# Statistical Analysis of Environment Canada's Wind Speed Data

Someshwar Singh
Department of Electrical and
Computer Engineering
University of New Brunswick-Fredericton
New Brunswick, Canada
Email: someshwar.singh@unb.ca

James H. Taylor
Department of Electrical and
Computer Engineering
University of New Brunswick-Fredericton
New Brunswick, Canada
Email: jtaylor@unb.ca

***Abstract*—Wind energy utilities use wind speed modeling and prediction to forecast their power production in order to participate in electricity markets. Time-series models which are indirectly based on a Weibull Distribution (WD) are used extensively to predict wind speed. The WD is converted into an approximately Gaussian distribution, as there are no rigorously developed time-series models for random variables possessing a WD. This conversion is performed using the parameters of the WD, a procedure that may negatively impact the accuracy of the forecast – research has demonstrated that WDs under- or over-fit the lower and upper ranges of wind speed histograms.**

**This paper reports on a study of the histories of wind speed forecasts and actual wind speed data available from Environment Canada and the resulting estimates of forecast error distributions and statistics. It is shown through statistical analysis that the hourly prediction error distributions are nearly Gaussian in nature.**

**It also appears to show that the statistics of the wind-speed prediction error do not increase significantly as time increases, which is in contrast to other researchers' arguments that the error increases over time. This result may warrant further investigation.**

## I. INTRODUCTION

The intermittent nature of wind power generation poses operational difficulties to electricity markets. An electricity market operated by an Independent System Operator (ISO) must always maintain a balance between supply and demand of electricity at each instant of time. If there is any variation in load, there must be reserves at the ISO's disposal. To maintain stable operation of the grid, the ISO accepts hourly bids starting at 9:00 am and ending at 11:00 am Atlantic Standard Time (AST), for the following day (Delivery Day, 00:00 to 23:59:59), from buyers and suppliers [1]. The system operator then runs an optimization algorithm to calculate the price at which maximum demand has been fulfilled at minimum cost.

The participants have to fulfill their obligation at the time of delivery. After the delivery day, deviations from the hourly accepted bid quantities are calculated for each market participant and financial penalties will be charged to the defaulters. The wind energy (WE) utility thus faces the challenge of producing accurate power generation forecasts before entering into the electricity market, as power forecast errors could have a significant impact on the WE utility's revenue.

Wind power prediction requires wind speed forecasting because the kinetic energy in the wind is converted into electric power by the wind power generator. Stationary time-series models are used extensively for modeling and forecasting wind speed [2], [3], [4]. The wind speeds are recorded at the site, then their distribution is plotted; the statistical distribution of the series does not change over time. It has been assumed that the wind speeds follow a Weibull Distribution (WD) [5]. Since there is no rigorously developed time-series models for random variables possessing a WD, the data is transformed into an approximately Gaussian Distribution (GD) [2], [3], [4].

García-Bustamante *et al.* [5] and Jamil *et al.* [6] have shown that the WD assumption for recorded wind-speed data is not appropriate – the WD under- or over-fits wind speed histograms, especially in lower- and upper-range wind speed intervals. The transformation from a WD to an approximate GD is carried out by raising each hourly wind speed to the power of $m$; the value of $m$ is calculated using shape and scale parameters of a WD. Since a WD fit is not quite appropriate for recorded wind-speed data, the transformation from a WD to an approximate GD may not be a realistic characterization of the recorded wind-speed data.

Holttinen [7] has shown that prediction error increases as time increases; we have not observed this in our data set, however. In such cases, the WE utility revenue could increase by 7% if wind power is traded 2 hours before actual delivery [7] rather than 16 hours, as is presently the case in New Brunswick.

## II. PREDICTION ERROR CALCULATION

Environment Canada's Fredericton station provides weather forecasts every day at 08:00 for the next 48 hours in 3 hour blocks in Gridded Binary (GRIB) format. The forecasts are available at multiple resolutions for over 817 Canadian stations or sample points [8]. GRIB is a concise data format commonly used by the meteorological institutes of the world to store and share historical and forecasted weather data.

