Gaussian Mixture Model for Estimating Solar Irradiance Probability Density

Maisam Wahbah
Department of Biomedical Engineering
Khalifa University of Science and Technology
Abu Dhabi, United Arab Emirates
maisam.wahbah@ku.ac.ae

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

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

2020 IEEE Electric Power and Energy Conference (EPEC) | 978-1-7281-6489-2/20/$31.00 ©2020 IEEE | DOI: 10.1109/EPEC48502.2020.9320011

***Abstract*—The increasing penetration of photovoltaic generation resources make it imperative for power network designers to assess the available resources by obtaining accurate estimates of solar irradiance at a given site/geographical area. The parametric Beta distribution has long been a popular choice in such studies; however, the use of parametric functions for probability density estimation (such as the Beta distribution) can be problematic and may lead to model mis-specification. The Gaussian Mixture Model (GMM) is proposed in this paper to provide a more robust estimation of solar irradiance probability density at a certain site. Multi-year solar data from eight locations in the United States is utilized to evaluate the accuracy of the GMM estimate and compare its performance with the popular Beta distribution. Assessments are carried out using three standard measures of error, coefficient of determination, and the Kolmogorov–Smirnov goodness-of-fit test for distributional adequacy. Results demonstrate that the GMM estimate produces a more robust estimation with better performance metrics when compared with the Beta distribution.**

***Index Terms*—Gaussian mixture model, parametric statistics, probability density estimation, solar irradiance models.**

I. Introduction

Robust statistical models of solar irradiance levels are needed when making decisions about the location of solar farms, and evaluating the effect of the large-scale integration of Photovoltaic (PV) generation [1] into the local power network. The Beta distribution has long been the parametric distribution of choice used in network design and planning studies to model the variation of solar resources due to its ease of use and simplicity. Such studies have been widely reported in the literature, for example, research works on network planning and design [2], [3], distribution system partitioning into microgrids [4], and voltage stability for networks with distributed generation [5]. Examples of more recent studies employing the Beta distribution include, but are not limited to, clustering and design of microgrids [6], storage planning in distribution networks [7], and energy management studies [8], [9]. However, the use of a simple parametric distribution (such as the Beta distribution) can possibly lead to the risk of mis-specification of the model (in which the best estimators fail to be robust) and potentially results in wrong planning decisions. To overcome such limitations, recent works have suggested new models for estimating solar irradiance including: a model [10] based on a nonparametric approach (which is data-driven and not defined by a small number of parameters), and a hybrid model combining parametric and nonparametric approaches [11]. These methods potentially produce robust estimates, but require multiple trials or heavy computational effort.

In this paper, the Gaussian Mixture Model (GMM) is proposed for estimating solar irradiance probability density as a more accurate alternative to the Beta distribution, and simpler to implement compared with recent methods in the literature. The GMM estimate can be denoted as a linear combination of Gaussian densities with the careful selection of appropriate mixture components being of paramount importance. The Expectation–Maximization (EM) framework [12], which is an advanced iterative clustering technique, is adopted in this work to evaluate the parameters of the GMM with consideration of a simple yet effective unsupervised information theoretic approach, called the Bayesian Information Criterion (BIC) [13] for correctly determining the number of mixture components in an optimal way.

The performances of the GMM and the popular Beta distribution are evaluated and compared using hourly solar irradiance data (2011–2015) at eight locations in the United States. The two estimation models (the GMM and the Beta distribution) are evaluated using coefficient of determination (R^2), a goodness-of-fit test: the Kolmogorov–Smirnov (K–S) test, and three standard error measures: Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Square Error (RMSE). Results show that the GMM provides more accurate estimations and better performance metrics with higher R^2 indices and substantially lower error values than the Beta distribution.

This work is supported by Khalifa University, Abu Dhabi, United Arab Emirates, under the Advanced Power and Energy Center.

