Evaluation of Parametric Statistical Models for Wind Speed Probability Density Estimation

Maisam Wahbah
Electrical and Computer Engineering
Khalifa University
Abu Dhabi, United Arab Emirates
maisam.wahbah@ku.ac.ae

Omar Alhussein
Electrical and Computer Engineering
University of Waterloo
Ontario, Canada
oalhussein@uwaterloo.ca

Tarek H.M. EL-Fouly
Electrical and Computer Engineering
Khalifa University
Abu Dhabi, United Arab Emirates
tarek.elfouly@ku.ac.ae

Bashar Zahawi
Electrical and Computer Engineering
Khalifa University
Abu Dhabi, United Arab Emirates
bashar.zahawi@ku.ac.ae

Sami Muhaidat
Electrical and Computer Engineering
Khalifa University
Abu Dhabi, United Arab Emirates
sami.muhaidat@ku.ac.ae

***Abstract*—An accurate statistical estimation of wind speed probability density at a given site is crucial when making power network planning decisions involving wind generation resources. The use of parametric probability density functions, such as the Rayleigh, Weibull and Gaussian distributions, can be problematic as it can lead to model mis-specification at a given site. In this paper, the use of the Gaussian Mixture Model (GMM) to estimate wind speed variability is investigated and compared with the above three popular parametric models using wind speed data for six sites in northwest Europe. Results show that the GMM produces the lowest error values with the highest percentage improvements, and is the only model that consistently fails to reject the null hypothesis when conducting the K–S goodness-of-fit test.**

***Index Terms*—Density estimation, Gaussian mixture model, statistical analysis, wind speed models.**

I. Introduction

A reliable probabilistic model of wind speed is required to assess the impact of the large-scale integration of this intermittent source of renewable energy into the power system. Small differences in estimation can lead to significantly different decisions in terms of turbine location and network planning studies [1], [2]. Parametric Families of Probability Distributions (PFPDs), such as the Rayleigh, Weibull and Gaussian families of distributions, have been widely used in the power engineering literature to model wind speed variability. Such studies include, but are not limited to, wind power potential assessment in power system planning [3], voltage stability for networks with distributed generation [4], wind turbine power productivity prediction [5], optimal allocation of energy storage systems in distribution networks [6], power reliability assessment in hybrid power generating units [7], and multiple microgrid design and clustering studies [8]. However, the use of simple PFPDs can be problematic, leading to the risk of model mis-specification.

In this paper, we propose the use of the Gaussian Mixture Model (GMM) [9], which can be expressed as a finite convex linear combination of Gaussian densities (each with different probability density parameters). In fitting a finite mixture distribution, the determination of an appropriate number of mixture components is of paramount importance. Selecting a small number of components yields an inaccurate representation, while a very large number of components unnecessarily results in an over-fitting and an increase in complexity. In this paper, we adopt the Expectation–Maximization (EM) framework to estimate the parameters of the GMM and further consider the Bayesian Information Criterion (BIC) in order to optimally determine the number of mixture components.

The performance of the GMM is compared with three parametric probability distributions (Rayleigh, Weibull, and Gaussian distributions) which are most widely used to estimate wind speed probability density using hourly wind speed data at six sites in northwest Europe. Evaluations are carried out using Kolmogorov–Smirnov (K–S) goodness-of-fit test, and two statistical error measures: Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). Results show that GMM yields the lowest MAE and RMSE values with percentage improvements of up to 77.9% when compared with the Rayleigh distribution (MAE), and up to 42.8% when compared with the Weibull distribution (RMSE).

The remainder of the paper is organized as follows. Section II gives statistical information about wind speed data sets used in the study. Section III provides a general overview of popular parametric probability density distributions used in estimating wind speed variability. A description of the proposed Gaussian mixture model is given in Section IV. Section V provides a full analysis of the resulting probability density estimates including an evaluation of the performance of the proposed EM-based GMM approach. Conclusions are presented in Section VI.

