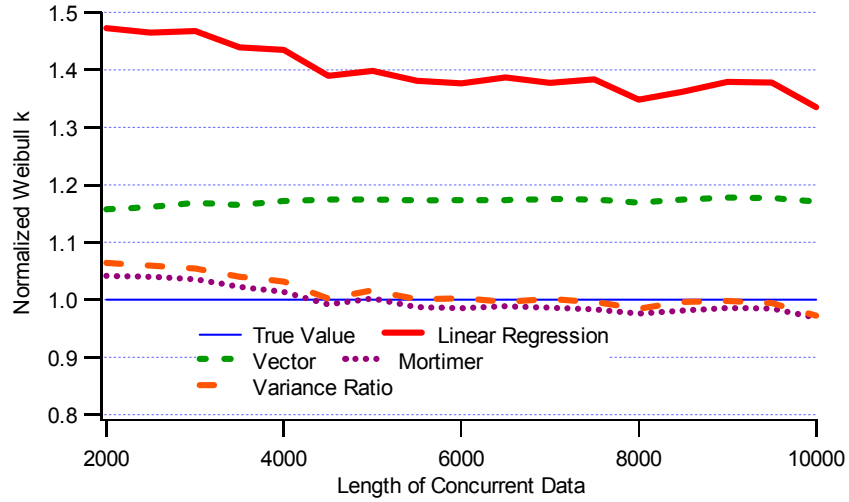


Figure 3. Results of Weibull k metric, m_2 , for the four methods, including true value (thin solid line), using the Logan-44013 data set. Methods include: Linear Regression (solid line), Vector (short dashes), Mortimer (dots) and Variance Ratio (long dashes). Only the Mortimer and Variance Ratio methods provide unbiased estimates of k with small standard deviations.

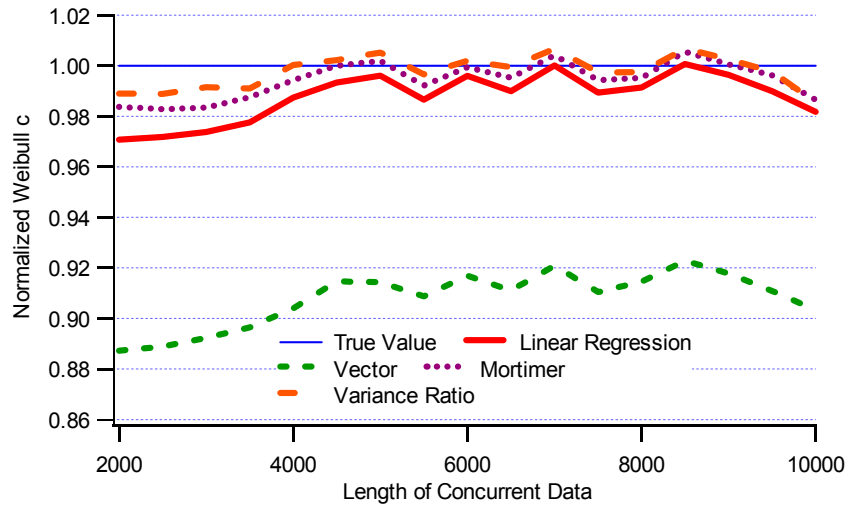


Figure 4. Results of Weibull c metric, m_3 , for the four methods, including true value (thin solid line), using the Logan-44013 data set. Methods include: Linear Regression (solid line), Vector (short dashes), Mortimer (dots) and Variance Ratio (long dashes). All methods except the Vector method provide unbiased estimates of c .

