

Modelling power output from correlated wind farms

Bill Whiten¹Barry McDonald²Ray Bedford³

(Received 30 January 2012; revised 22 November 2013)

Abstract

Wind data from fifteen potential wind farm sites covering New Zealand were analysed. The data was fitted with a generalised gamma distribution and transformed to a Gaussian distribution. A regression was used to model the temporal correlation and the spatial correlation was modelled with a multivariate Gaussian distribution. This enabled prediction of the distribution of power generated and of the changes in power over ten minute intervals. For predicting up to two hours ahead, prediction of wind speed was moderate, but the steep relation of wind to power meant that power predictions were not good.

Subject class: 62P30

Keywords: Wind prediction, Multiple wind farms

<http://journal.austms.org.au/ojs/index.php/ANZIAMJ/article/view/5194> gives this article, © Austral. Mathematical Soc. 2013. Published November 28, 2013, as part of the Proceedings of the 2011 Mathematics and Statistics in Industry Study Group. ISSN 1446-8735. (Print two pages per sheet of paper.) Copies of this article must not be made otherwise available on the internet; instead link directly to this URL for this article.

Contents

1	Introduction	M105
2	A probabilistic model of wind data	M106
3	Simulation of power generation probabilities	M117
4	Prediction using simulation	M118
5	Prediction by data-matching	M120
6	Prediction using a local model	M127
7	Prediction using power	M127
8	Additional effects	M132
9	Conclusions	M137
A	Notation	M139

1 Introduction

The MISG project sponsor Transpower is the owner of the New Zealand national electric power grid that delivers power from the power stations to the local suppliers of electricity to the end users. Transpower is also responsible for balancing the demands of power users with the supply of electricity.

Wind power, although available day and night, is variable and is thus a source of imbalance on the power network. An understanding of the nature of its variation is needed, particularly as more wind farms are added to the network.

One problem of interest is to predict wind velocity and hence power from

a given site. A second issue, and the primary focus of this report, is the combined effect of multiple generating sites. Transpower is interested in the size of changes in power that need to be accommodated by the electricity grid. As more wind farms are added to the grid the variation, measured in megawatts, in power generation will increase, but due to averaging effects the variation expressed as a proportion of total wind power will decrease.

This report examines the ability to predict the wind velocity and the power produced. A large set of wind speed data was provided for examination by this MISG project ([Electricity Authority, 2011](#)). The data is synthetic in the sense that it was generated for potential wind farm sites from numerical weather forecasts that had been matched to actual measurements where these were available. The use of synthetic data was necessitated by the fact measurements are not generally available from the wind farm sites and particularly not at the height of typical wind turbines, which is 80 m at the hub. The data consisted of wind speeds at ten minute intervals for 15 sites located around New Zealand, and extended over five years.

Section 2 outlines a method for simulating the distribution of wind data. Section 3 carries this simulation on to power generation. Sections 4, 5 and 6 examine how well the simulated data can predict a real data sequence. Section 7 considers modelling the power distribution based on previous power data without the use of wind data. Section 8 considers some additional effects that could improve the regression models.

2 A probabilistic model of wind data

This section demonstrates a simulation of the wind speed velocities at the various wind farm sites. Predictions of both the systematic wind behaviour and the random component within the wind are included in the simulation. The simulation can be used to make predictions of future wind speeds, and uses Monte Carlo methods to give accuracy estimates for the short term

predictions and to give insight into longer term statistical distributions. This work extends ideas presented by [Zakaria et al. \(2010\)](#) for combining rainfall data from two sites.

An examination of the data showed that, if we ignore temporal correlation, a reasonable fit to the wind speed distribution at each site is obtained with a generalised gamma distribution ([Wikipedia, 2011](#); [Stacy, 1962](#)). This was implemented as a gamma distribution fitted to the 2.5 power of the wind velocity. [Table 1](#) gives the shape and scale parameters for the gamma distribution, [Figure 1](#) shows these fits for the fifteen wind farm sites and [Figure 2](#) shows the difference in probability density between the data and the predicted distribution. There are noticeable differences at low velocities, but for power generation these are not important as no power is generated for winds below about 5 ms^{-1} . At higher velocities the fit appears to be good.

To model time dependence the data is simulated using a random walk and for this a scale more closely related to a Gaussian (Normal) distribution is used. The 2004 MISG ([Whiten and Tsoularis, 2004](#)) demonstrated the generation of a data sequence with very similar frequency distribution to the wind data, using a simple correlated sequence generator.

Using the generalised gamma distribution the data $x_{i,j}$ is transformed to a probability scale $u_{i,j}$ (0 to 1) which will be approximately uniform. This is then transformed to Gaussian distributed values $y_{i,j}$ for further analysis. In symbolic form this is (see notation [Appendix A](#) for symbol definitions)

$$u_{ij} = \Gamma(x_{ij}^{2.5}, \alpha_j, \beta_j), \quad (1)$$

$$y_{ij} = \Phi^{-1}(u_{ij}). \quad (2)$$

This Gaussian data is then suitable for the simulation as it is no longer restricted in range, and correlated Gaussian distributions are readily simulated.

As correlations between the Gaussian values (and also the original wind velocities) are positive or close to zero; nonnegative regression, resulting in only positive coefficients, is used to predict the ten minute Gaussian values

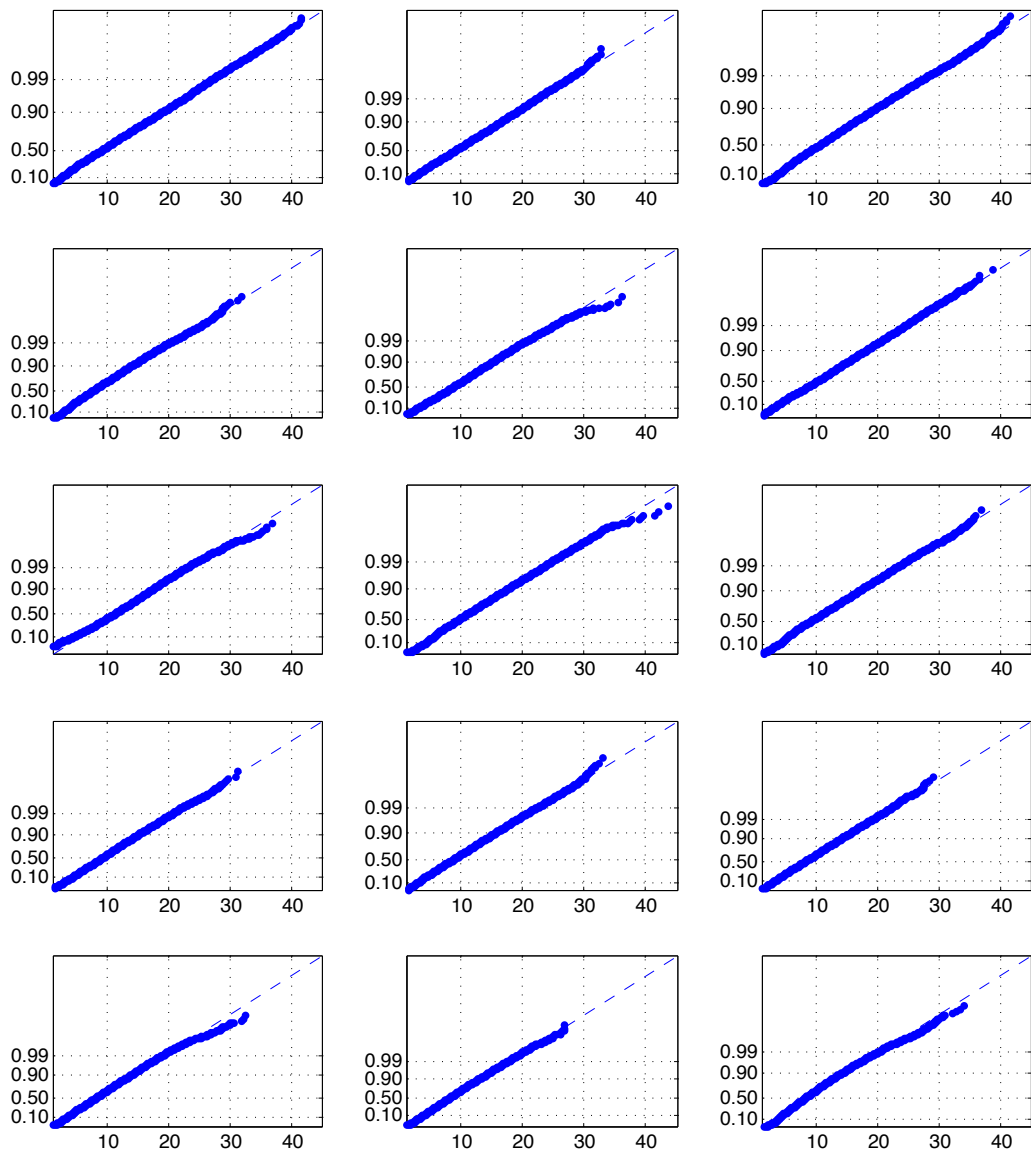


Figure 1: Fit to experimental data to power of 2.5 using gamma distributions (horizontal axis wind speed ms^{-1} , vertical axis probability).