II. Data Summary

Hourly wind speed data (for the year 2014) was obtained from the interactive web platform www.renewables.ninja [10] for six sites in northwest Europe: Copenhagen in Denmark, Kiel in Germany, Malin Head in Ireland, Black Law, Whitelee and Schaw in Scotland. Fig. 1 shows the geographical locations of the six sites, and Fig. 2 shows the time series plots of hourly wind speed data for a period of a week. Table I lists the latitude and longitude coordinates in addition to wind speed statistical information including the maximum, minimum, mean, median, and standard deviation.

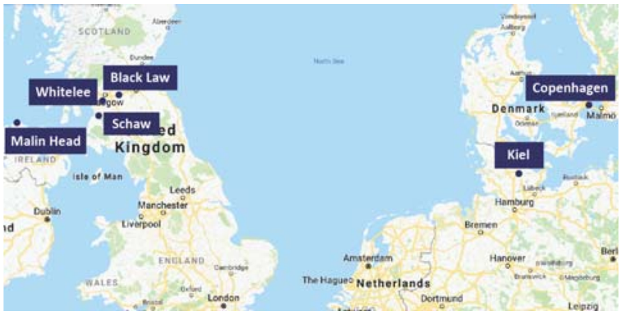


Fig. 1: Geographical locations of the six selected wind speed sites.

III. Popular Parametric Models for Wind Speed Density Estimation

Wind speed probability distribution can be estimated using the Rayleigh [6], Weibull [3], and Gaussian distributions [5]. The formulas for the Probability Density Functions (PDFs) for the three selected distributions are listed in Table II, where X is the wind speed random variable.

IV. Gaussian Mixture Model

Let the j^{th} entry of the random wind speed data $\boldsymbol{x}$=$\{x_1, \ldots, x_n\}$ be modeled as a finite convex linear combination of Gaussian densities,

$$f_X(x_j|\theta) = \sum_{i=1}^{C} \omega_i \, \phi(x_j, \theta_i), \quad x_j \geq 0, \, j = 1, \ldots, N \tag{1}$$

where C is the number of mixture components. Each i^{th} mixture component is expressed as

$$\phi\left(x_j, \theta_i\right) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x_j - \mu_i)^2}{2\sigma_i^2}\right), \tag{2}$$

where the weight of the i^{th} component is $\omega_i > 0$, with $\sum_i^C \omega_i = 1$, and $\theta = (\{\omega_i, \mu_i, \sigma_i^2\}_{i=1}^{C})$. The parameters μ_i and σ_i^2 correspond to the mean and variance of the i^{th} component, respectively.

The GMM inherits the advantages of the Gaussian distribution, where its individual densities are efficiently characterized by the first two moments [11], [12], and have the so-called "universal-approximation" property. This was proven by the Weiner's approximation theory [9] which proves that the GMM distribution can approximate any arbitrarily shaped non-Gaussian density. The EM framework is employed in this paper to estimate the GMM parameters [13].

Ideally, one would like to determine the parameters (θ) that maximize the log-likelihood function ($\ln \Pr(\boldsymbol{x}|\theta, C)$). Maximizing the log-likelihood function is analytically intractable [13], [14]. Instead, the EM algorithm solves the Maximum Likelihood Estimation (MLE) problem through two iterative steps: (1) Expectation Step and (2) Maximization Step. By adopting the EM framework, the MLE problem is approached by maximizing the expected log-likelihood of the data through an iterative approach. The EM algorithm, however, requires the number of mixture components (C) as an *a priori* input. As C is increased, the log-likelihood function can be maximized further at the expense of increased complexity in the model, and can lead to over-fitting.