The prediction errors were calculated by comparing the forecasted wind speeds in the GRIB files with the actual wind speeds for the same prediction time; 291 GRIB files and the actual wind speeds for the year 2003 provided by Environment Canada (EC) Fredericton were used. It has been demonstrated that wind speed forecasts are more accurate if the forecasting techniques incorporate local weather conditions and knowledge of prediction errors [7], [9], [10], [11], so this is very adventageous. The actual data is in hourly blocks, while forecast data is in three-hour blocks; therefore the missing two hour data points of each block in the GRIB files were filled using the persistence technique. The persistence technique assumes that speed will be the same at $t + k$ hours as at $t$ hours, $k = 1, 2$ [7].

As mentioned above, Holttinen [7] has shown that error in the wind power prediction increases as the forecast grows older; therefore power prediction error can be reduced significantly if the time between bid close and delivery is short. Since the ISO accepts bids for the following day (Delivery Day, 00:00 to 23:59:59), the wind-speed forecasts provided by Environment Canada are 16 to 40 hour old over the Delivery Day (DD), as the forecast is available at 08:00 for the next 48 hours.

One can then intuitively think that the mean and the variance of wind-speed forecast error should increase as prediction time increases. To validate this, the wind-speed forecast errors were calculated by comparing forecasted wind speeds ($v_f$) with actual wind speeds ($v_a$) for a prediction horizon of 48 hours, according to the scheme shown in Fig. 1, where **w** is the number of hours covered in a single prediction error distribution, e.g., six hours as shown. The Forecasts $1, \ldots, N$ were compared with the actual data $1, \ldots, N$, where $N = 291$. For example, given **w** = 6 hours and a prediction horizon of 48 hours, then there are 8 wind-speed forecast error distributions of 6 hours; the first distribution covers the wind-speed forecast error data from 08:00 to 13:59:59 AST, the second distribution covers 14:00 to 19.59:59 AST, and so on. It should be noted that there is a time overlap between the forecasted and the actual data (Fig. 1) but data corresponding to them belongs to different categories. The time 08:00 at day 2 in Forecast 1 is 24 hours old while 08:00 at day 2 in Forecast 2 is 00:00 hours old.

### A. Mean and Variance for Wind-Speed Forecast Error

Let $v_f(i,j)$ represent the value of the wind speed in Forecast $i$ at the $j^{th}$ hour, and let $v_a(i,j)$ represent the corresponding actual wind speed. Also, let $\hat{\mu}^{(r)}$ represent the sample mean and $\hat{S}^{(r)}$ represent the sample variance of the wind-speed forecast error of the $r^{th}$ distribution, where $r = 1, 2, \ldots, (P_h/\mathbf{w})$; $P_h$ is the prediction horizon. For example, given **w** = 6 hours and a prediction horizon of 48, then there are 8 wind-speed forecast error distributions of 6 hours, thus there are 8 sample means and sample variances, $r = 1, 2, \ldots, 8$. The statistics $\hat{\mu}^{(r)}$ and $\hat{S}^{(r)}$ are calculated using equation (1) and (2) respectively:

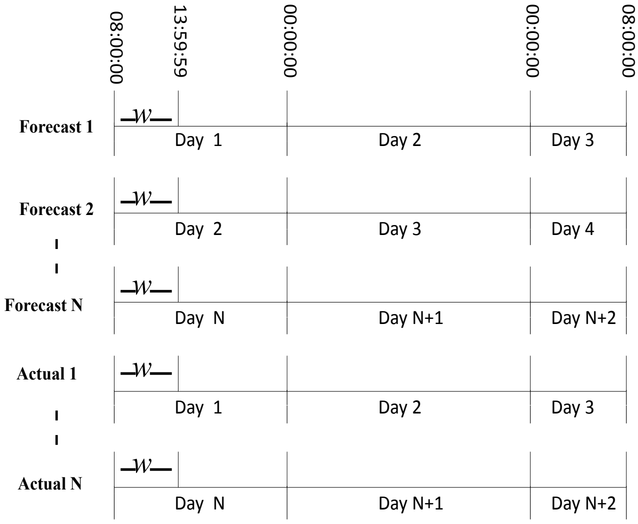


Fig. 1. Scheme for data collection

$$\hat{\mu}^{(r)} = \frac{1}{N\mathbf{w}} \sum_{j=8+(r-1)\mathbf{w}}^{8+(r\mathbf{w}-1)} \sum_{i=1}^{N} (v_f(i,j) - v_a(i,j)) \qquad (1)$$