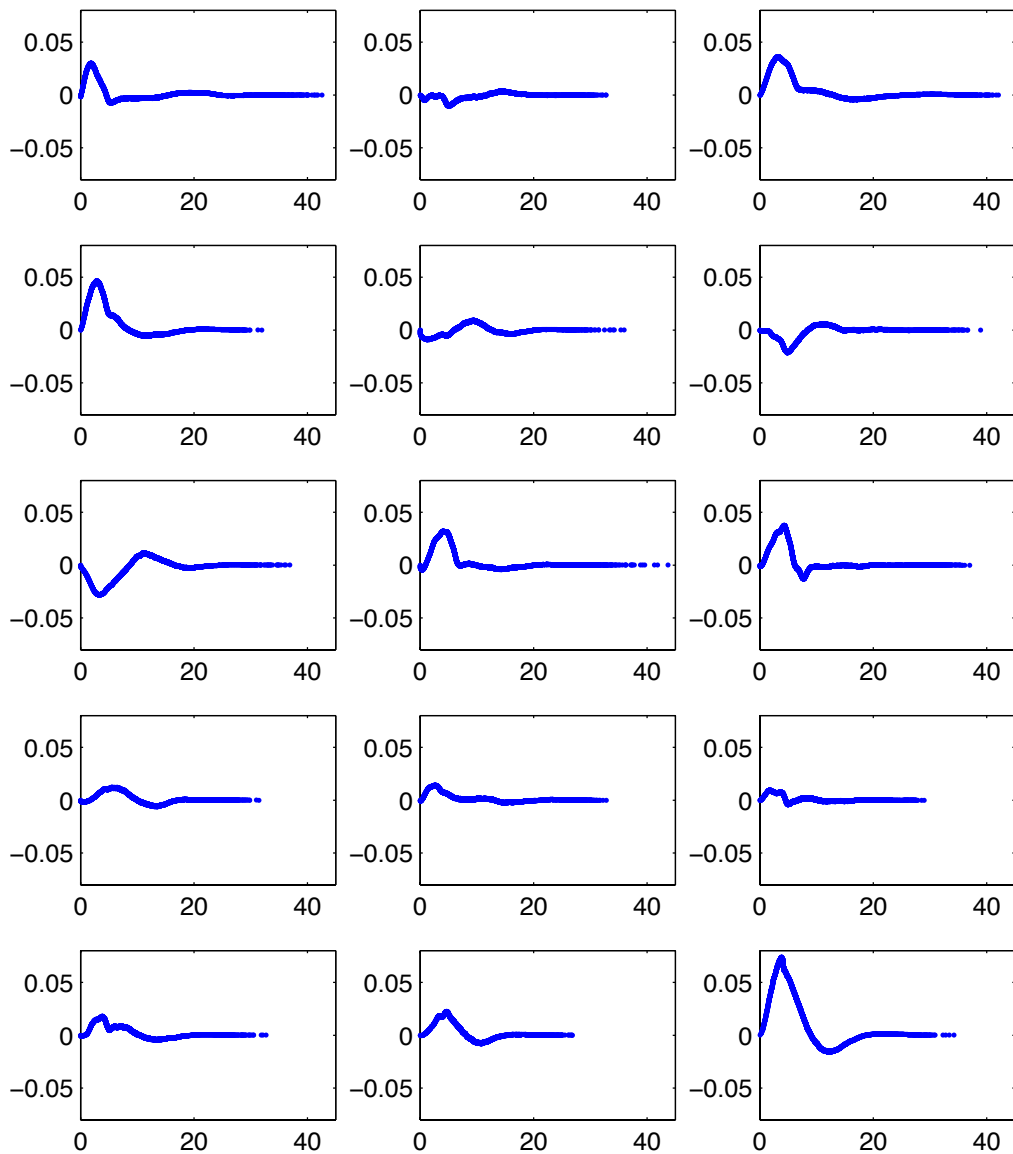


Figure 2: Probability difference (vertical axis) data (probability minus fitted probability), versus wind speed ms^{-1} .

Table 1: Parameters α shape and β scale for the gamma distributions for wind speed to the power 2.5, fitted to the fifteen wind farm sites (sites range over the length of New Zealand, see [Electricity Authority \(2011\)](#) for details). The last column gives regression coefficients $\hat{\alpha}_j$ from equation (3) for previous ten minute value.

Code	α (shape)	β (scale)	$\hat{\alpha}_j$ (3)
STH1	0.476	1215.9	0.904
STH2	0.708	608.7	0.915
STH3	0.589	1184.5	0.953
CTY1	0.602	508.3	0.959
CKS1	0.779	471.2	0.963
CKS2	0.758	770.4	0.883
CKS3	1.085	533.2	0.889
MWT1	0.682	792.1	0.862
MWT2	0.632	760.1	0.825
MWT3	0.895	437.6	0.893
CNI1	0.639	618.0	0.937
CNI2	0.764	402.0	0.922
NTH1	0.778	383.7	0.897
NTH2	0.933	290.8	0.843
NTH3	0.674	479.5	0.971

from previous values. For all sites the major term in the prediction is the previous ten minute value from the same site, with a typical coefficient of about 0.9. It was found, by some trial and error, that including in the regression, from the fifteen sites, the sum of the Gaussian values from three to five 10 minute intervals before the value being predicted was needed for a satisfactory result. The values two 10 minute intervals before were found not to improve predictions. That is, for each site $j = 1, \dots, 15$, and time i we

have the regression model

$$y_{i,j} = \hat{\alpha}_j y_{i-1,j} + \sum_{k=1}^{15} \hat{\beta}_{j,k} (y_{i-3,k} + y_{i-4,k} + y_{i-5,k}) + \epsilon_{i,j}, \quad (3)$$

where the coefficients $\hat{\alpha}_j$ and $\hat{\beta}_{j,k}$ minimise $\sum_{i,j} \epsilon_{i,j}^2$. For nearby wind farms it is tempting to interpret these extra predictors $y_{i-3,k} + y_{i-4,k} + y_{i-5,k}$ as representing the delay in wind travelling from one site to the next. However, in some cases the winds are also associated with distant sites, where it would take far more than half an hour for a wind gust to travel between them, even if the wind was blowing in the right direction. So perhaps these extra predictors are better interpreted as representing the passage (or absence) of widespread weather systems across the country. The inclusion of these additional predictors from the fifteen sites made only a modest reduction of between 2% and 0.3% for the root mean squared error in the regression, but made the distributions of the predicted wind speed and changes in wind speed much closer to the distributions given by the data. The exact reasons for this is a subject that needs further investigation. Table 1 gives the regression coefficients for prediction from the previous ten minute interval, and Table 2 gives the coefficients of the sum of the third to fifth lagged ten minute intervals.

Given the coefficient estimates from the autoregression equation (3), residuals are calculated as

$$e_{i,j} = y_{i,j} - \left[\hat{\alpha}_j y_{i-1,j} + \sum_{k=1}^{15} \hat{\beta}_{j,k} (y_{i-3,k} + y_{i-4,k} + y_{i-5,k}) \right]. \quad (4)$$

The residuals are compared to Gaussian quantiles in Figure 3. The plotted points are linear over the central three or four probits each side of zero, which means the tails are less than 0.1% of the data. Plotting the residuals against the wind speed, Figure 4 shows these tails occur at the very low wind speeds. The transformation from wind speed to Gaussian distribution considerably expands the low part of the scale, resulting in the larger residuals due to a lack of resolution. These values have been left in the data used for this

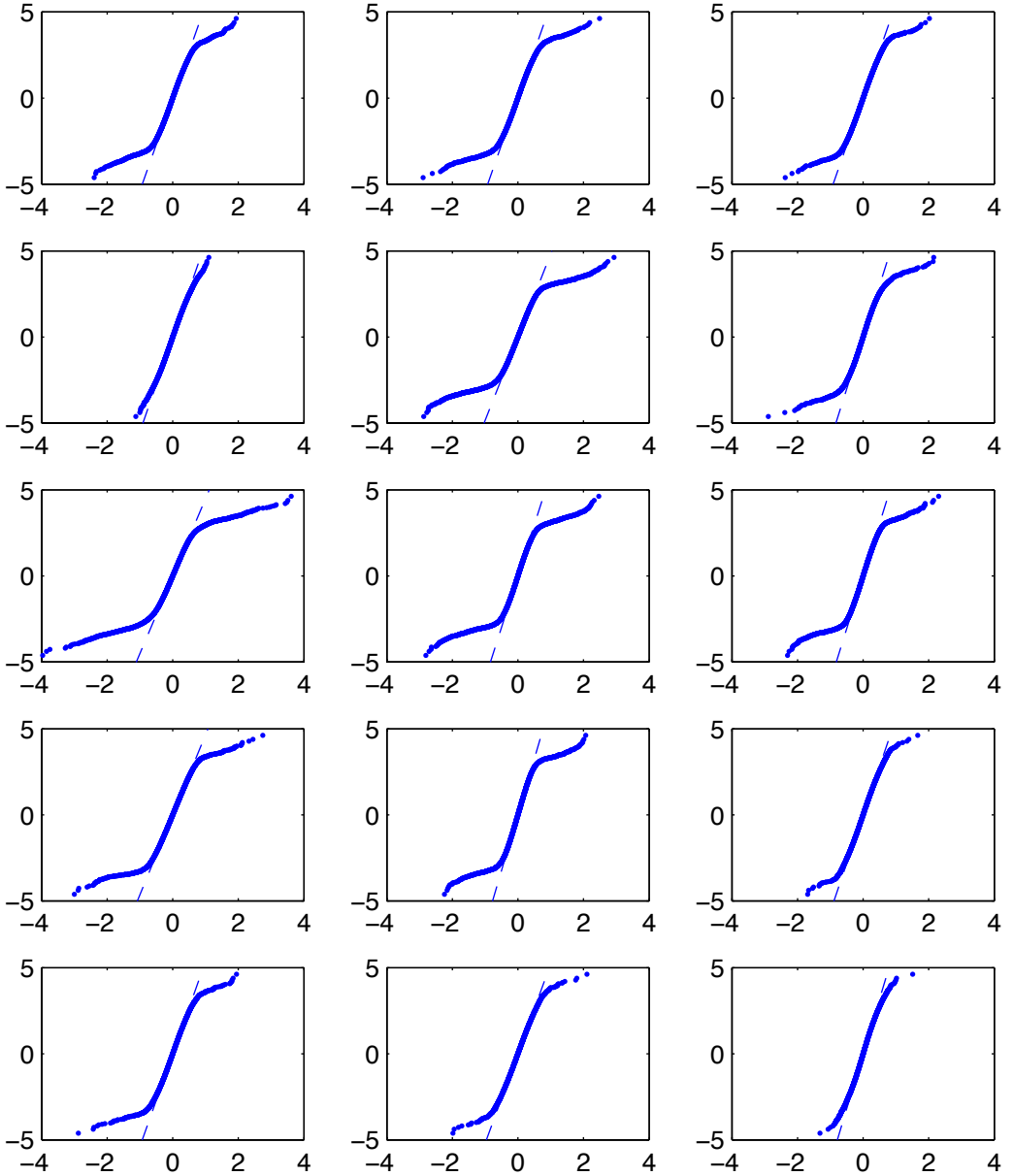


Figure 3: Gaussian (Normal) probability plots of residuals ms^{-1} , horizontal axis is residual ms^{-1} and vertical axis is probits.