To resolve this issue, a simple yet effective unsupervised information theoretic approach, called the BIC approach is adopted [15], [16]. The BIC adds a penalty term to the log-likelihood function as C in increased, as follows

$$\mathrm{BIC}(C) = -2 \ln \Pr(\hat{\theta}|\boldsymbol{x}, \, C) + C \, \ln(n) \tag{3}$$

where $\hat{\theta}$ is the parameter such that the log-likelihood function is maximized. In this paper, the BIC is employed in conjunction with the EM algorithm to find an appropriate number of component, while maximizing the log-likelihood function. The EM framework is terminated when the BIC measure does not improve. In a large-sample setting, the number of components determined by the minimum BIC is asymptotically optimal from the perspective of the Bayesian posterior probability ($\Pr(\theta|\boldsymbol{x}, \, C)$) [15]. It is noteworthy to point out that the EM algorithm is guaranteed not to get worse as it iterates [13]. Moreover, it has an advantage of being a completely unsupervised learning algorithm, which makes it very convenient for wind speed probability density estimation.

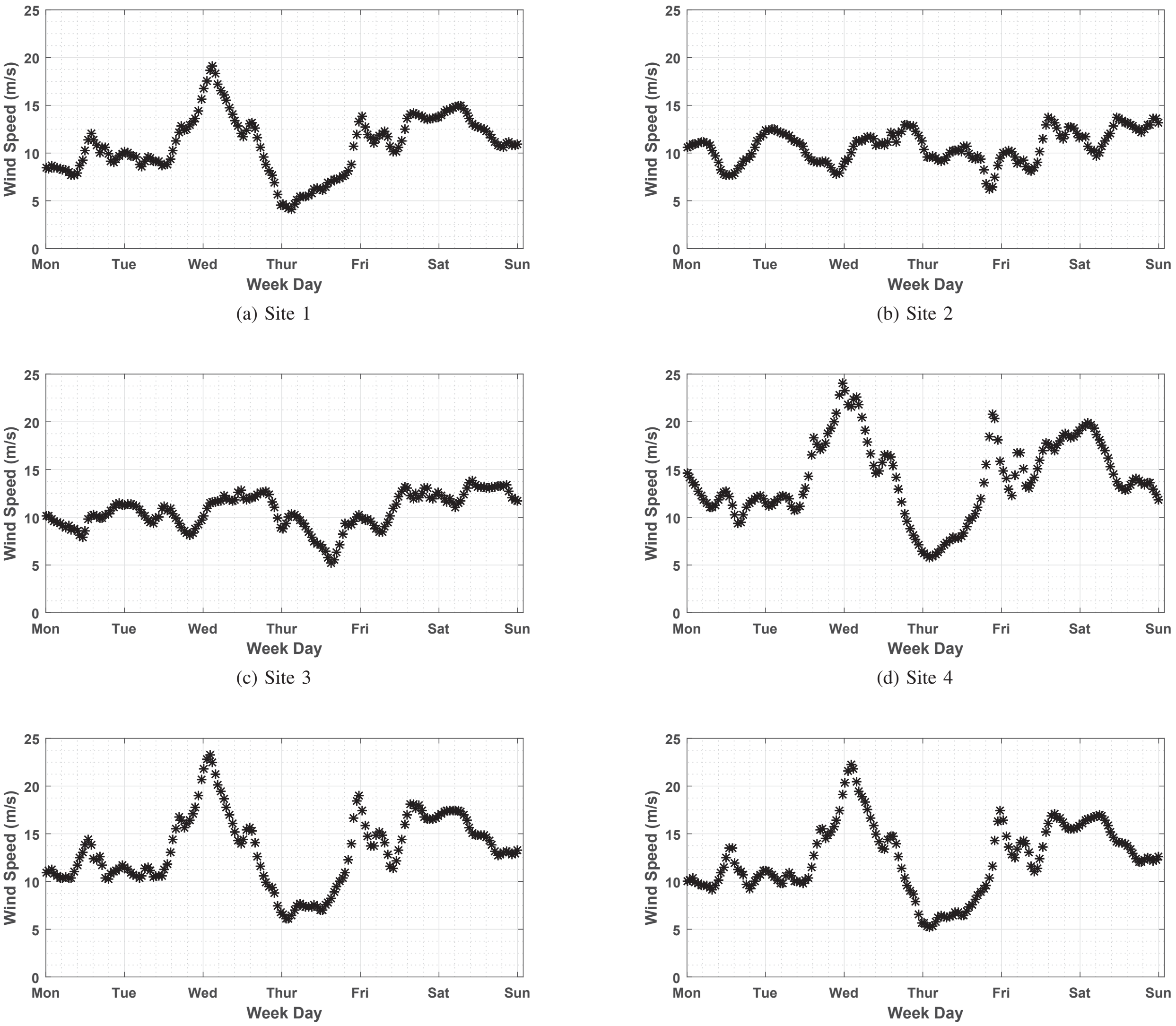
V. Results

In this section, the performance of the EM-based GMM approach for wind speed probability density estimation is assessed via comparisons with three popular parametric probability density models (Rayleigh, Weibull, and Gaussian distributions) using hourly wind speed data from six sites in northwest Europe. The BIC-assisted EM-based GMM routine, described in Section IV, was written in MATLAB. In addition, the function `raylfit`, `wblfit`, `mle`, `raylpdf`, `wblpdf` and `normpdf` were used to obtain the parameters and the density estimates for the three PFPDs in Table II. Data histograms were produced using MATLAB's `histogram` function with density scaling.