978-1-7281-6489-2/20/$31.00 ©2020 IEEE

The remainder of the paper is organized as follows. Statistical information about the used solar data in the study is presented in Section II. The Beta distribution is descried in Section III, and the proposed GMM is introduced in Section IV. The resulting probability density estimates are presented in Section V together with the results obtained from the statistical test and measures (the K–S test, R^2 index, and the three measures of error). Conclusions are lastly presented in Section VI.

II. Solar Global Horizontal Irradiance Multi-Year Data

Hourly Global Horizontal Irradiance (GHI) solar data was acquired from the National Renewable Energy Laboratory (NREL) [14] at eight sites (solar PV power stations) located in the West of the United States for the five year period 2011–2015. Table I lists statistical information of GHI data and the Global Positioning System coordinates (i.e. latitude and longitude) of solar sites.

A pre-processing technique was implemented to eliminate data corresponding to night-time hours [15], [16] based on the Solar Elevation Angle (SEA). Solar GHI data with SEA $\leq 1^\circ$ are eliminated. The normalized values of the resulting hourly solar GHI multi-year data ($\approx 21.6\times10^3$ points) were computed based on the maximum value at each corresponding site.

III. The Beta Probability Density Function

The Beta distribution is the most common parametric family of probability distributions used to estimate the probability density of Normalized Solar Irradiance (NSI) data. The Probability Density Function (PDF) f_{Beta} [17], [18] is expressed by:

$$f_{Beta}(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\; x^{\alpha-1}\;(1-x)^{\beta-1} \tag{1}$$

where $x \in (0,1)$, Γ is the Gamma function, and $\alpha, \beta \in (0,\infty)$ are the first and second shape parameters, respectively, given by:

$$\begin{aligned} \alpha &= \frac{\gamma\times\beta}{1-\gamma}, \\ \beta &= (1-\gamma)\;\left(\frac{\gamma\,(1-\gamma)}{\sigma^2}-1\right), \end{aligned} \tag{2}$$

where σ and γ are the standard deviation and mean of the random variable, respectively.

IV. Gaussian Mixture Model

The GMM is a finite convex linear combination of Gaussian densities with different parameters of each probability density [19]. The careful selection of a correct number of mixture components is of top importance when fitting a finite mixture distribution. A very large number of mixture components may result in an increase in complexity and over-fitting, while a small number of components yields an inaccurate representation. The EM framework is adopted in this paper to obtain an estimation of the parameters of the GMM PDF, and the BIC approach is considered to determine the number of components in an optimal manner.

Similar to the Gaussian distribution, the individual densities of the GMM retain the "universal-approximation" property, and are adequately characterized by the first two moments [20], [21]. The probability density distribution of the GMM can approximate any arbitrarily shaped non-Gaussian density as proven by the Wiener's approximation theory [19]. Suppose that the observed solar irradiance values are denoted by $\boldsymbol{x} = x_1,\ldots,x_n$, and let the j^{th} entry of the random variable be modeled as a finite convex linear combination of Gaussian densities,

$$f_X(x_j|\theta) = \sum_{i=1}^{C} \omega_i\,\phi(x_j,\theta_i), \quad x_j \geq 0,\, j=1,\ldots,N \tag{3}$$

where C is the number of mixture components, and each i^{th} component is given by

$$\phi\,(x_j,\theta_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left(-\frac{(x_j-\mu_i)^2}{2\sigma_i^2}\right), \tag{4}$$

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

In practice, it is advantageous to compute the parameters θ which maximize the log-likelihood function ($\ln \Pr(\boldsymbol{x}|\theta,C)$), which is, however, analytically intractable [12], [22]. In this paper, the EM framework is used to solve the Maximum Likelihood Estimation (MLE) problem, and ultimately evaluate the GMM PDF parameters [12]. The MLE problem is addressed through two iterative steps to maximize the expected log-likelihood of the data: (1) Expectation Step and (2) Maximization Step. The EM algorithm, however, requires an *a priori* input as the number of mixture components (C), which when increased, the log-likelihood function can be further maximized at the cost of over-fitting and increased model complexity.