$$\hat{S}^{(r)} = \frac{1}{N\mathbf{w}-1} \sum_{j=8+(r-1)\mathbf{w}}^{8+(r\mathbf{w}-1)} \sum_{i=1}^{N} (v_f(i,j) - v_a(i,j) - \hat{\mu}^{(r)})^2 \qquad (2)$$

The mean and variance for a prediction time block **w** = 1 are shown in Fig. 2.

It can be observed that the mean and variance of the prediction error for this data do not increase significantly as prediction time increases; in fact they follow 24 hour cycles. Therefore, although the data used for modeling distribution will be 16 to 40 hours old at the time of its usage, it would appear to have little affect on the prediction results as compared to when it was 00:00 hours old. The main reason for this counterintuitive result could be the limitation of the model discussed in [7] for forecasting wind speeds, or the size of our data set.

The probability distribution of wind-speed forecast error for various prediction time blocks were then analyzed so that an appropriate method could be chosen to model the probability distributions. The probability distributions for 24h, 12h, 6h, 3h,

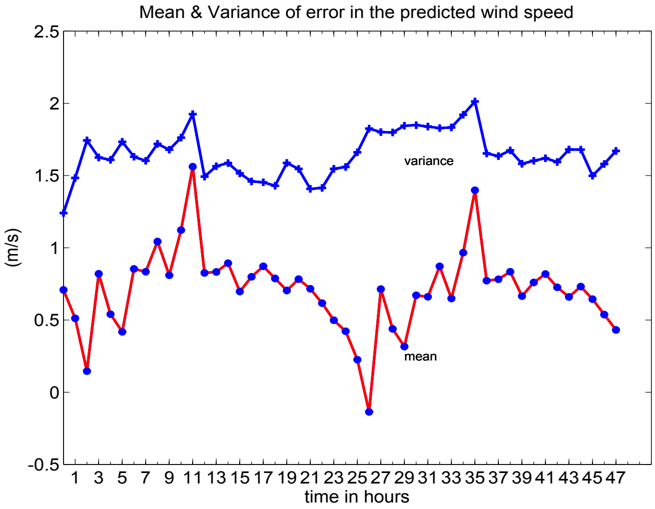


Fig. 2. Mean and variance of error in forecasted wind speed

and 1h prediction time blocks in a 24 hour prediction horizon (00:00 to 23:59:59) were then tested for normality using two methods: a statistical Chi-Square ($\chi^2$) test and by comparing the empirical cumulative distribution with that of a Gaussian process having the same mean and variance.

### B. Chi-Square Test

The $\chi^2$ method tests a null hypothesis $H_0$ that a sample data set follows a specified distribution, i.e., that there is no significant difference between their distributions. It divides the samples into $k$ bins and then calculates a test statistic $\chi^2$ given as,

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{3}$$

where $O_i$ = observed frequency and $E_i$ = expected frequency. The expected frequency is given as,

$$E_i = n \int_{\underline{x}_i}^{\bar{x}_i} f(x)dx = n(F(\bar{x}_i) - F(\underline{x}_i)) \tag{4}$$

where $f(x)$ is the specified probability density function, $F(x)$ is the corresponding cumulative distribution, and $\underline{x}_i$, $\overline{x}_i$ bound the $i^{th}$ bin. Note that the cumulative distribution function is given as:

$$F(x) = P(X \leq x) \tag{5}$$

where $F(x)$ is the corresponding cumulative probability function and $x$ denotes an instance of the random variable $X$. The test statistic follows, approximately, a $\chi^2$ distribution with a degree of freedom $k-3$, and the hypothesis is defined as follows:

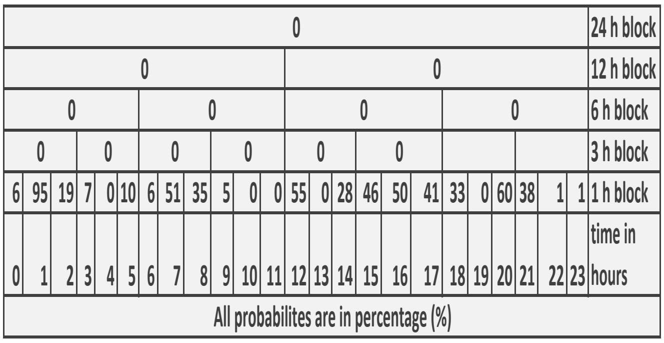
$$\text{if } \chi^2 < \chi^2_\alpha \text{ , } H_0 = 0, \text{ the Hypothesis will be accepted.} \tag{6}$$

otherwise,

$$H_0 = 1, \text{ the Hypothesis will be rejected.} \tag{7}$$

where $\chi^2_\alpha$ corresponds to the known distribution with $k-3$ degree of freedom and an $\alpha$ level of significance. The chi-square test results, using the MATLAB® command `chi2gof` for various time blocks are summarized in Table I. The table shows the probability (in %) that a given distribution is normal (the probabilities less than $10^{-4}$ were rounded off to zero).

TABLE I
SUMMARY OF TEST RESULTS



| | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | | | | | | | | | | | | | | | | | | | | | | | | 24 h block |
| 0 | | | | | | | | | | | | 0 | | | | | | | | | | | | 12 h block |
| 0 | | | | | | 0 | | | | | | 0 | | | | | | 0 | | | | | | 6 h block |
| 0 | | | 0 | | | 0 | | | 0 | | | 0 | | | 0 | | | | | | | | | 3 h block |
| 6 | 95 | 19 | 7 | 0 | 10 | 6 | 51 | 35 | 5 | 0 | 0 | 55 | 0 | 28 | 46 | 50 | 41 | 33 | 0 | 60 | 38 | 1 | 1 | 1 h block |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | time in hours |
| All probabilites are in percentage (%) | | | | | | | | | | | | | | | | | | | | | | | | |

Lange [12] and Landberg [10] have claimed that the wind-speed forecast error follows a normal distribution for 12h and 24h prediction. But the results in Table I show that wind-speed forecast error will not be normally distributed for 12h and 24h prediction time blocks, so the claim by Lange and Landberg is not valid for the given data; the reason could be that wind-speed error distributions vary from region to region [10]. But 11 one hourly prediction distributions are normally distributed with probability ranging from 33 to 95%, 2 ranging from 10 to 33% and 11 below 10%. At this point, in a case where probabilities are low, it is usually desirable to repeat the test with a larger sample size, irrespective of whether the initial statistical chi-square test gives low probability that the distribution is normal. Given the limitation on our data set (291 points for each hour), and the coarseness of the $\chi^2$ test (due to sorting data into bins, which obscures the fine-grain detail of the data set) a different approach to investigate the normality of wind-speed forecast error distribution was considered.

### C. Gaussian Distribution Fitting

To investigate the normality of wind-speed forecast error cumulative distributions further, each one-hourly distribution was fitted by a GD using the MATLAB® command `normfit`. Then, using the statistics obtained from `normfit`, 1000 synthetic normal forecasting error samples were generated for each hourly distribution. The resulting cumulative distribution for the synthetic normal sample was plotted along with the actual distribution for the same bin width for each hour, as shown in Figs. 3 and 4; all 24 cases are also shown in the Appendix (section V).

The plots in section V show that the normal cumulative distributions closely fit the actual cumulative distributions for each one-hour data window. It is particularly noteworthy that the fits are in good agreement on the “tails” of the distribution ($F(x) \leq 0.1$ and $F(x) \geq 0.9$), where such fitting processes are most likely to be problematic.