Fig. 3 shows the normalized histogram plots of wind speed data at the six sites, overlaid with the three selected parametric probability distributions and the Gaussian mixture PDF. The parameters of the three selected popular distributions were estimated using maximum likelihood estimation, and are listed

TABLE I: Wind Speed Data Information

Site	**Location**	**Geographical Coordinates (°)**		**Wind Speed (m/s)**				
		Latitude	**Longitude**	**Maximum**	**Minimum**	**Mean**	**Median**	**Standard Deviation**
Site 1	Copenhagen, Denmark	55.69	12.57	23.69	2.41	9.10	9.09	2.82
Site 2	Kiel, Germany	54.46	10.20	19.17	2.63	8.87	8.90	2.73
Site 3	Black Law, Scotland	55.83	-3.82	19.10	1.85	7.40	7.01	2.79
Site 4	Whitelee, Scotland	55.71	-4.34	22.20	2.06	8.43	7.95	3.23
Site 5	Schaw, Scotland	55.46	-4.46	23.26	2.40	9.06	8.63	3.31
Site 6	Malin Head, Ireland	55.38	-7.40	24.32	2.37	10.05	9.53	3.80



(a) Site 1

(b) Site 2

(c) Site 3

(d) Site 4

(e) Site 5

(f) Site 6

Fig. 2: Time series of hourly wind speed data for a period of a week.

TABLE II: Popular Parametric Families of Probability Distributions

Distribution	**Probability Density Function**	**Domain**	**Parameters**
Rayleigh	$f_X(x) = \frac{x}{b^2}\, e^{-(x^2/2b^2)}$	$x \in [0, \infty)$	Scale: $b \in (0, \infty)$
Weibull	$f_X(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}$	$x \in [0, \infty)$	Scale: $\lambda \in (0, \infty)$ Shape: $k \in (0, \infty)$
Gaussian	$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$x \in (-\infty, \infty)$	Mean: $\mu \in (-\infty, \infty)$ Variance: $\sigma^2 \in (0, \infty)$

TABLE III: Parametric Probability Distributions Parameters

Site	**Rayleigh**	**Weibull**		**Gaussian**	
	b	λ	k	μ	σ^2
Site 1	6.73	10.09	3.50	9.09	7.89
Site 2	6.54	9.81	3.55	8.84	7.35
Site 3	5.62	8.35	2.83	7.43	7.86
Site 4	6.41	9.52	2.78	8.46	10.56
Site 5	6.85	10.21	2.91	9.10	11.09
Site 6	7.64	11.35	2.82	10.10	14.59