To resolve this issue, the BIC approach [13], [23] is adopted in this paper because it adds to the log-likelihood function a penalty term as C is increased, following

$$\text{BIC}(C) = -2\ln\Pr(\hat{\theta}|\boldsymbol{x},\,C) + C\,\ln(n), \tag{5}$$

where $\hat{\theta}$ is the parameter such that the log-likelihood function is maximized. The EM algorithm is used in this paper to maximize the log-likelihood function, along with the BIC approach used to find the appropriate number of mixture components. The algorithm is terminated when the BIC measure does not show any improvement. In a large-sample setting, like the one we have in this study, the number of mixture components computed by the minimum BIC is asymptotically optimal when the Bayesian posterior probability ($\Pr(\theta|\boldsymbol{x},\,C)$)

Table I: Statistical Information of Solar GHI Data

Site	Location	Geographical Coordinates		Solar Irradiance (W/m^2)			
		Latitude	Longitude	Maximum	Mean	Median	Standard Deviation
1	Clark County, Nevada	36.41°N	114.9°W	1089	478.27	464	302.43
2	Boulder City, Nevada	35.89°N	114.98°W	1095	481.18	474	302.30
3	Riverside County, California	33.85°N	115.42°W	1088	504.88	516	304.47
4	San Bernardino County, California	35.57°N	115.42°W	1103	489.95	486	305.51
5	Imperial County, California	32.69°N	115.62°W	1091	503.21	513	305.70
6	Phoenix, Arizona	33.45°N	112.06°W	1101	497.53	505	304.64
7	Sierra County, New Mexico	33.13°N	107.06°W	1130	498.11	497	316.07
8	El Paso, Texas	31.77°N	106.50°W	1126	501.22	495	316.15

is considered [13]. It is important to note that the EM algorithm is guaranteed not to get worse as it iterates [12], while having the merit of being a fully unsupervised learning algorithm producing robust solar irradiance probability density estimates.

V. Results and Performance Evaluations

An assessment of the robustness of the proposed GMM and the common Beta distribution for characterizing the variability of solar data is presented in this section. Performance evaluation of the computed probability density estimates is performed using a goodness-of-fit test (the K–S test), R^2 index, and three statistical measures (MAE, MAPE and RMSE). The proposed GMM routine was coded in MATLAB with the PDF parameters computed using the EM algorithm and the optimal number of mixture components found using the BIC approach. The Beta distribution shape parameters and density estimates were computed in MATLAB using the functions `betafit` and `betapdf` (see Table II).

Fig. 1 shows the normalized histograms (with density scaling) of NSI data at the eight selected sites (obtained using MATLAB's `histogram`), overlaid with the two density estimates. Visually, these plots show substantial discrepancies between the data histogram and the Beta distribution. As depicted, the proposed GMM provides a more accurate estimate of NSI probability density.

The K–S goodness-of-fit test [24] is employed to decide if the observed set of data is independently sampled from a specific statistical distribution by providing a test decision for the null hypothesis, against the alternative that the data does not come from such a distribution. Each K–S test returns a p-value which, informally, is a measure of the evidence against the null hypothesis. When the observed data set does not provide sufficient evidence validating the hypothesis that it is associated with the statistical model, the K–S test rejects the null hypothesis at the α_{K-S} significance level.

Table III lists the p-values produced by the K–S test at the chosen $\alpha_{K-S} = 1\%$ significance level for each of the two models with bold-highlighted values indicating a failure to reject the null hypothesis. For all eight sites, the returned p-values by the K–S test when considering the Beta distribution are lower than α_{K-S} (i.e. the null hypothesis is rejected), illustrating that the Beta distribution does not provide a reliable fit for solar irradiance data.

The proposed BIC-assisted GMM applying the EM algorithm produces p-values which indicate a failure to reject the null hypothesis at the α_{K-S} significance level for all sites, suggesting that the proposed model is a more robust approach for estimating solar GHI data.