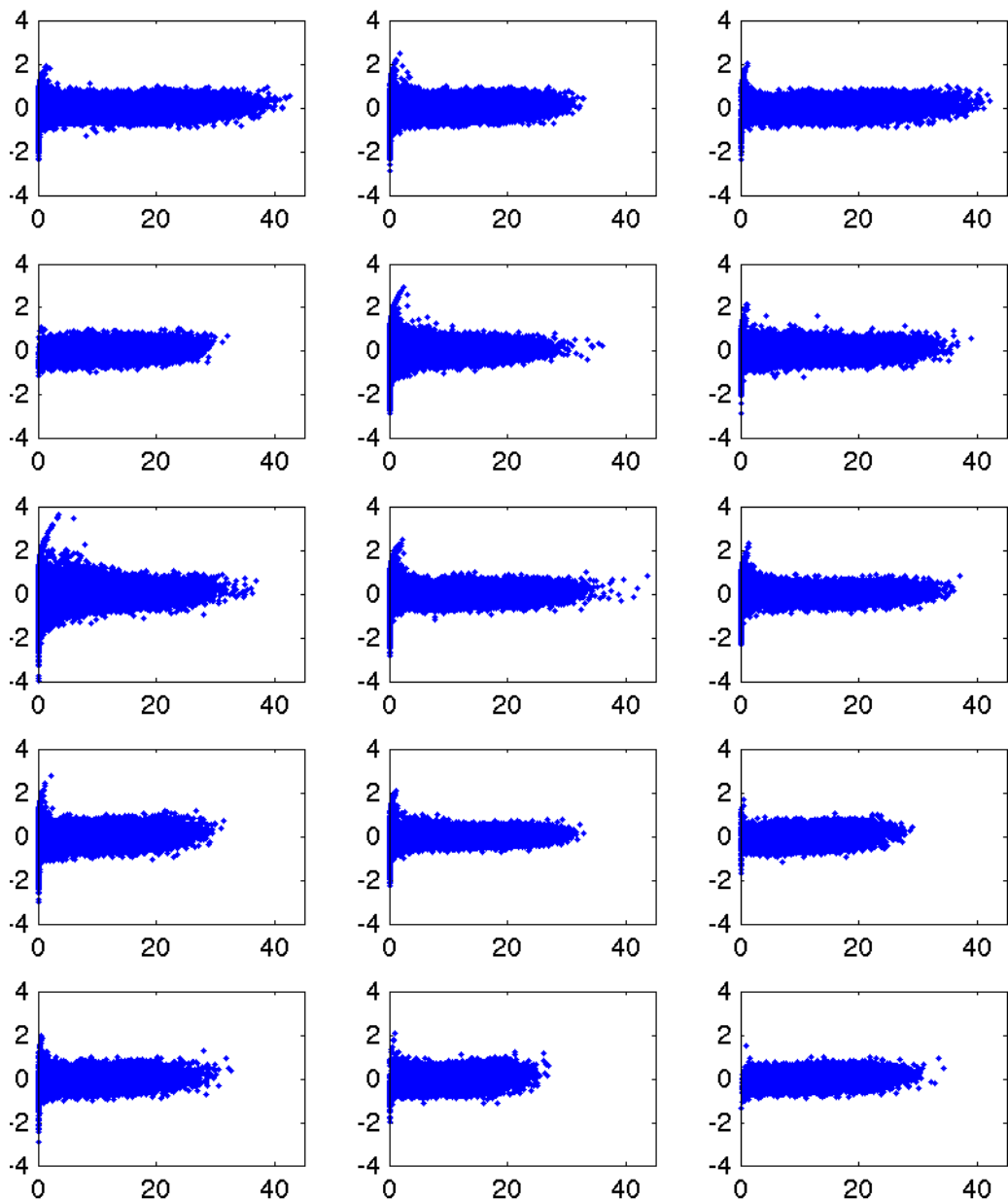


Figure 4: Residuals vertical axis ms^{-1} , versus wind speed horizontal axis ms^{-1} .

Table 2: Regression coefficient matrix $\mathbf{b}_{j,k}$ of row vector $\mathbf{y}_{i-2,k} + \mathbf{y}_{i-3,k} + \mathbf{y}_{i-4,k}$ times 100000. Each column gives coefficients for one wind site. The order of sites is the same as Table 1.

2369	124	200	58	0	0	0	0	0	0	0	0	0	0	0
251	1755	544	0	0	0	28	0	4	25	0	82	9	0	9
333	544	116	218	0	0	0	0	0	180	0	0	0	4	15
219	0	359	0	43	63	203	0	53	99	66	48	72	0	0
106	0	115	119	0	141	266	55	0	172	0	0	8	10	0
74	0	0	250	173	2851	361	0	0	0	109	0	103	0	0
119	0	0	286	512	678	1937	0	0	0	75	206	0	0	0
119	0	137	28	163	100	0	3058	482	424	122	92	0	0	0
71	37	0	113	77	0	143	340	4717	15	92	232	0	14	15
0	0	248	0	188	82	133	241	35	1986	263	190	62	73	92
0	0	0	46	22	135	137	0	67	186	1205	67	0	110	106
0	196	6	20	0	0	61	36	241	216	187	1444	79	66	83
0	119	0	94	84	0	0	0	0	115	0	70	2450	295	170
2	0	1	0	32	0	74	0	0	77	130	14	241	4165	409
0	0	0	0	0	0	0	0	0	3	91	22	93	420	0

demonstration of the simulation method. It is expected that their removal would not have a large effect.

The residuals \mathbf{e}_{ij} form the basis for stochastic simulation of the data. Using \mathbf{e}_{ij} directly generates the original data. The fifteen residuals for each ten minute time interval are correlated (Table 3) so this correlation needs to be simulated. There are several possible ways this can be done:

- M1. The time sequence of the calculated ten minute residuals can be randomised to produce a new sequence;
- M2. The residuals can be replaced by Gaussian samples generated with the covariance seen in the residuals $\text{cov}(\mathbf{e}_{ij})$;

Table 3: Covariance matrix for residuals times 10000. The order of sites is the same as Table 1.

406	21	21	15	22	-3	24	10	0	-58	-14	-24	-48	8	-50
21	400	33	12	11	-23	12	-18	9	32	-15	30	25	-4	-19
21	33	371	27	22	-42	-90	29	-2	40	-9	12	5	19	-22
15	12	27	358	27	23	42	19	16	30	20	20	22	-14	-21
22	11	22	27	535	26	45	40	15	40	12	-60	26	27	-18
-3	-23	-42	23	26	327	78	14	-15	11	20	-46	4	-9	-24
24	12	-90	42	45	78	644	-38	14	25	28	31	-118	25	-53
10	-18	29	19	40	14	-38	386	35	48	17	20	-9	8	-17
0	9	-2	16	15	-15	14	35	330	21	14	18	-4	5	-6
-58	32	40	30	40	11	25	48	21	509	30	29	22	26	3
-14	-15	-9	20	12	20	28	17	14	30	279	15	-14	18	3
-24	30	12	20	-60	-46	31	20	18	29	15	348	16	13	-1
-48	25	5	22	26	4	-118	-9	-4	22	-14	16	388	12	0
8	-4	19	-14	27	-9	25	8	5	26	18	13	12	399	9
-50	-19	-22	-21	-18	-24	-53	-17	-6	3	3	-1	0	9	275

M3. The residuals can be replaced by Gaussian samples generated with a covariance calculated to give the observed covariance of the Gaussian transformed wind speeds. This covariance is

$$\text{cov}(\mathbf{y}_{i,j}) - \text{cov} \left[\hat{\mathbf{a}}_j \mathbf{y}_{i-1,j} + \sum_{k=1}^{15} \hat{\mathbf{b}}_{j,k} (\mathbf{y}_{i-3,k} + \mathbf{y}_{i-4,k} + \mathbf{y}_{i-5,k}) \right]. \quad (5)$$

With the regression equations generated as above all three methods give very similar results for the chosen predictor variables. For other choices of the predictors the three methods gave differing results. In most cases the third method is preferred as it ensures the simulated Gaussian values have a correlation matrix that agrees with the experimental value. With limited

data the difference in method M3 may not be positive definite and in that case the method can not be used.

So the steps in a Monte Carlo regeneration of wind speeds are the following.

0. Initially five previous wind speed readings at ten minute intervals for the fifteen sites are needed to start the simulation. These are transformed to the Gaussian values. Alternatively the simulation could be started from zero or random values on the Gaussian scale, and the initial simulated values neglected. Call these initial values

$$\mathbf{y}_{1,j}^*, \mathbf{y}_{2,j}^*, \mathbf{y}_{3,j}^*, \mathbf{y}_{4,j}^* \text{ and } \mathbf{y}_{5,j}^*, \quad (6)$$

so that the first values to be simulated will be $\mathbf{y}_{6,j}^*$ and $\mathbf{x}_{6,j}^*$;

1. Simulate new error values using one of the three methods M1, M2 or M3, giving

$$\epsilon_{ij}^*; \quad (7)$$

2. Use the regression equations to predict the Gaussian values for the next ten minute interval and add the generated residual values, to get the simulated Gaussian values

$$\mathbf{y}_{i,j}^* = \hat{\alpha}_j \mathbf{y}_{i-1,j}^* + \sum_{k=1}^{15} \hat{\beta}_{i,k} (\mathbf{y}_{i-3,k}^* + \mathbf{y}_{i-4,k}^* + \mathbf{y}_{i-5,k}^*) + \epsilon_{i,j}^*; \quad (8)$$

3. Convert the simulated Gaussian values from step 2 to a uniform distribution

$$\mathbf{u}_{i,j}^* = \Phi(\mathbf{y}_{i,j}^*); \quad (9)$$

4. Convert the uniform distribution using the generalised gamma distribution, back to wind speeds

$$\mathbf{x}_{i,j}^* = \Gamma^{-1}(\mathbf{u}_{i,j}^*, \alpha_j, \beta_j)^{1/2.5}. \quad (10)$$

Steps 1 to 4 are repeated to produce the required number of time steps.

3 Simulation of power generation probabilities

We are interested in the effect of the correlated winds on the total power generated, so the wind speeds must be converted to power and summed over all the generating sites. For this demonstration the wind speed-power relation for a single turbine is assumed to be cubic until 12 ms^{-1} and then constant from 12 ms^{-1} until 25 ms^{-1} when it drops to zero. Furthermore the wind farm has been assumed to have a linear wind variation between turbines across the farm of 5 ms^{-1} giving the conversion curve for the total wind farm in Figure 5. We assumed that all fifteen sites have the same maximum power generation. The results have been normalised to show the fraction of maximum possible power.

Figure 6 shows the cumulative distribution of power from the original wind data and ten simulations that indicate the extent of variation that can be expected, for example at the median power (50% on probability scale) the power output is estimated as being between about 46% to 52% of the maximum. Figure 7 expands the ends of the probability scale, and adds curves for a single wind farm and four wind farms. As more wind farms are added the distribution becomes flatter, corresponding to less extreme relative variation. However, as the power from multiple wind farms is greater, the actual variation in power is greater.

The power network operators are also interested in how fast changes in power generation can occur. Figure 8 shows the distribution of changes in power generation between one ten minute interval and the next, again for the fifteen sites with equal maximum generating capacity. The vertical scale is change in power as a fraction of total generating capacity. The graph suggests that around 95% of the time the relative power change is within $\pm 5\%$, but occasionally changes of between 10% and 15% will occur, even though the time between measurements is only ten minutes. Figure 9 expands the ends of the probability scale, and adds the relative changes for one and four wind farms. Again the relative changes decrease with more wind farms, but actual changes

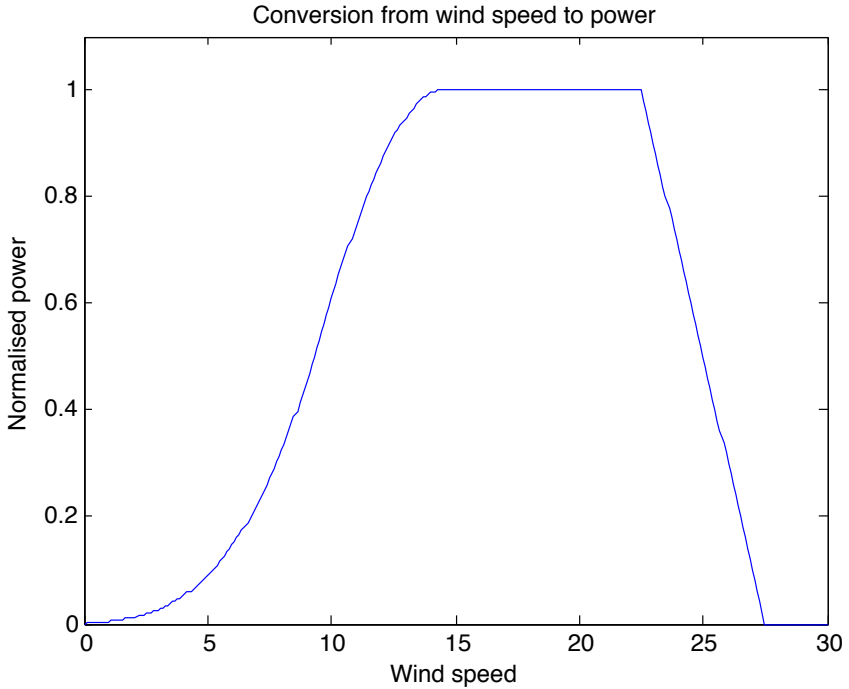


Figure 5: Relation between wind speed ms^{-1} and total power generated for the wind farm. Power scale is normalised by the maximum power possible.

will increase. The Monte Carlo method appears to slightly overestimate the magnitude of relative power change in the mid-tails of the distribution ($0.001 < \text{probability} < 0.1$ and $0.9 < \text{probability} < 0.999$). The reasons for this have not been investigated.