TABLE IV: p-values for Kolmogorov–Smirnov Test

Site	**Rayleigh**	**Weibull**	**Gaussian**	**GMM**
Site 1	6.41×10^{-80}	$\mathbf{24.53\times10^{-02}}$	$\mathbf{58.67\times10^{-02}}$	$\mathbf{6.70\times10^{-02}}$
Site 2	2.00×10^{-90}	$\mathbf{5.31\times10^{-02}}$	$\mathbf{27.27\times10^{-02}}$	$\mathbf{5.57\times10^{-02}}$
Site 3	4.18×10^{-39}	6.85×10^{-10}	2.37×10^{-13}	$\mathbf{84.42\times10^{-02}}$
Site 4	9.28×10^{-38}	2.59×10^{-09}	5.15×10^{-15}	$\mathbf{49.65\times10^{-02}}$
Site 5	9.01×10^{-49}	2.66×10^{-12}	9.36×10^{-14}	$\mathbf{18.71\times10^{-02}}$
Site 6	8.88×10^{-34}	2.76×10^{-08}	7.91×10^{-13}	$\mathbf{23.18\times10^{-02}}$

in Table III. The parameters of the Gaussian mixture PDF were estimated using EM algorithm and the number of optimal mixture components are determined using the BIC approach. Visually, these plots show significant discrepancies between all three popular probability distributions and the data histogram, and more faithful estimates from the GMM method. In the remainder of this section, the performance of the GMM is evaluated alongside the three selected parametric distributions using K–S goodness-of-fit test, and two standard statistical error measures: RMSE and MAE.

A. Kolmogorov–Smirnov Goodness-of-fit Test

The Kolmogorov–Smirnov test is used to test if a given data set comes from a particular statistical model. For each of the four distributions under consideration, the question is asked whether the null hypothesis that the observed data are independently sampled from that model can be rejected. The fundamentals behind the K–S goodness-of-fit test can be found in standard statistics textbooks [17]. Results are presented in terms of a p-value which, informally, is a measure of the evidence against the null hypothesis, with low p-values (e.g. $p < 0.01$) corresponding to strong evidence against the null hypothesis. It is important to note, however, that a high p-value may not be interpreted as an evidence supporting a particular model. p-values resulting from the K–S test for each model are presented in Table IV, where the values highlighted in bold indicate a failure to reject the null hypothesis. The GMM is the only model that consistently produces p-values that indicate a failure to reject the null hypothesis. On the other hand, the three popular parametric models produce low p-values, suggesting they are unsuitable for modeling wind speed data. As such, the proposed approach is a more suitable candidate for modeling the given wind speed data.

B. Root Mean Square Error

Root Mean Square Error provides a general-purpose error measure and is a common tool for numerical comparisons when assessing PDF estimates. RMSE is defined as the square root of the average of the square of the errors between the observed and expected probabilities:

$$RMSE = \sqrt{\frac{1}{t}\sum_{i=1}^{t}(y_i - \hat{y_i})^2} \tag{4}$$

where t is the number of bins of data, y_i is the probability of wind speed being within bin i calculated from the data set, and $\hat{y_i}$ is the probability within the same bin calculated from the estimated data set ($i = 1, \ldots, t$).

Table V lists the computed RMSE values for the six sites together with percentage improvements calculated with respect to the Rayleigh distribution error. The proposed GMM gives the lowest RMSE value among all the evaluated methods for all six sites with percentage improvements between 51.5%-77.2% (at an average of 61.3%). As expected, parametric distributions produced the highest RMSE values.