Table II: The Beta distribution shape parameters

Site	α	β
1	0.98	1.27
2	0.99	1.29
3	1.01	1.20
4	1.00	1.28
5	1.00	1.21
6	1.00	1.26
7	0.96	1.26
8	0.96	1.24

Table III: p-values of the K–S goodness-of-fit test

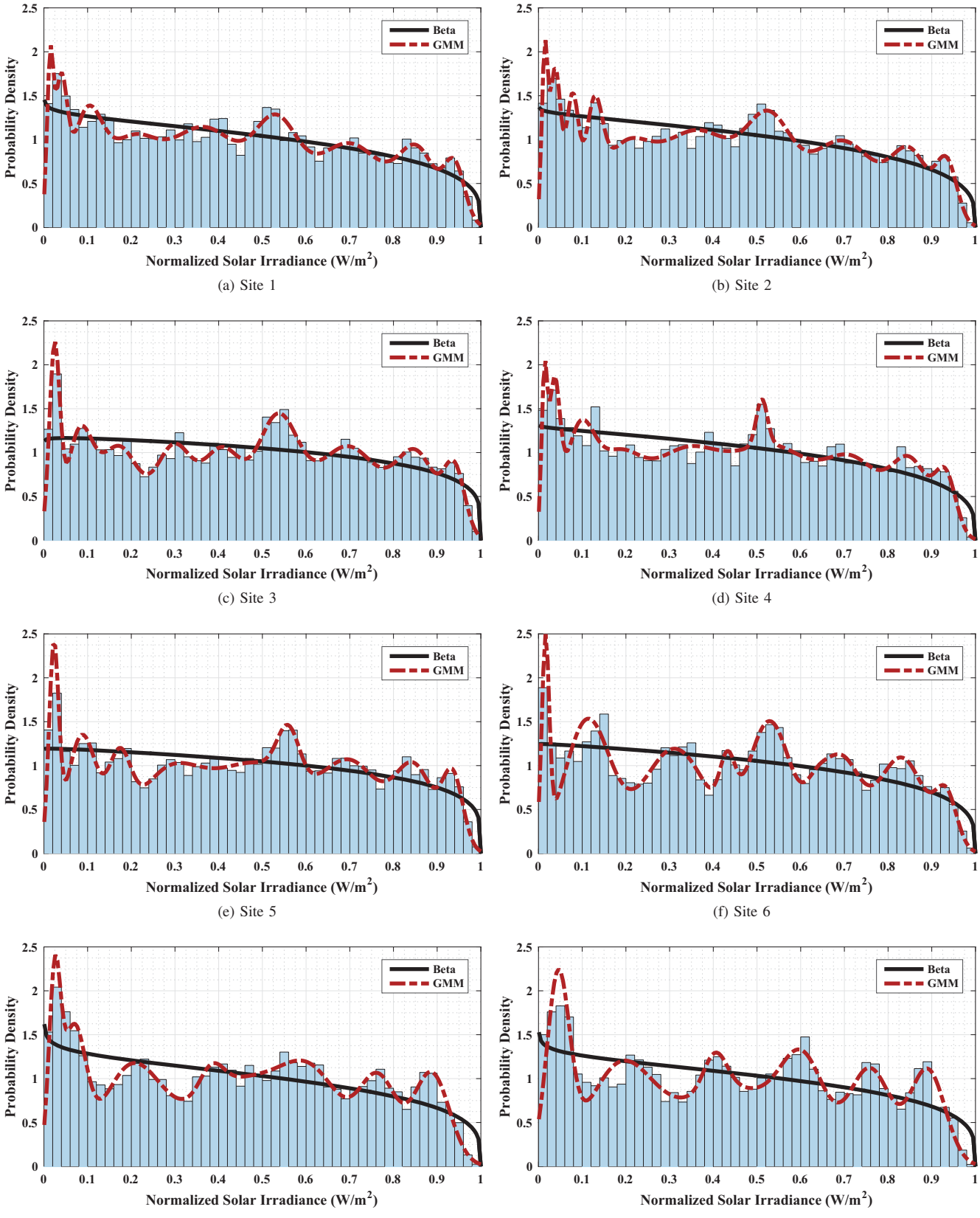
Site	Beta	GMM
1	1.10×10^{-12}	$\mathbf{39.46\times10^{-2}}$
2	2.80×10^{-19}	$\mathbf{23.72\times10^{-2}}$
3	1.15×10^{-31}	$\mathbf{29.05\times10^{-2}}$
4	4.29×10^{-24}	$\mathbf{15.80\times10^{-2}}$
5	3.98×10^{-30}	$\mathbf{34.32\times10^{-2}}$
6	2.15×10^{-36}	$\mathbf{4.41\times10^{-2}}$
7	4.83×10^{-32}	$\mathbf{33.59\times10^{-2}}$
8	9.93×10^{-30}	$\mathbf{18.52\times10^{-2}}$