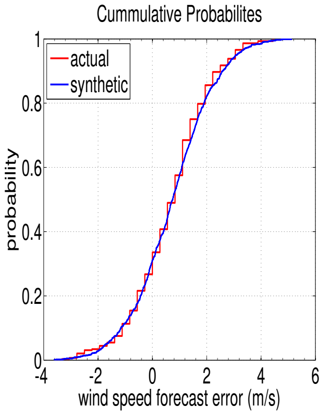


Fig. 3. 01:00:00 to 1:59:59

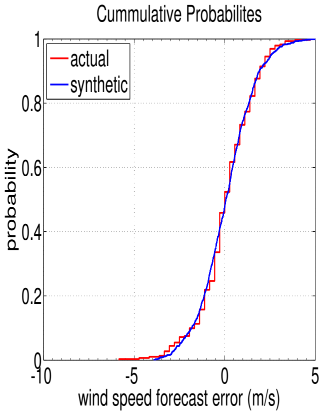


Fig. 4. 09:00:00 9:59:59

Therefore, we believe it can be assumed that all of the hourly distributions are GDs; given the nature of wind-speed forecasting, the very slight variation between the actual and cumulative distribution at a few points will not have a significant impact on the quality of modeling and forecasting.

## III. CONCLUSION

This paper has shown through statistical analysis that the hourly wind-speed prediction error distributions were approximately Gaussian in nature. However, it is important to provide a rigorous validation of the Gaussian assumption by having larger forecast and actual data sets to process. The advantage of our approach to modeling prediction error is that the local weather conditions are already considered by EC. Also, since prediction error follows a normal distribution, it can be modeled accurately by one of the many well developed time-series models for random variables characterized by a GD.

It was also shown that the statistics of the wind-speed prediction error in our data sets do not increase significantly as time increases. This is significant since the models developed using EC’s prediction error distributions can be directly applicable to the existing electricity market rules.

## IV. ACKNOWLEDGEMENT

The first-named author is grateful to the Atlantic Innovation Fund, Canada and to Dr. Liuchen Chang for the financial support afforded during the course of this work.

## REFERENCES

[1] N. B. System Operator, “New Brunswick electricity market rules,” Tech. Rep., September 2007.

[2] E. Cadenas and W. Rivera, “Wind speed forecasting in the south coast of Oaxaca, Mexico,” *Renewable Energy*, vol. 32, no. 12, pp. 2116–2128, 2007.

[3] J. Torres, A. García, M. D. Blas, and A. D. Francisco, “Forecast of hourly average wind speed with ARMA models in Navarre (Spain),” *Solar Energy*, vol. 79, no. 1, pp. 65–77, 2005.

[4] B. G. Brown, R. W. Katz, and A. H. Murphy, “Time series models to simulate and forecast wind speed and wind power,” *Journal of Climate and Applied Meteorology*, vol. 23, no. 8, pp. 1184–1195, 1984.

[5] E. García-Bustamante, J. F. González-Rouco, P. A. Jiménez, J. Navarro, and J. P. Montávez, “The influence of the Weibull assumption in monthly wind energy estimation,” *Wind Energy*, vol. 11, no. 5, pp. 483–502, 2008.

[6] M. Jamil, S. Parsa, and M. Majidi, “Wind power statistics and an evaluation of wind energy density,” *Renewable Energy*, vol. 6, no. 5-6, pp. 623–628, 1995.

[7] H. Holttinen, “Optimal electricity market for wind power,” *Energy Policy*, vol. 33, no. 16, pp. 2052–2063, 2005.

[8] Environment Canada, “Low-resolution CMC GRIB database,” http://www.weatheroffice.gc.ca/grib/.

[9] S. Salcedo-Sanz, A. M. Perez-Bellido, E. G. Ortiz-Garcia, A. Portilla-Figueras, L. Prieto, and D. Paredes, “Hybridizing the fifth generation mesoscale model with artificial neural networks for short-term wind spe ed prediction,” *Renewable Energy*, vol. 34, no. 6, pp. 1451–1457, 2009.