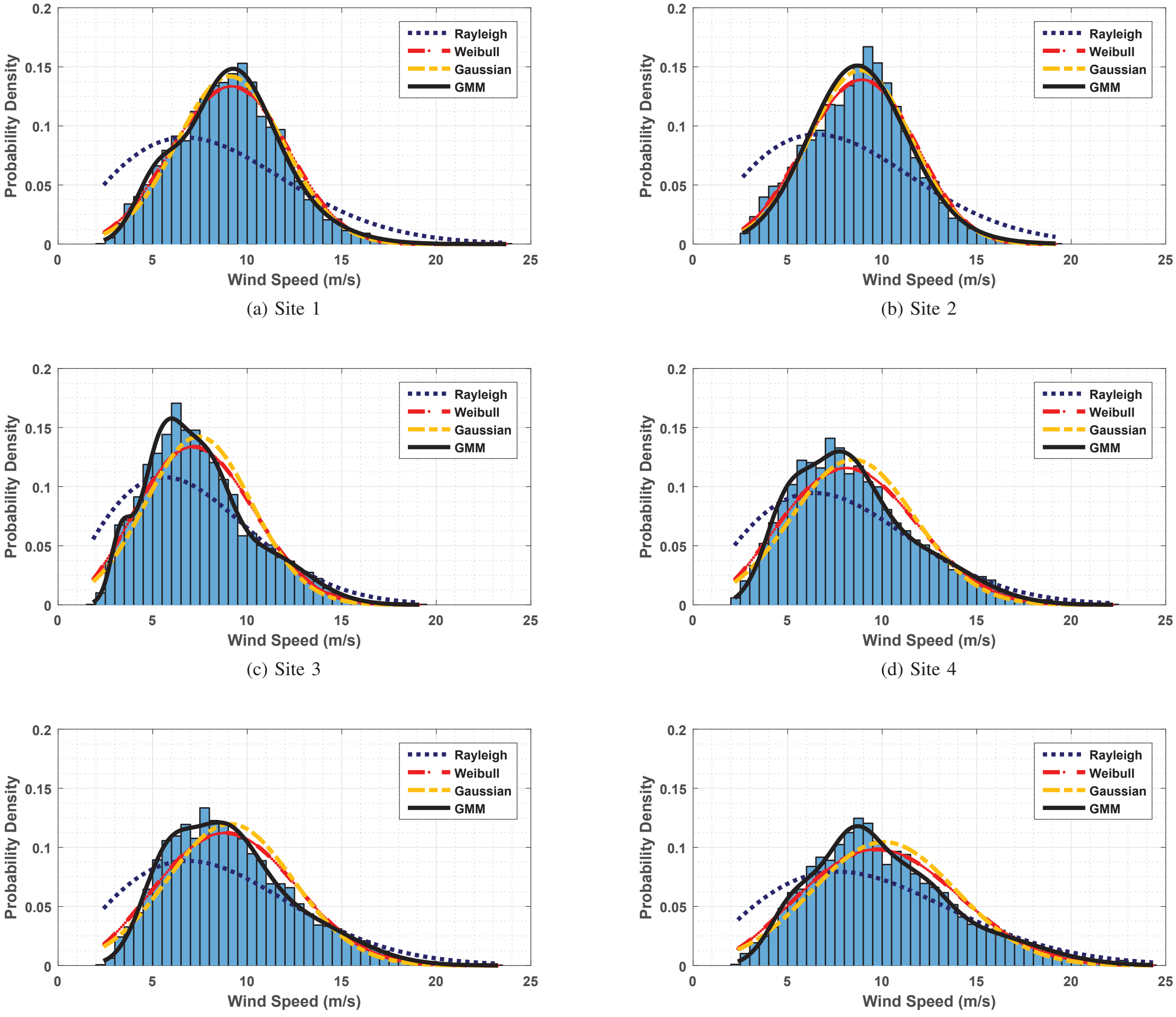
C. Mean Absolute Error

Mean Absolute Error is a global error measure metric that calculates the average of the absolute values of the deviations of the probability within bin i calculated from the estimated data set, from the probability of wind speed being within the same bin, and is given by:

$$MAE = \frac{1}{t}\sum_{i=1}^{t}|y_i - \hat{y_i}| \tag{5}$$

where t, y_i, and $\hat{y_i}$ are the same as in (4).