4 Prediction using simulation

The simulation was used to produce predictions over a short time interval given initial values of wind speed over the previous four, 10 minute intervals.

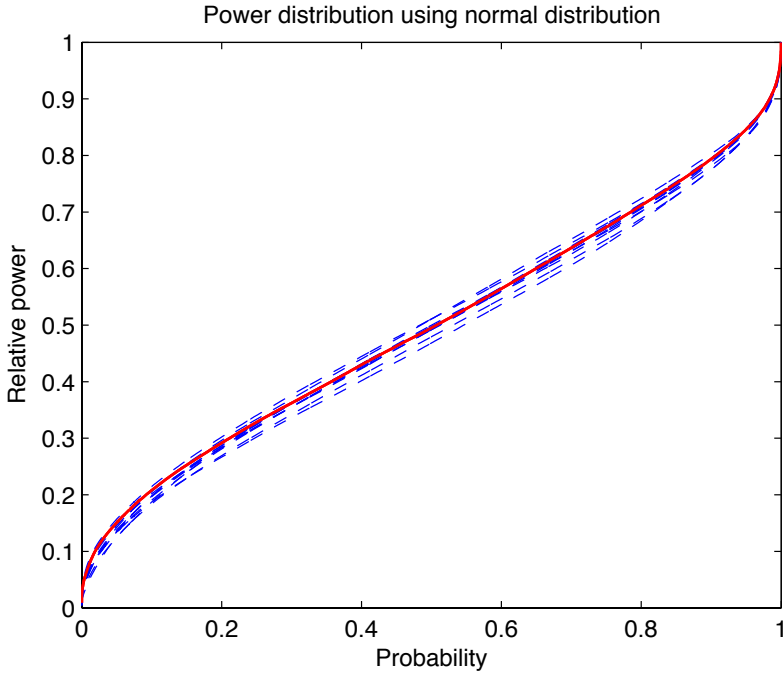


Figure 6: Cumulative distribution of relative power output; sum of fifteen equal sites with maximum output normalised to one. Solid red line is from the wind data and the blue dashed lines are ten Monte Carlo repeats.

Figure 10 shows ten simulations for each of the fifteen sites over the first 120 minutes of prediction. As might be expected the variation predicted by the simulations progressively increases as the time increases and in most cases showed a significant amount of variation after 120 minutes.

Figure 11 shows the wind speeds from Figure 10 converted to power generation expressed relative to the maximum power output. In this example many of the sites are operating at wind speeds that correspond to step parts of the speed to power graph of Figure 5. The result is a large amount of scatter in the power, and the prediction rapidly becomes useless.

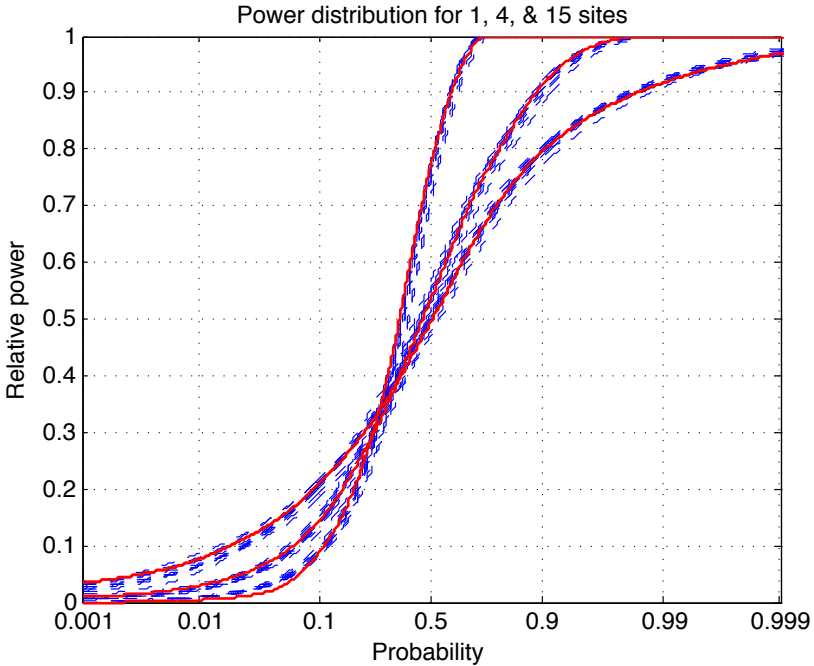


Figure 7: Cumulative distribution of power output; for one, four, and fifteen equal sites with maximum output normalised to one and an expanded probability scale. Lower slopes correspond to more wind farm sites. Solid red lines from the wind data and the blue dashed lines are ten Monte Carlo repeats.

5 Prediction by data-matching

The data available is plentiful so it could be expected that weather conditions in the file could be matched with the current wind values, and the previous behaviour used to predict the future behaviour from the current conditions. The current wind values used for testing are the same as the initial conditions used in the previous section and four years of data from the file of ten minute values was used as the historical record.

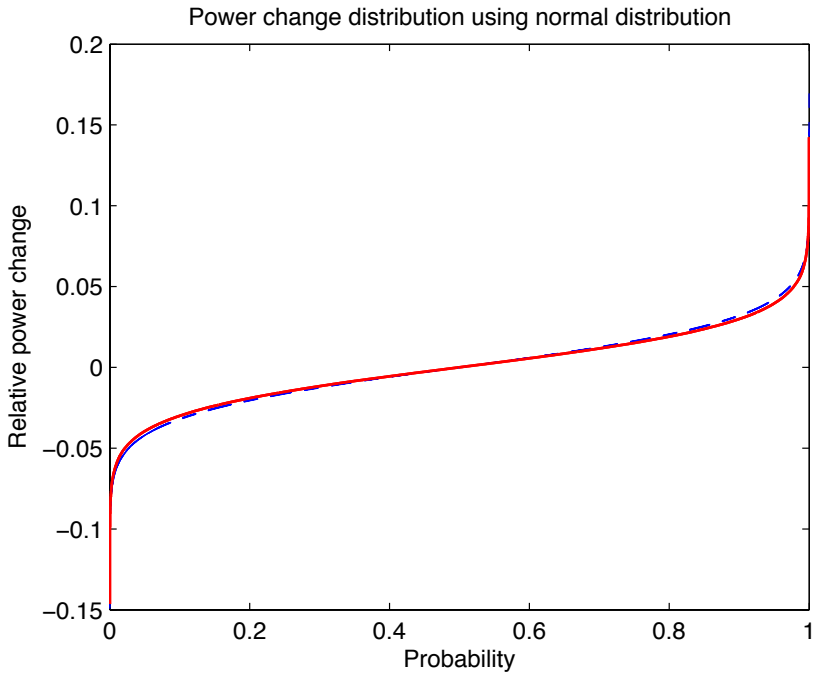


Figure 8: Cumulative distribution of changes in power output relative to the maximum power, in ten minutes; fifteen equal sites with maximum output normalised to one. Solid red line is from the wind data and the blue dashed lines are ten Monte Carlo repeats.

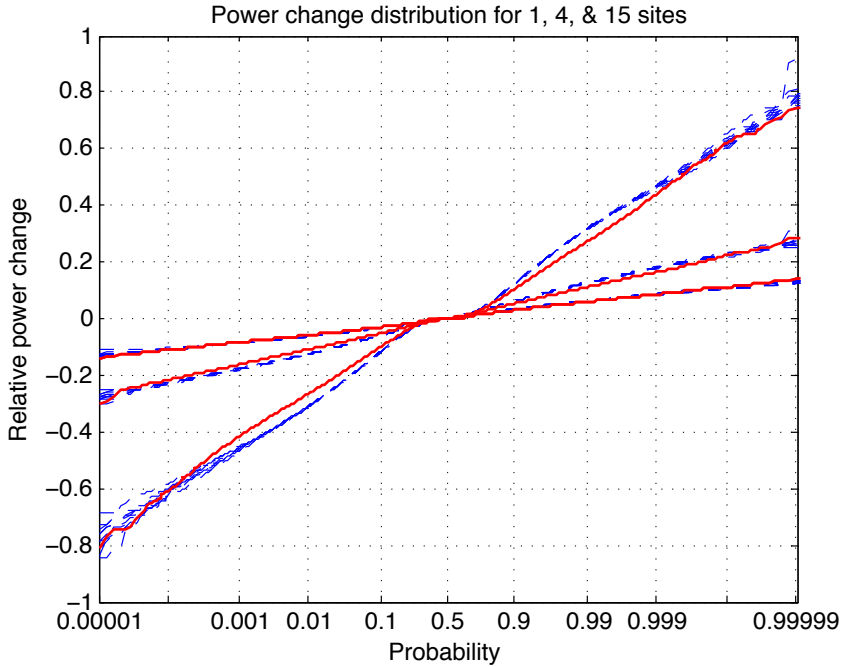


Figure 9: Cumulative distribution of changes in power output relative to the maximum power, in ten minutes; one, four, and fifteen equal sites with maximum output normalised to one and an expanded probability scale. More wind farms sites give a lower slope. Solid red line is from the wind data and the blue dashed lines are ten Monte Carlo repeats.