In addition to the goodness-of-fit test, the performance of the proposed GMM PDF is evaluated against f_{Beta} using the R^2 index, and three standard measures: MAE, MAPE and RMSE given by [25]:

$$R^2 = 1 - \frac{\sum_{i=1}^{t}(y_i - \hat{y_i})^2}{\sum_{i=1}^{t}(y_i - \bar{y})^2}, \tag{6}$$

$$MAE = \frac{1}{t}\sum_{i=1}^{t}\left|y_i - \hat{y_i}\right|, \tag{7}$$

$$MAPE = \frac{1}{t}\sum_{i=1}^{t}\left|\frac{y_i - \hat{y_i}}{y_i}\right|, \tag{8}$$



(a) Site 1

(b) Site 2

(c) Site 3

(d) Site 4

(e) Site 5

(f) Site 6

(g) Site 7

(h) Site 8

Figure 1: Probability density plots and histograms.

$$RMSE = \sqrt{\frac{1}{t}\sum_{i=1}^{t}(y_i - \hat{y_i})^2}, \quad (9)$$

where t is the number of data bins chosen using the square root rule, y_i is the probability of solar GHI data being within bin i calculated from the data set, $\hat{y_i}$ is the probability within the same bin calculated from the estimated data set, $\bar{y} = \frac{1}{t}\sum_{i=1}^{t} y_i$, and $i = 1, \ldots, t$.

Table IV–VII list the computed R^2, MAE, MAPE and RMSE values with error percentage improvements computed with respect to the Beta distribution for all sites. The proposed GMM produces the highest R^2 indices and the lowest error values for all eight sites at the expense of additional computational and mathematical sophistication, with MAE percentage improvements between 24.1%–42.4% (at an average of 33.1%), MAPE percentage improvements between 44.5%–73.5% (at an average of 64.0%), and RMSE percentage improvements between 22.2%–37.7% (at an average of 30.0%) with respect to the Beta distribution.

The proposed GMM produces the best performance metrics and the highest percentage improvements providing a more accurate estimation of solar irradiance compared with the popular Beta distribution.

Table IV: R^2 values

Site	Beta	GMM
1	0.5698	0.7396
2	0.5041	0.7571
3	0.3029	0.6962
4	0.4565	0.6888
5	0.3136	0.6913
6	0.3418	0.6183
7	0.4374	0.7618
8	0.3707	0.7558

Table V: MAE with % improvements

Site	Beta	GMM
1	8.00×10^{-4}	6.08×10^{-4} (24.1%)
2	9.43×10^{-4}	6.17×10^{-4} (34.5%)
3	9.55×10^{-4}	6.39×10^{-4} (33.1%)
4	9.74×10^{-4}	6.94×10^{-4} (28.7%)
5	9.88×10^{-4}	6.51×10^{-4} (34.1%)
6	11.10×10^{-4}	7.86×10^{-4} (29.2%)
7	10.80×10^{-4}	6.66×10^{-4} (38.3%)
8	11.91×10^{-4}	6.86×10^{-4} (42.4%)

VI. Conclusion

A Gaussian Mixture Model (GMM) is proposed in this paper for estimating solar irradiance probability density at a given site. The proposed GMM parameters are obtained using the Expectation–Maximization (EM) algorithm, and the optimal number of mixture components is determined using the Bayesian Information Criterion (BIC).