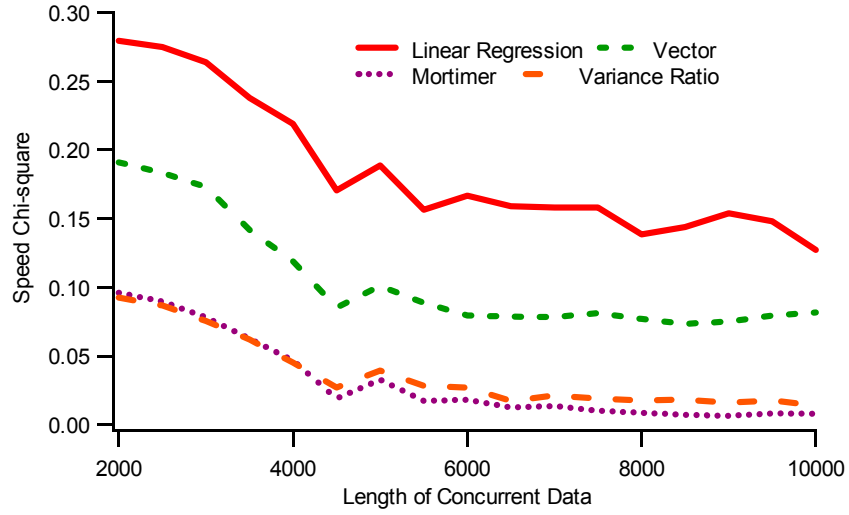


Figure 5. Results of speed Chi-square metric, m_4 , for the four methods, using the Logan-44013 data set. Methods include: Linear Regression (solid line), Vector (short dashes), Mortimer (dots) and Variance Ratio (long dashes). The Mortimer and Variance Ratio methods result in the lowest values of this metric.

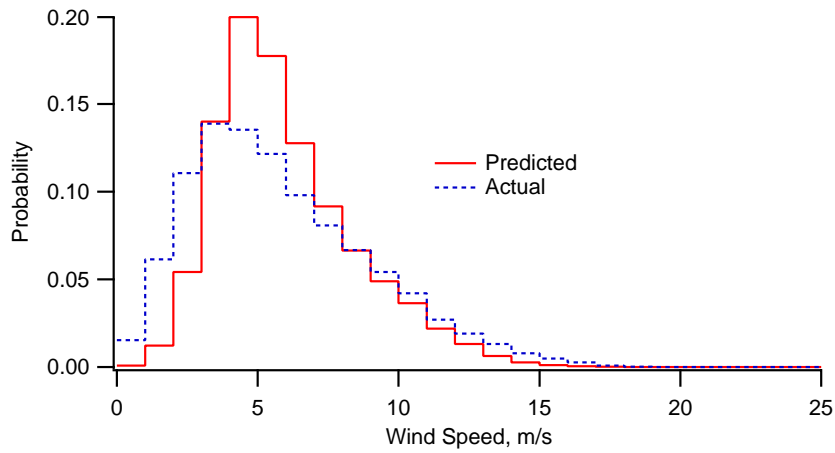


Figure 6. Comparison of predicted wind speed distribution using the Linear Regression method (thin solid line) and actual wind speed distribution (dashed line), using the BUZM3-44013 data set. The predicted distribution using the Linear Regression method indicates a smaller standard deviation in values than the actual distribution.

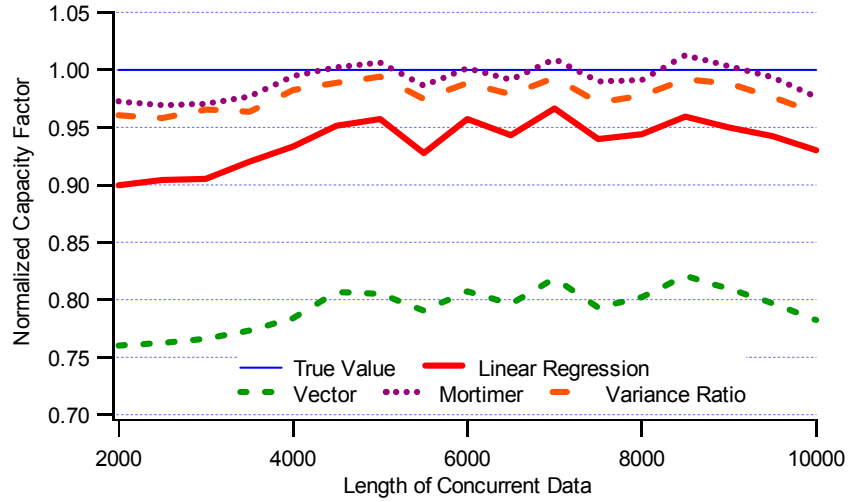


Figure 7. Results of capacity factor metric, m_6 , for the four methods, including true value (thin solid line), using the Logan-44013 data set. Methods include: Linear Regression (solid line), Vector (short dashes), Mortimer (dots) and Variance Ratio (long dashes). The Linear Regression, Mortimer and Variance Ratio methods all provide unbiased capacity factor estimates. The Vector method provides biased results.