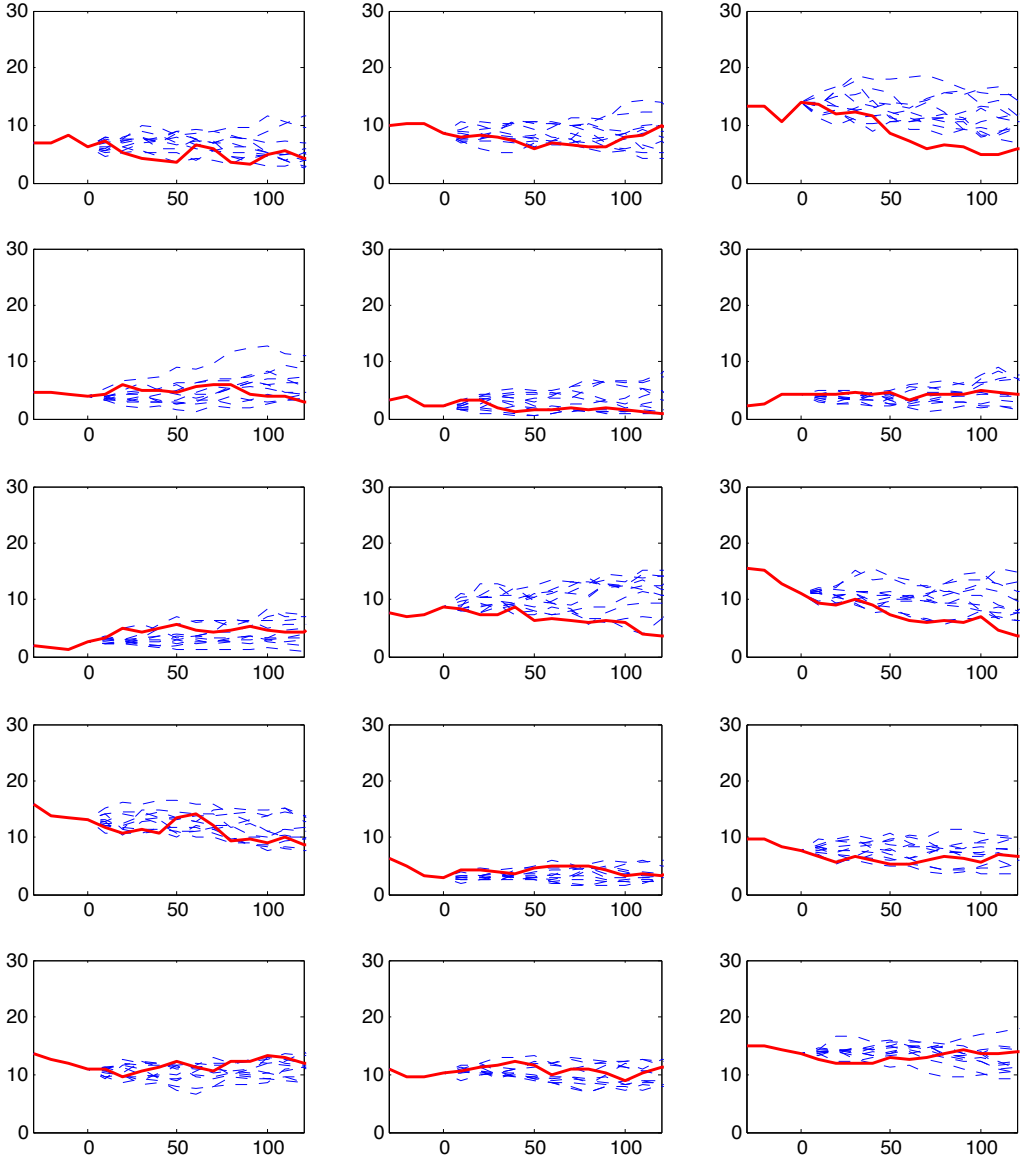


Figure 10: Simulation of wind (vertical scale ms^{-1}) over 120 minutes (horizontal scale) from a given initial position. Dotted blue lines are simulations, and solid red is actual data.

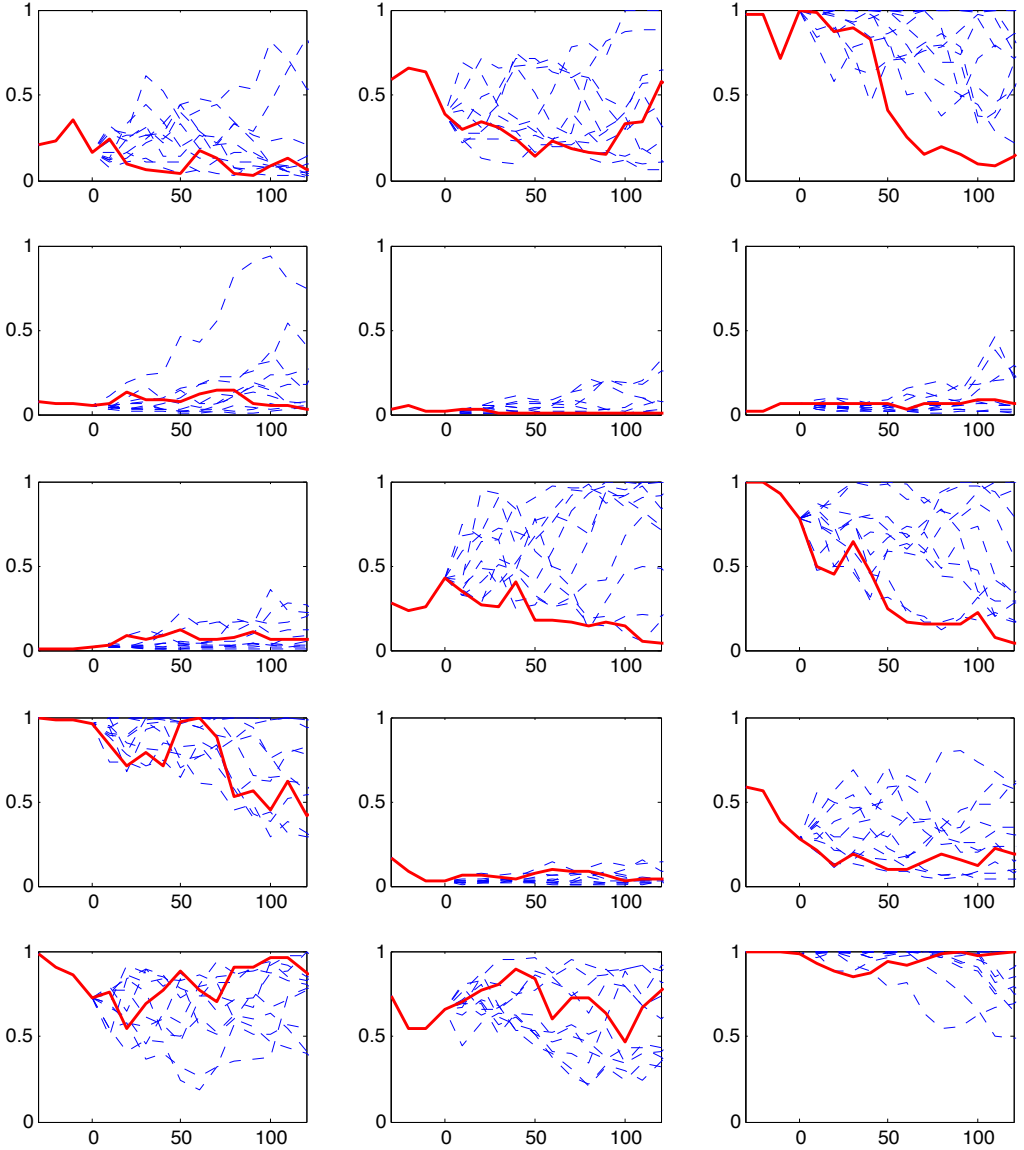


Figure 11: Simulation of power relative to maximum, over 120 minutes (horizontal scale) from a given initial position. Dotted blue lines are simulations, and solid red is actual data.

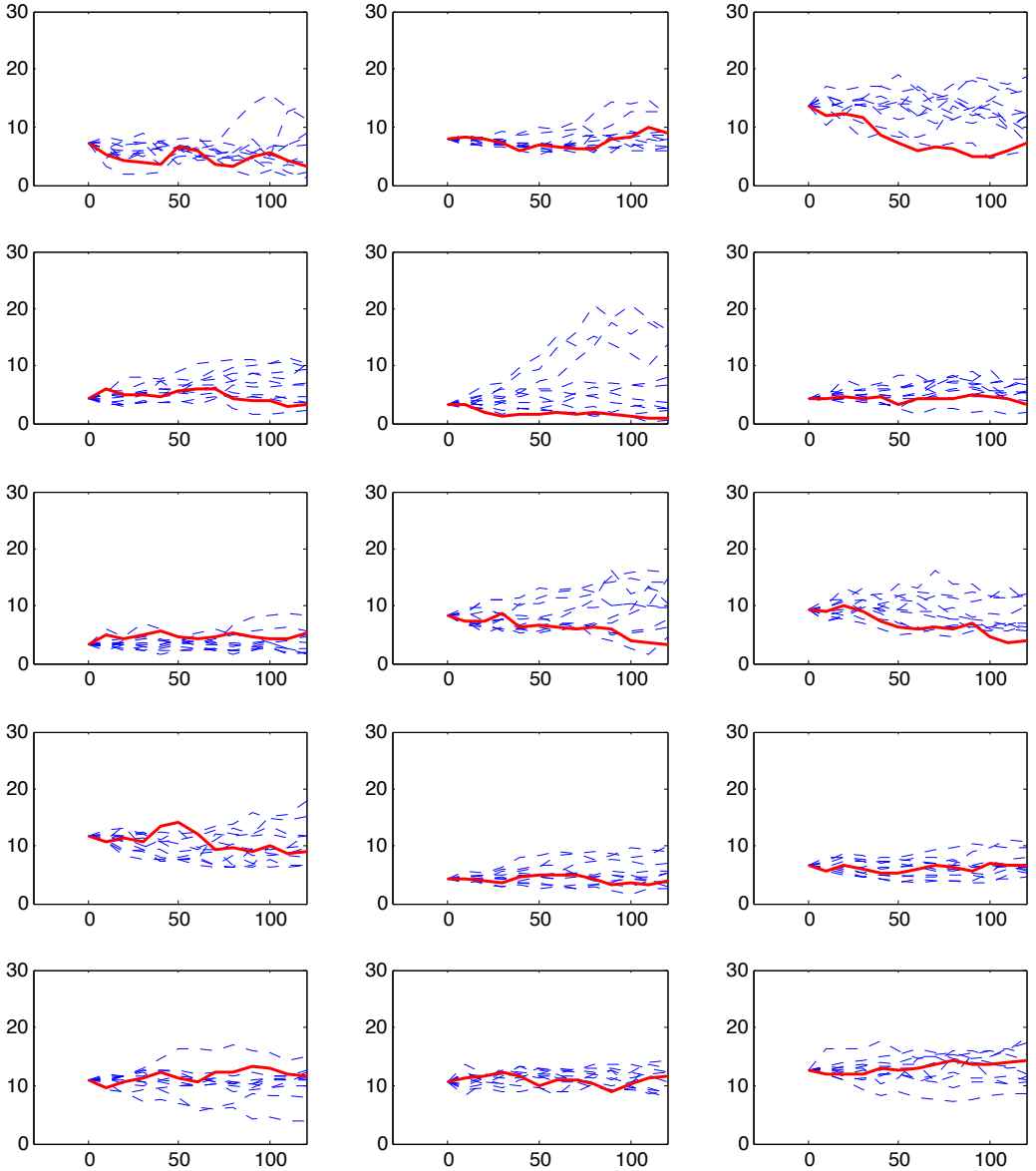


Figure 12: Prediction of wind (vertical scale ms^{-1}), using similar cases from data, over 120 minutes from a given initial position. Dotted blue lines are simulations, and solid red is actual data.