Table VI: MAPE with % improvements

Site	Beta	GMM
1	0.2474	0.1373 (44.5%)
2	0.3535	0.1422 (59.8%)
3	0.3585	0.1500 (58.2%)
4	0.6101	0.1732 (71.6%)
5	0.5850	0.1767 (69.8%)
6	0.4594	0.1733 (62.3%)
7	0.6795	0.1802 (73.5%)
8	0.7366	0.2056 (72.1%)

Table VII: RMSE with % improvements

Site	Beta	GMM
1	10.41×10^{-4}	8.10×10^{-4} (22.2%)
2	11.97×10^{-4}	8.38×10^{-4} (30.0%)
3	12.90×10^{-4}	8.51×10^{-4} (34.0%)
4	12.45×10^{-4}	9.42×10^{-4} (24.3%)
5	13.19×10^{-4}	8.84×10^{-4} (32.9%)
6	14.52×10^{-4}	11.05×10^{-4} (23.8%)
7	13.93×10^{-4}	9.07×10^{-4} (34.9%)
8	14.98×10^{-4}	9.33×10^{-4} (37.7%)

The performance of the proposed BIC-assisted EM-based GMM is assessed against the popular parametric Beta distribution using multi-year solar data at eight sites located in the United States. Assessments are performed using a goodness-of-fit test (the K–S test), R^2 index, and three standard statistical error measures (MAE, MAPE and RMSE). Results confirm the robustness of the GMM in estimating solar irradiance probability density with higher R^2 indices and lower error values than the Beta distribution with percentage improvements of up to 42.4% (MAE), 73.5% (MAPE), and 37.7% (RMSE). When conducting the goodness-of-fit test, the common Beta distribution consistently produced p-values that do not sufficiently support the null hypothesis, i.e. it produces poor estimates of solar GHI probability density. The GMM, on the other hand, consistently returns p-values indicating a failure to reject the null hypothesis, implying that it is a more adequate and robust model for estimating solar GHI probability density.

References

[1] Z. Ren, W. Yan, X. Zhao, W. Li, and J. Yu, "Chronological probability model of photovoltaic generation," *IEEE Trans. Power Syst.*, vol. 29, no. 3, pp. 1077–1088, May 2014.

[2] G. Mokryani, "Active distribution networks planning with integration of demand response," *Sol. Energy*, vol. 122, pp. 1362–1370, Dec. 2015.

[3] A. Soroudi, M. Aien, and M. Ehsan, "A probabilistic modeling of photo voltaic modules and wind power generation impact on distribution networks," *IEEE Syst. J.*, vol. 6, no. 2, pp. 254–259, Jun. 2012.

[4] F. S. Gazijahani and J. Salehi, "Stochastic multi-objective framework for optimal dynamic planning of interconnected microgrids," *IET Renew. Power Gener.*, vol. 11, no. 14, pp. 1749–1759, Dec. 2017.

[5] 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 Trans. Power Syst.*, vol. 28, no. 1, pp. 326–334, Feb. 2013.

[6] R. A. Osama, A. F. Zobaa, and A. Y. Abdelaziz, "A planning framework for optimal partitioning of distribution networks into microgrids," *IEEE Syst. J.*, vol. 14, no. 1, pp. 916–926, Mar. 2020.

[7] N. M. Nor, A. Ali, T. Ibrahim, and M. F. Romlie, "Battery storage for the utility-scale distributed photovoltaic generations," *IEEE Access*, vol. 6, pp. 1137–1154, Nov. 2018.