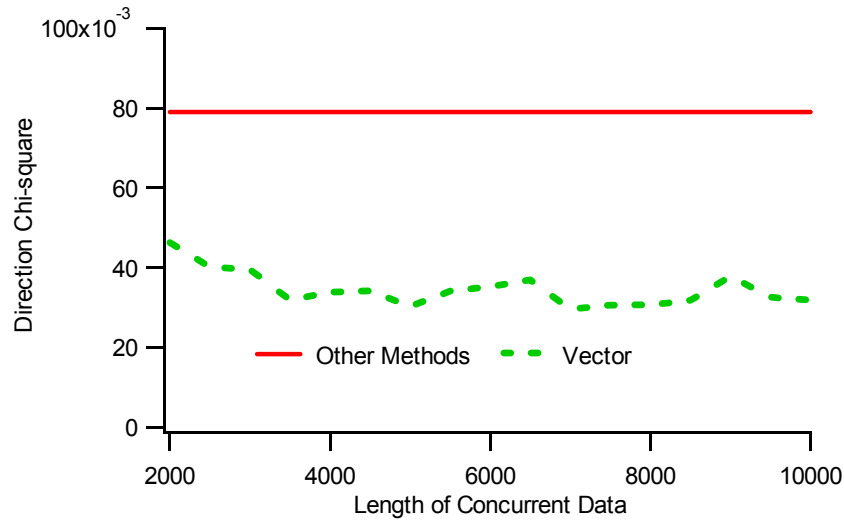


Figure 8. Results of direction Chi-square metric, m_5 , for the four methods, using the Logan-44013 data set. Methods include: Linear Regression, Mortimer and Variance Ratio (all the same solid line) and Vector (short dashes). Only the Vector method improves on the assumption that the wind direction distribution at the target site is that of the reference site.

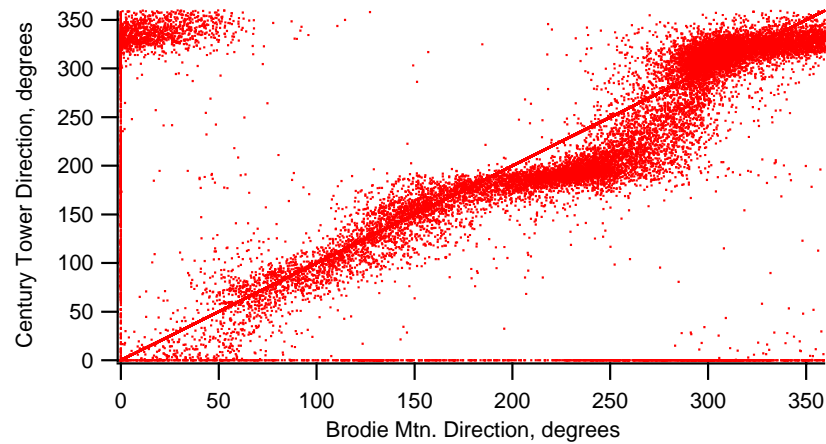


Figure 9 Wind direction relationship between Century Tower and Brodie Mtn. data. The data indicate significantly skewed winds, under some circumstances, at the reference site with respect to the target site in this data set.

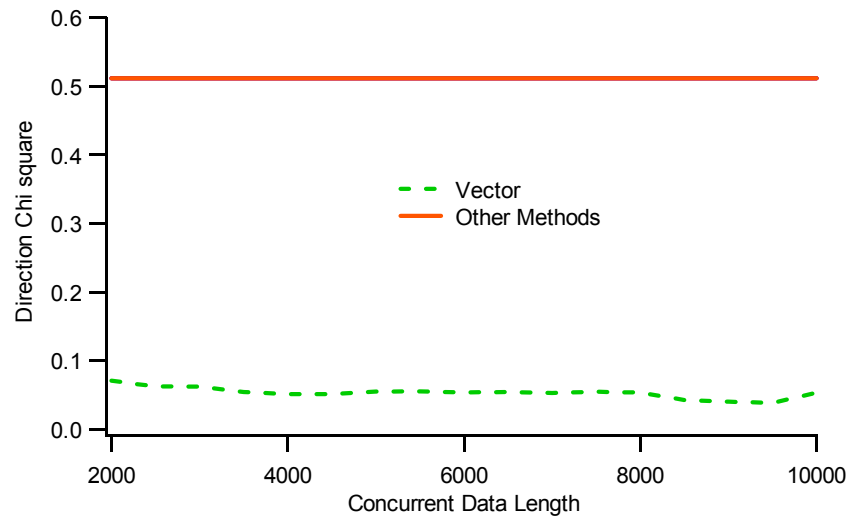


Figure 10. Direction Chi-square metric, m_6 , from Century Tower - Brodie Mountain data. Methods include: Linear Regression, Mortimer and Variance Ratio (all the same solid line) and Vector (short dashes). The graph illustrates the superior performance of the Vector method for predicting wind direction, compared with the assumption that the wind directions at the two sites are the same.