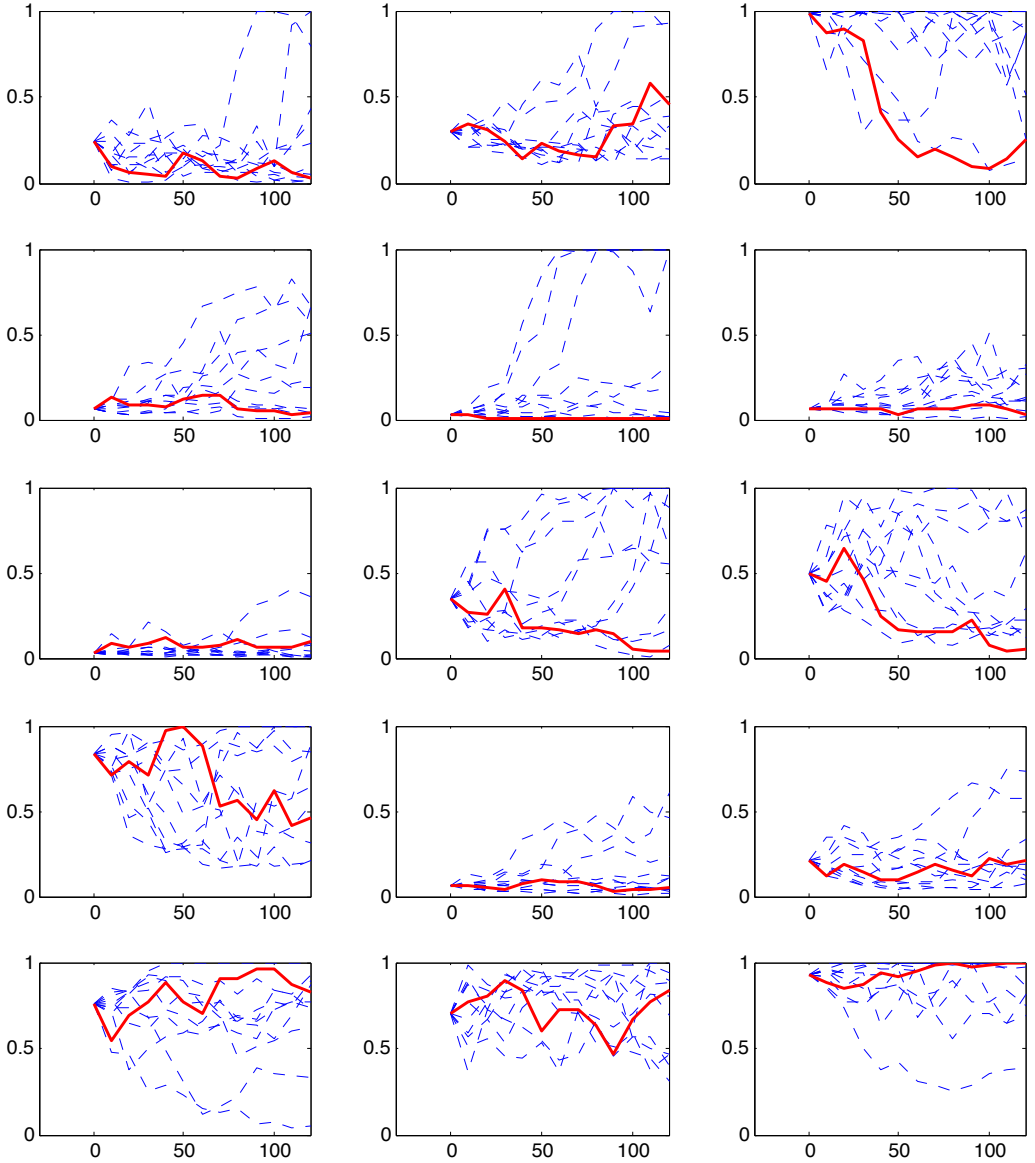


Figure 13: Prediction of power relative to maximum power, using similar cases from data, over 120 minutes from a given initial position. Dotted blue lines are simulations, and solid red is actual data.

The root mean square error was calculated between the current ten minute values for the fifteen sites, and the ten minute values for the fifteen sites in the four years of data. The ten smallest root mean square values (1.80 to 2.15) were located and the wind for the next two hours following the match point scaled so that the first values matched the current wind values.

Figure 12 shows the predictions obtained for the wind and Figure 13 shows the relative power predictions. Comparing these with Figures 10 and 11 it is seen that the variation is perhaps slightly smaller than for the simulated data, but overall the results are basically similar and the power prediction rapidly becomes useless.

6 Prediction using a local model

It was also investigated whether the simulation model could produce better predictions if the parameter estimation was based on data similar to the case to be predicted. Here, sections of the data starting with wind conditions similar to the current wind values were selected and the 120 minutes following these times were used to develop a model of the same form as used above.

To get a stable model it was necessary to use the covariance of the residuals. Ten cases of possibly overlapping sections of the data were used. Figure 14 shows the wind predictions and Figure 15 shows the relative power predictions. Again the predictions for power rapidly spread out to cover essentially the whole possible range.

7 Prediction using power

While simulating the wind and generating the power follows the physical mechanisms, the wind speed and the previous values of the wind speed may

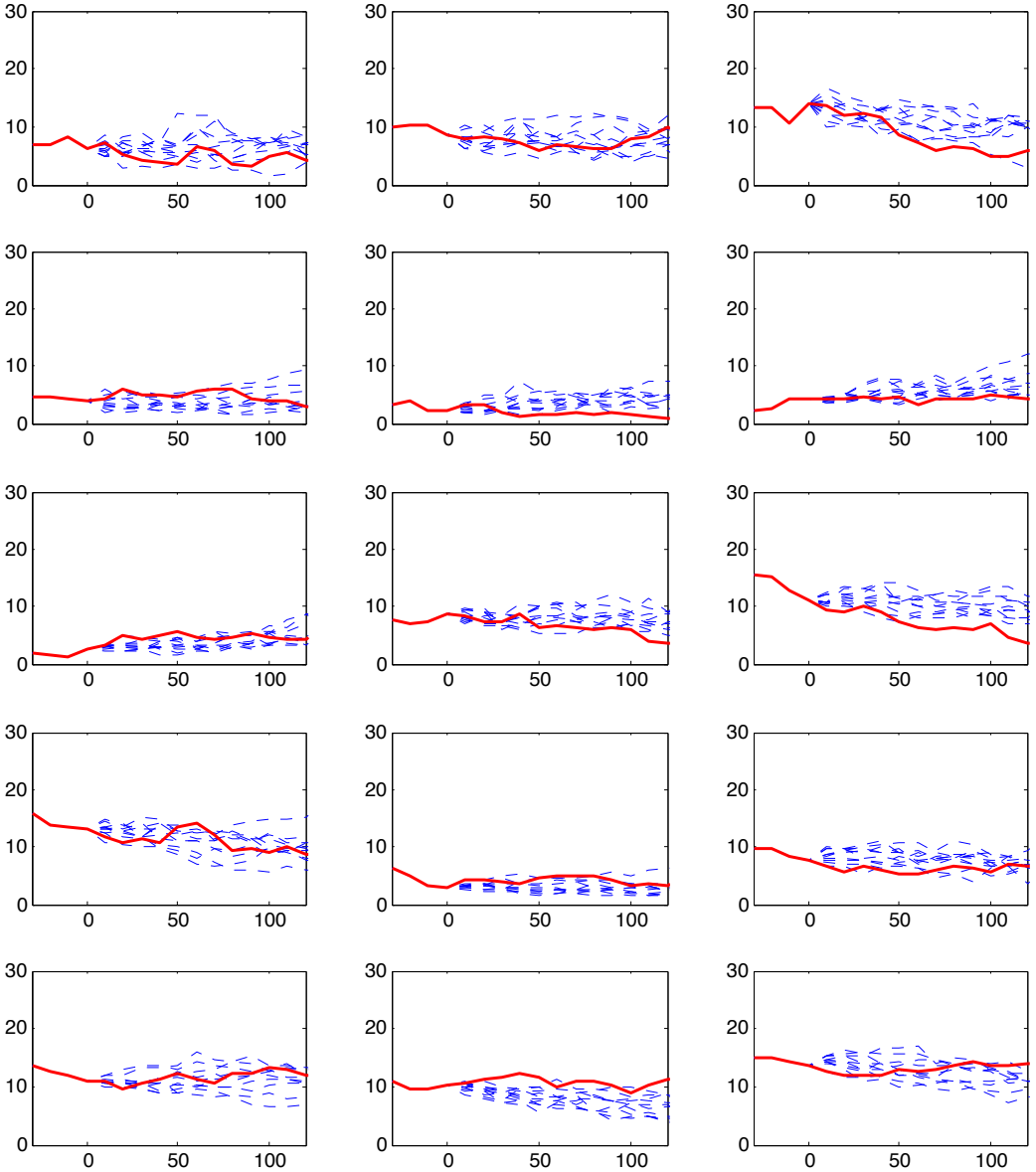


Figure 14: Simulation of wind (vertical scale ms^{-1}), using a model based on data similar to the initial wind, over 120 minutes from a given initial position. Dotted blue lines are simulations, and solid red is actual data.

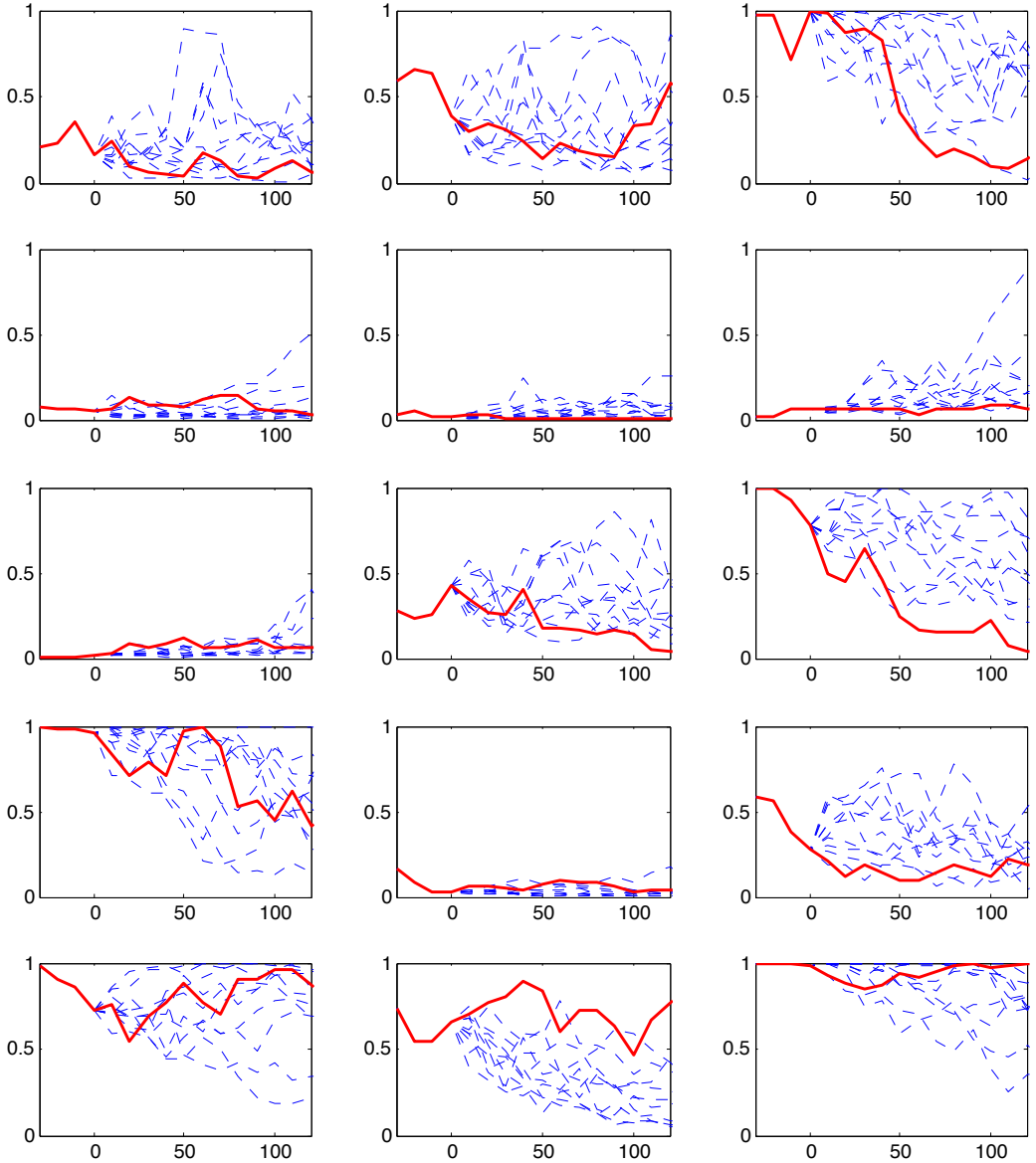


Figure 15: Simulation of power relative to maximum power, using a model based on data similar to the initial wind, over 120 minutes from a given initial position. Dotted blue lines are simulations, and solid red is actual data.