Table VI lists the computed MAE values and the corresponding percentage improvements (with respect to the Rayleigh distribution) for the six sites. Again, the proposed



(a) Site 1

(b) Site 2

(c) Site 3

(d) Site 4

(e) Site 5

(f) Site 6

Fig. 3: Probability density plots and histograms.

TABLE V: Root Mean Square Error and Percentage Improvements

Site	Rayleigh	Weibull	Gaussian	GMM
Site 1	1.24×10^{-2} (–)	3.07×10^{-3} (75.2%)	2.95×10^{-3} (76.3%)	2.83×10^{-3} **(77.2%)**
Site 2	1.19×10^{-2} (–)	4.20×10^{-3} (64.6%)	4.00×10^{-3} (66.3%)	3.81×10^{-3} **(67.9%)**
Site 3	8.85×10^{-3} (–)	5.76×10^{-3} (35.0%)	6.45×10^{-3} (27.1%)	3.29×10^{-3} **(62.8%)**
Site 4	8.64×10^{-3} (–)	5.88×10^{-3} (32.0%)	6.75×10^{-3} (21.9%)	4.19×10^{-3} **(51.5%)**
Site 5	9.05×10^{-3} (–)	5.77×10^{-3} (36.2%)	6.34×10^{-3} (29.9%)	3.89×10^{-3} **(56.9%)**
Site 6	8.22×10^{-3} (–)	5.33×10^{-3} (35.2%)	6.09×10^{-3} (25.9%)	3.98×10^{-3} **(51.6%)**

TABLE VI: Mean Absolute Error and Percentage Improvements

Site	Rayleigh	Weibull	Gaussian	GMM
Site 1	9.02×10^{-3} (–)	2.26×10^{-3} (75.0%)	2.05×10^{-3} (77.3%)	2.00×10^{-3} **(77.9%)**
Site 2	8.78×10^{-3} (–)	2.91×10^{-3} (66.8%)	2.79×10^{-3} (68.2%)	2.42×10^{-3} **(72.4%)**
Site 3	5.81×10^{-3} (–)	3.94×10^{-3} (32.2%)	4.69×10^{-3} (19.3%)	2.34×10^{-3} **(59.7%)**
Site 4	5.82×10^{-3} (–)	4.01×10^{-3} (31.0%)	4.85×10^{-3} (16.6%)	2.66×10^{-3} **(54.2%)**
Site 5	6.57×10^{-3} (–)	4.07×10^{-3} (38.1%)	4.51×10^{-3} (31.3%)	2.86×10^{-3} **(56.5%)**
Site 6	5.56×10^{-3} (–)	3.83×10^{-3} (31.2%)	4.43×10^{-3} (20.3%)	2.93×10^{-3} **(47.3%)**

GMM produces the lowest errors with percentage improvements between 47.3%-77.9% (at an average of 61.3%), and consistent with previous results, all three selected parametric distributions produced much higher error values in line with our previous observations.

VI. Conclusion

This paper presents an evaluation of the performance of BIC-assisted EM-based GMM for obtaining the probability density estimate of wind speed at a given site for use in estimating the electric power generation from wind turbines/farms needed in power system planning and reliability studies. In this work, the EM algorithm is adopted to estimate the parameters of the Gaussian mixture PDF, and the BIC approach is considered to optimally determine the number of mixture components. The proposed model is assessed against three popular parametric wind speed models (Rayleigh, Weibull, and Gaussian distributions) using hourly wind speed data at six sites in northwest Europe. Evaluations are carried out using K–S goodness-of-fit test, and two standard statistical error measures. Based on the studied data sets, all three popular parametric models were shown to be inadequate for modeling wind speed data. Results also confirm the suitability of the GMM in obtaining accurate wind speed probability density estimates producing the lowest error values with the highest percentage improvements, and is the only model that consistently fails to reject the null hypothesis when conducting the goodness-of-fit test.