[10] L. Landberg, “Short-term prediction of local wind conditions,” *Journal of Wind Engineering and Industrial Aerodynamics*, vol. 89, no. 3-4, pp. 235–245, 2001.

[11] National Electricity Market Management Company Ltd. (NEMMCO), “Forecasting intermittent generation in the national electricity market,” Tech. Rep., February 2004.

[12] M. Lange, “On the uncertainty of wind power predictions—analysis of the forecast accuracy and statistical distribution of errors,” *Journal of Solar Energy Engineering*, vol. 127, no. 2, pp. 177–184, 2005.

## V. APPENDIX

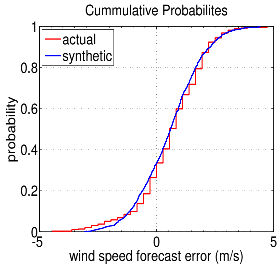


Fig. 5. 00:00h to 0:59:59h

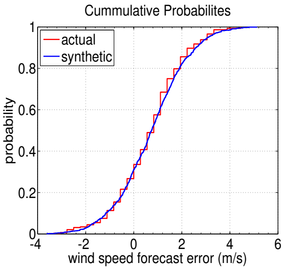


Fig. 6. 1:00h to 1:59:59h

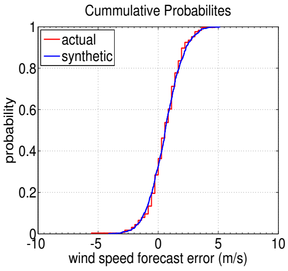


Fig. 7. 2:00h to 2:59:59h

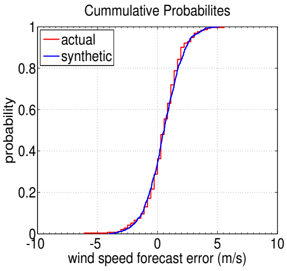


Fig. 8. 3:00h to 3:59:59h

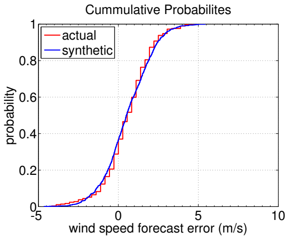


Fig. 9. 04:00h to 4:59:59h

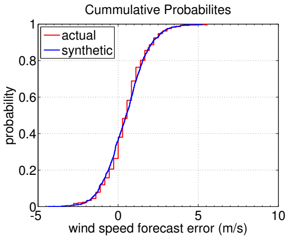


Fig. 10. 5:00h to 5:59:59h

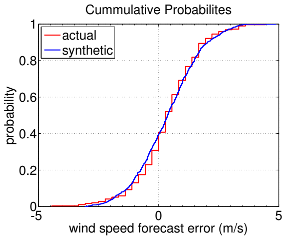


Fig. 11. 6:00h to 6:59:59h

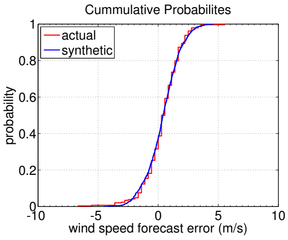


Fig. 12. 7:00h to 7:59:59h