not be generally available. However, the power from the wind farm is available. Prediction using power has some major disadvantages:

1. The wind to power curve is not monotonic;
2. There is a flat section of the power curve at intermediate wind speeds, that does not allow a unique conversion from power to wind speed;
3. At low winds speeds the power generation is zero not allowing estimation of low wind speeds;
4. It may not be known how many of the windmills are actually operational, due to scheduled or unscheduled down time.

We believe that it would be possible to estimate whether a reduction in power is due to low or high wind conditions. Assuming this is possible, the wind speeds were converted to power and back to an estimated wind speed. This set speeds between 14.5 and 22.5 to 18.5, and those over 27.5 to 29. These modified wind speeds were transformed to the Gaussian values and used to find the regression coefficients and the correlation matrix, so that predictions could be made. Figure 16 shows the result of predicting the distribution of relative power outputs using these values. The fit to the distribution from the data (solid red line) is not as good as obtained using the wind speeds (Figure 6).

The conversion from power output to wind speed used for Figure 16 is exact except where wind to power relation is flat. In practice this relation will not be exact and so predictions made from power data will be worse than seen here. So the use of power to predict future power outputs is only recommended if no other method is feasible.

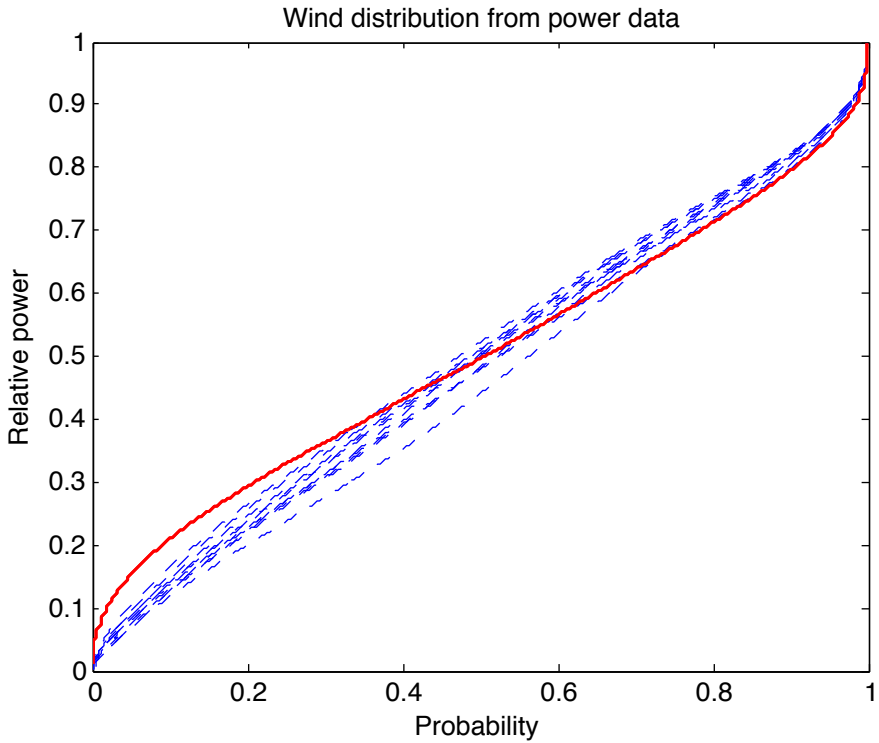


Figure 16: Ten repeats of prediction of power relative to maximum power output, using wind speed estimated using power outputs (blue dashes), and power output distribution from wind data (red curve).

8 Additional effects

During MISG a prospective approach was examined, to see whether one could at least predict whether winds near a critical value on the speed-power curve (15 or 25 ms^{-1}) were likely to go on to exceed that value. Episodes of wind were examined to see whether the wind behaviour could be used to predict whether the wind would exceed the crucial number or stay below. In the end no short-term pattern of speeds could be found to distinguish between episodes that went on to exceed the crucial number and those that did not.

The simulation model relies heavily on a regression of the current wind on that ten minutes earlier. Simulation based on winds further back in time (for example 30, 60, 120 or 300 minutes prior) is much less accurate. In that case it may be useful to include into the regression model terms involving the time of day, month of year and wind direction. During and subsequent to MISG considerable effort was put into investigating whether such terms would be useful.

Figures 17 and 18 illustrate the correlations between wind speed at the Manawatu 1 wind farm (MWT1) and time-lagged winds at the nearby Manuwatu 2 (MWT2) farm or Cook Strait 2 farm (CKS2). The numbers $i = 1 : 61$ refer to the time lag $10(i - 1)$ minutes ago. The black lines give the correlations ignoring wind direction (direction was not provided for the synthetic wind farm data). The coloured dashed lines refer to correlations where the data has been broken up into different subsets depending on the real (as opposed to synthetic) contemporary wind direction at the nearby Palmerston North Airport. For example East denotes north-easterly through to south-easterly winds. Note the direction data was incomplete, and when the Palmerston North wind speed was zero then no direction was available. For low speeds the direction was also highly variable, and one could also argue that wind direction at a lowland airport can not be expected to tell us much about winds 80 metres above a range of hills some kilometres away: but it was the best wind direction data available. The graphs show that MWT1 winds were

much more highly correlated with easterly winds (red dashed line) from the MWT2 farm or CKS2 than with winds from other directions. Disappointingly, ignoring the wind direction (black line) gave correlations almost as high as for easterly winds: this is probably due to the data limitations stated above. For MWT2, the shorter the lag, the stronger the correlation was with the nearby MWT1 winds, indicating the value of having recent data from nearby sites. By contrast Figure 18 for CKS2 (some 150 km south) the relationship between the correlation to MWT1 and the lag was much flatter, giving about the same correlation 300 minutes (five hours) ago as for 10 minutes ago. Interestingly these correlations were still stronger for easterly winds than for southerly (i.e., from the direction from Cook Strait) which is perhaps related to topography and perhaps to the relatively large size of weather systems, leading to west and east prevailing winds. This limited illustration does suggest that forecasts may be improved by using meteorological data, or having anemometers at the sites and at some distance east or west of the farms to give indication of wind direction and speed.

Finally, the effects of time of day, and month of year, were considered. These will be more important for predicting more than one lag in advance. We illustrate by considering three models:

$$\text{MWT1}_i = \beta_0 + \beta_1 (\text{MWT1}_{i-1}); \quad (11)$$

$$\text{MWT1}_i = \beta_0 + \beta_1 (\text{MWT1}_{i-1}) + \beta_2 (\text{MWT1}_{i-2}); \quad (12)$$

$$\begin{aligned} \text{MWT1}_i = & \beta_0 + \beta_1 (\text{MWT1}_{i-1}) + \beta_2 (\text{MWT1}_{i-2}) + \gamma(\mathbf{m}) \\ & + \sum_{k=1}^2 \left[A_k \sin \left(\frac{2k\pi i}{144} \right) + B_k \cos \left(\frac{2k\pi i}{144} \right) \right] \\ & + \sum_{k=1}^2 \left[\mathbf{a}(\mathbf{m})_k \sin \left(\frac{2k\pi i}{144} \right) + \mathbf{b}(\mathbf{m})_k \cos \left(\frac{2k\pi i}{144} \right) \right]. \quad (13) \end{aligned}$$

Equations (11) and (12) are regressions of wind speed on the l -lagged values for the Manuwatu 1 and 2 wind farms. In addition (13) includes an overall effect $\gamma(\mathbf{m})$ for the month, sine and cosine terms on the second line represent

Table 4: R^2 for Regression Models (11)–(13).

Model	lag 1 10 min.	lag 3 30 min.	lag 6 1 hour	lag 12 2 hour	lag 30 5 hour
(11)	0.9659	0.9150	0.8565	0.7638	0.5388
(12)	0.9663	0.9176	0.8632	0.7753	0.5549
(13)	0.9664	0.9182	0.8650	0.7809	0.5769

some (possibly asymmetrical) diurnal effect as the land heats during the day and cools during the night, and the third line coefficients $\mathbf{a}(\mathbf{m})_k$ and $\mathbf{b}(\mathbf{m})_k$ allow for a modification of the diurnal effect by month. Table 4 summarizes the results of fitting these three models to a two-year subset of the synthetic data (model fitted to the power law wind^{0.75} to improve symmetry of residuals). For the first ten minute lag, the R^2 for model fit is scarcely increased from model (11) to model (12) but in fact the regression coefficient is significant with P-value $< 10^{-15}$. Again the R^2 scarcely increased for model (13) but the first diurnal component of model (13) was significant (P-value = 5×10^{-13}) and the monthly effect almost significant (P-value = 0.0602): however, the month–diurnal interaction was not significant (P-value = 0.9532). For all higher lags, the month–diurnal interaction in model (13) was highly significant, and the additional terms in models (12) and (13) start to have an appreciable effect on the R^2 . These results illustrate that some gains can be made by allowing for the effect of time of day, and month of year, in regression models. However it will require considerable modelling effort and more detailed data, in order to have an appreciable effect on the accuracy of predicting the wind, let alone the power.