References

[1] J. Kabouris and F. D. Kanellos, “Impacts of large-scale wind penetration on designing and operation of electric power systems,” *IEEE Transactions on Sustainable Energy*, vol. 1, no. 2, pp. 107–114, 2010.

[2] K. De Vos, J. Morbee, J. Driesen, and R. Belmans, “Impact of wind power on sizing and allocation of reserve requirements,” *IET Renewable Power Generation*, vol. 7, no. 1, pp. 1–9, 2013.

[3] B. G. Kumaraswamy, B. K. Keshavan, and S. H. Jangamshetti, “A statistical analysis of wind speed data in west central part of karnataka based on weibull distribution function,” in *2009 IEEE Electrical Power Energy Conference (EPEC)*, pp. 1–4, Oct 2009.

[4] R. S. Al Abri, E. F. El-Saadany, and Y. M. Atwa, “Optimal placement and sizing method to improve the voltage stability margin in a distribution system using distributed generation,” *IEEE Transactions on Power Systems*, vol. 28, no. 1, pp. 326–334, 2013.

[5] K. Sunderland, T. Woolmington, J. Blackledge, and M. Conlon, “Small wind turbines in turbulent (urban) environments: A consideration of normal and weibull distributions for power prediction,” *Journal of Wind Engineering and Industrial Aerodynamics*, vol. 121, pp. 70 – 81, 2013.

[6] A. S. Awad, T. H. El-Fouly, and M. M. Salama, “Optimal ESS allocation for benefit maximization in distribution networks,” *IEEE Transactions on Smart Grid*, vol. 8, no. 4, pp. 1668–1678, 2017.

[7] A. Kumar, M. Zaman, N. Goel, N. Goel, and R. Church, “Probabilistic reliability assessment in the optimization of hybrid power generating units,” in *2012 IEEE Electrical Power and Energy Conference*, pp. 75–79, Oct 2012.

[8] S. A. Arefifar, Y. A.-r. I. Mohamed, and T. El-Fouly, “Optimized multiple microgrid-based clustering of active distribution systems considering communication and control requirements,” *IEEE Transactions on Industrial Electronics*, vol. 62, no. 2, pp. 711–723, 2015.

[9] N. Kostantinos and Stergiopoulos, *Gaussian mixtures and their applications to signal processing*. CRC Press, 2000.

[10] I. Staffell and S. Pfenninger, “https://www.renewables.ninja/,” 2016.

[11] E. Patrick, *Fundamentals of Pattern Recognition*. Prentice-Hall information and system sciences series, Pearson Education, Limited, 1972.

[12] P. Ales, *Signal analysis and prediction*. Birkhauser, 1998.

[13] A. Dempster, N. Laird, and D. Rubin, “Maximum likelihood from incomplete data via the em algorithm.,” *Journal of the Royal Statistical Society*, vol. 29, pp. 1–38, 1976.

[14] B. Selim, O. Alhussein, S. Muhaidat, G. K. Karagiannidis, and J. Liang, “Modeling and analysis of wireless channels via the mixture of gaussian distribution,” *IEEE Transactions on Vehicular Technology*, vol. 65, pp. 8309–8321, Oct 2016.

[15] G. Schwarz, “Estimating the dimension of a model,” *The Annals of Statistics*, vol. 6, no. 2, p. 461–464, 1978.

[16] O. Alhussein, I. Ahmed, J. Liang, and S. Muhaidat, “Unified analysis of diversity reception in the presence of impulsive noise,” *IEEE Transactions on Vehicular Technology*, vol. 66, pp. 1408–1417, Feb 2017.

[17] M. DeGroot and M. Schervish, *Probability and Statistics*. London: Pearson Education, Inc., 4th ed., 2012.