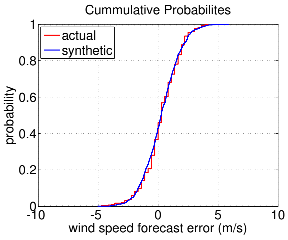


Fig. 13. 08:00h to 8:59:59h

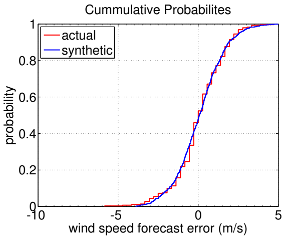


Fig. 14. 9:00h to 9:59:59h

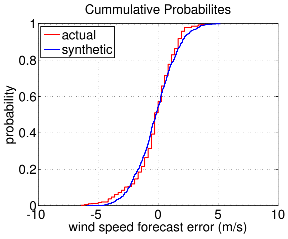


Fig. 15. 10:00h to 10:59:59h

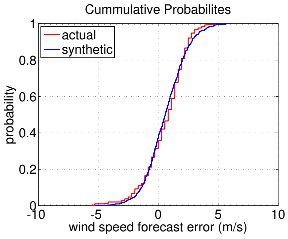


Fig. 16. 11:00h to 11:59:59h

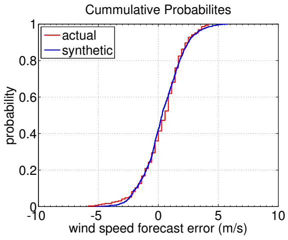


Fig. 17. 12:00h to 12:59:59h

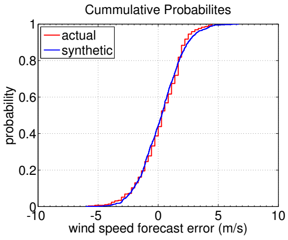


Fig. 18. 13:00h to 13:59:59h

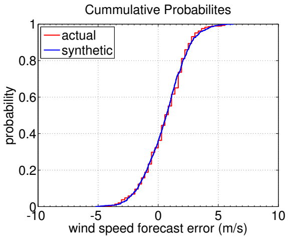


Fig. 19. 14:00h to 14:59:59h

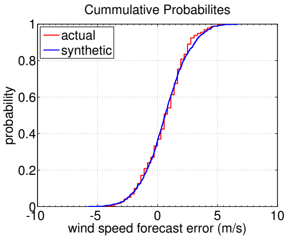


Fig. 20. 15:00h to 15:59:59h

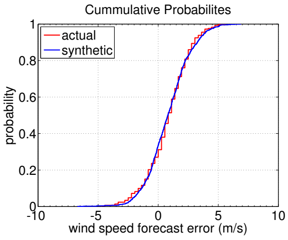


Fig. 21. 16:00h to 16:59:59h

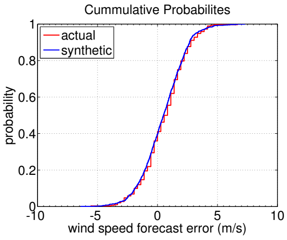


Fig. 22. 17:00h to 17:59:59h

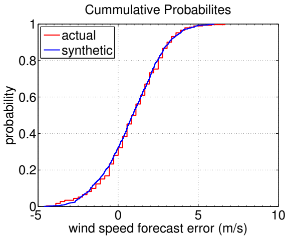


Fig. 23. 18:00h to 18:59:59h

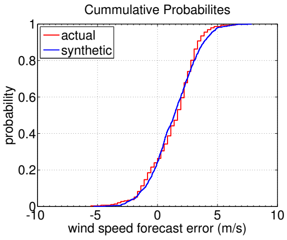


Fig. 24. 19:00h to 19:59:59h

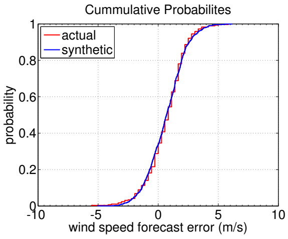


Fig. 25. 20:00h to 20:59:59h

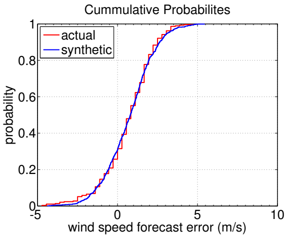


Fig. 26. 21:00h to 21:59:59h

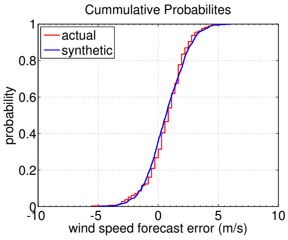


Fig. 27. 22:00h to 22:59:59h

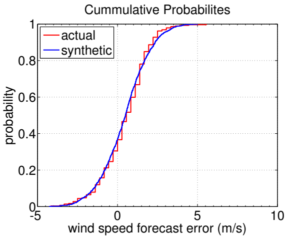


Fig. 28. 23:00h to 23:59:59h