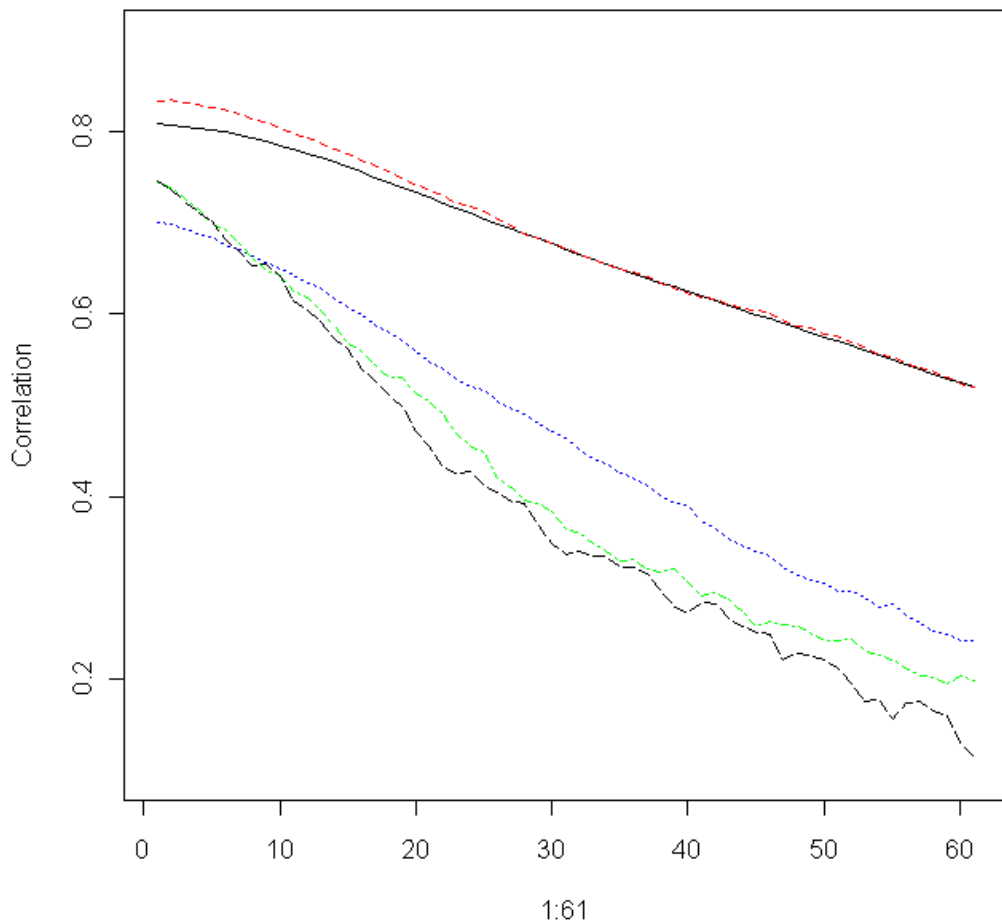


Figure 17: Correlations of MWT1 winds with time-lagged MWT2 winds, versus time lag, namely $10(i - 1)$ minutes for $i = 1 : 61$. Split is by wind direction at Palmerston North.

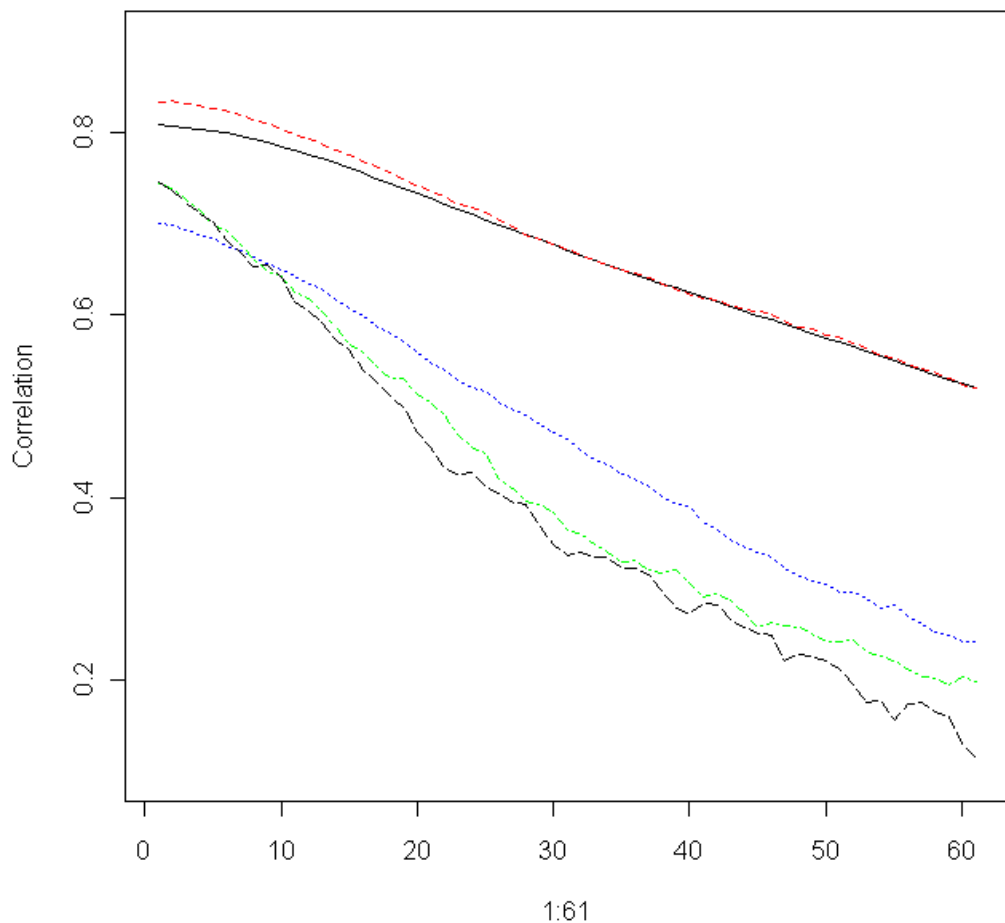


Figure 18: Correlations of MWT1 winds with time-lagged CKS2 winds, versus time lag, namely $10(i - 1)$ minutes for $i = 1 : 61$. Split is by wind direction at Palmerston North.

9 Conclusions

This paper illustrates the modelling of correlated wind speeds and hence the power output for fifteen existing or potential New Zealand wind farms. The method demonstrated was able to predict the distribution of power generation and the distribution of changes in ten minute power generation, to a usable level of accuracy. The simulations also provided estimates of the accuracy of these distributions.

The short term predictions of wind from given initial wind conditions showed a progressive loss of accuracy over two hours. However due to the steep relations in the conversion from wind speed to power, predictions of short term power are disappointing.

Predictions made by locating similar wind conditions over the fifteen sites gave largely similar results, as did fitting a local model to the similar wind conditions. Again the conversion from wind speed to power greatly increases the variation in the power predictions.

Predicting using power values, has several potential problems. An attempt using the power values to estimate wind speed for a simulation produced results worse than using wind directly. Using power without conversion to wind speed may be possible but as it would lose information on high versus low wind speeds a useful improvement in prediction accuracy is not expected.

There are multiple possibilities for improvement to the predictions, in particular the lag between the different sites has been set at a single value whereas a value dependent on distance seems appropriate. Tests on the use of a local model developed using hopefully similar wind conditions did not produce the desired improvement; however, further investigation along these lines seems promising.

The most promising approach from those considered seems to be the use of a local model developed from similar wind conditions. Additional data, such as additional near sites and wind direction, to assist in getting good matches to

the current conditions should improve the predictions. Experience with the current algorithms indicates this could easily be done sufficiently quickly to make real time predictions.

An open question is how much information from weather forecasts would assist in prediction. Short-term predictions should be improved by information on wind speed and direction from nearby sites. The data used in this study was produced using numerical weather forecasts matched to known conditions. The thesis by [Cutler \(2009\)](#) takes a similar approach using numerical weather predictions around the wind farm site.

Acknowledgements The industry contact Conrad Edwards from Transpower (NZ) Ltd, Conrad.Edwards@transpower.co.nz took time out a busy schedule to prepare data, present and explain the problem to us. The team that worked on the problem included Allan Brown, Andreas Ernst, Lynne McArthur, and Simon Watt. Allan Brown continued to work on the project long after the end of the MISG. The work of John Shepherd and his team in organising MISG 2011 is greatly appreciated. A referee provided useful input that improved the paper.

A Notation

α_j	coefficient of the previous value $y_{i,j}$ in the regression for $y_{i+1,j}$
$b_{j,k}$	coefficients of $(y_{i-2,k} + y_{i-3,k} + y_{i-4,k})$ in the regression for $y_{i+1,j}$
$\text{cov}(x_{i,j})$	covariance function of the matrix $x_{i,j}$
$\epsilon_{i,j}$	the regression error term on the Gaussian scale for the i th data row and j th site
$e_{i,j}$	the residual from the regression for the i th data row and j th site
$\epsilon_{i,j}^*$	the simulated residual on the Gaussian scale for the i th data row and j th site
$\Gamma(x, \alpha, \beta)$	the gamma cumulative distribution function, shape α and scale β
$\Gamma^{-1}(u, \alpha, \beta)$	the inverse gamma cumulative distribution function, shape α and scale β
$\Phi(x)$	the Gaussian cumulative distribution function with mean zero and unit variance
$\Phi^{-1}(u)$	the inverse of the Gaussian cumulative distribution function
u_{ij}	the uniform distributed value calculated from x_{ij}
u_{ij}^*	the simulated uniform distributed value calculated from y_{ij}^*
x_{ij}	the wind speed data at time i (10 minute intervals), and site j
x_{ij}^*	the simulated wind speed at time i (10 minute intervals), and site j , calculated from $u_{i,j}^*$
y_{ij}	the Gaussian distributed value calculated from u_{ij}
y_{ij}^*	the simulated Gaussian distributed value used to calculate to u_{ij}^*
$\gamma(m)$	parameter representing the main effect of month on wind
A_k, B_k	coefficients for sinusoidal terms representing diurnal effects on wind
$a(m)_k, b(m)_k$	coefficients representing how diurnal sinusoidal effects are modified according to month

References

- Electricity Authority, 2011. <http://www.ea.govt.nz/industry/modelling/analysis-of-wind-integration/synthetic-wind-data/> M106, M110
- Cutler, N., 2009. Characterising the Uncertainty in Potential Large Rapid Changes in Wind Power Generation, PhD thesis, The University of New South Wales. http://www.ceem.unsw.edu.au/windfornem/nicholas_cutler.cfm M138
- Stacy, E.W., 1962. A generalisation of the gamma distribution, *Annals of Mathematical Statistics* 33(3):1187–1192. <http://www.jstor.org/stable/2237889> M107
- Whiten, B. and Tsoularis, T., 2004. Prediction of power generation from a wind farm, MISG2004. <http://www.maths-in-industry.org/miis/39/> M107
- Wikipedia, 2011. http://en.wikipedia.org/wiki/Generalized_gamma_distribution M107
- Zakaria, R., Metcalfe, A.V., Howlett, P., Piantadosi, J. and Boland, J., 2010. Using the skew-t copula to model bivariate rainfall distribution, ANZIAM J. (EMAC2009) pp. C231–C246. <http://journal.austms.org.au/ojs/index.php/ANZIAMJ/article/view/2030> M107

Author addresses

1. **Bill Whiten**, 4 Magnet Cl, Brisbane 4074
<mailto:W.Whiten@uq.edu.au>
2. **Barry McDonald**, Massey University, INMS, Auckland 0632
<mailto:B.McDonald@massey.ac.nz>

3. **Ray Bedford**, RMIT, Mathematical Sciences, Melbourne 3000
<mailto:ray.bedford@rmit.edu.au>