[8] S. Paul and N. P. Padhy, "Resilient scheduling portfolio of residential devices and plug-in electric vehicle by minimizing conditional value at risk," *IEEE Trans. Ind. Informat.*, vol. 15, no. 3, pp. 1566–1578, Mar. 2019.

[9] M. Shafie-Khah and P. Siano, "A stochastic home energy management system considering satisfaction cost and response fatigue," *IEEE Trans. Ind. Informat.*, vol. 14, no. 2, pp. 629–638, Feb. 2018.

[10] M. Wahbah, S. Feng, T. H. M. EL-Fouly, and B. Zahawi, "Root-transformed local linear regression for solar irradiance probability density estimation," *IEEE Trans. Power Syst.*, vol. 35, no. 1, pp. 652–661, Jan. 2020.

[11] M. Wahbah, T. H. M. EL-Fouly, B. Zahawi, and S. Feng, "Hybrid Beta-KDE model for solar irradiance probability density estimation," *IEEE Trans. Sustain. Energy*, vol. 11, no. 2, pp. 1110–1113, Apr. 2020.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stat. Soc. Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, Sep. 1977.

[13] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, Mar. 1978.

[14] National Renewable Energy Laboratory (NREL). National Solar Radiation Database (NSRDB). U.S. Department of Energy (DOE)/NREL/ALLIANCE. Washington, DC, USA. [Online]. Available: https://nsrdb.nrel.gov/

[15] C. Voyant, T. Soubdhan, P. Lauret, M. David, and M. Muselli, "Statistical parameters as a means to a priori assess the accuracy of solar forecasting models," *Energy*, vol. 90, pp. 671–679, Oct. 2015.

[16] J. R. Trapero, "Calculation of solar irradiation prediction intervals combining volatility and kernel density estimates," *Energy*, vol. 114, pp. 266–274, Nov. 2016.

[17] M. Brabec, E. Pelikán, P. Krč, K. Eben, and P. Musilek, "Statistical modeling of energy production by photovoltaic farms," in *2010 IEEE Electrical Power and Energy Conference (EPEC)*, Halifax, NS, Canada, Aug. 25–27, 2010, pp. 1–6.

[18] F. Youcef Ettoumi, A. Mefti, A. Adane, and M. Bouroubi, "Statistical analysis of solar measurements in Algeria using beta distributions," *Renewable Energy*, vol. 26, no. 1, pp. 47–67, May 2002.

[19] K. N. Plataniotis and D. Hatzinakos, *Gaussian Mixtures and their Applications to Signal Processing*, 1st ed., ser. Electrical Engineering & Applied Signal Processing Series. Boca Raton, Florida: CRC Press, Dec. 2000, ch. 3, pp. 1–32.

[20] E. A. Patrick, *Fundamentals of Pattern Recognition*, 1st ed., ser. Prentice-Hall Information and System Sciences Series. Prentice-Hall, 1972.

[21] A. Prochazka, N. Kingsbury, P. Payner, and J. Uhlir, *Signal Analysis and Prediction*, 1st ed., ser. Applied and Numerical Harmonic Analysis. Boston, MA: Birkhäuser Basel, 1998.

[22] 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 Trans. Veh. Technol.*, vol. 65, no. 10, pp. 8309–8321, Oct. 2016.

[23] O. Alhussein, I. Ahmed, J. Liang, and S. Muhaidat, "Unified analysis of diversity reception in the presence of impulsive noise," *IEEE Trans. Veh. Technol.*, vol. 66, no. 2, pp. 1408–1417, Feb. 2017.

[24] M. H. DeGroot and M. J. Schervish, *Probability and Statistics: Pearson New International Edition*, 4th ed. Harlow, UK: Pearson Education Limited, Jul. 2013.

[25] J. Kleissl, *Solar Energy Forecasting and Resource Assessment*, 1st ed. Boston, MA, USA: Academic Press, Jul. 2013.