# Study Project

# Math F266

# Exploration of Statistical Distributions for Energy Data Analysis

Final Report


Submitted by

Karthik Periasamy – 2022B4AA0621P

Submitted to

Dr. Sumanta Pasari, Associate Professor




Birla Institute of Technology & Science, Pilani

Vidyavihar, Pilani, Rajasthan – 333031

# 1. Introduction

The global energy landscape is witnessing a significant shift towards renewable energy sources, driven by growing environmental concerns, technological advancements, and supportive government policies. As of 2023, the global electricity production is 28,884 TWh. Wind and Solar power make up 7% and 5% of this figure respectively. Over recent years, there has been a marked increase in the adoption of renewables such as solar and wind. As a result, renewable energy is not only seen as a solution for reducing carbon emissions but also as a vital component of energy security and job creation in a new green economy. Moreover, the fall in prices of solar cells and wind turbines due to a reduce in production costs and subsidies provided by various governments. This paradigm shift is fostering a more sustainable energy future, with renewables expected to make up a significant portion of global energy generation in the coming decades. [1] [2] [3]

To assess the potential for a wind or solar farm at a certain location, it is important to estimate the yearly power output at that location. For wind speed the metric that is used is the wind power density (P). It depends on the spatiotemporal variability of wind speed (x) and air density (ρ).

$$P = \frac{1}{2}\rho x^3$$

Since P is proportional to the cube of the wind speed x, a small change in wind speed leads to a large change in power density. Therefore, we must have the probability density function (f(x)) of wind speed to accurately estimate the power output. To optimize the use of wind energy, the accurate estimation of wind resources available in a certain area is necessary. Wind turbine power output ($P_W$) is an estimation of the power as a function of wind speed x. This relation can be interpreted through power curves, which relate wind speed and power output and are wind turbine-specific. All power curves have the following characteristics

i) Cut-in speed: the speed at which the wind turbine starts to generate usable power
ii) Rated output speed: the speed at which the maximum power output is generated
iii) Cut-out speed: the speed at which the wind turbine shuts down to prevent damage

The average wind turbine power output $\overline{P_W}$ can be computed using the power curves by –

$$\overline{P_W} = \int_0^\infty P_w(x) f(x) dx$$

Similarly, to assess the energy potential for solar energy we use the metric of Global Horizontal Irradiance (GHI). This is a measure of irradiance, and it measures the power per unit area ($W/m^2$) received by sun light. The GHI at a particular location is measured using a pyranometer. The relationship between the solar power output ($P_S$) and GHI is linear and power output increases with increase in GHI. Unlike wind speed, which may have cut-out values where turbines stop operating due to excessively high wind speeds for safety reasons, solar cells don't have such limits in their power curves. This means that solar panels can continue to generate power as long as there is sunlight falling on them, without reaching a point where they shut down due to extreme conditions. GHI consists of the following components –

i) Direct Normal Irradiance (DNI): Solar radiation received from the sun without having been scattered by the atmosphere.
ii) Diffuse Horizontal Irradiance (DHI): Solar radiation that has been scattered by the atmosphere.

GHI is related to DNI and DHI by the following formula -

$$GHI = DHI + DNI \cos(z) \text{ [3]}$$

Dealing with GHI data, however, is somewhat more complex than dealing with wind speed data as it varies far more with time-of-day and season. Predicting GHI trends at a certain point in time require machine learning (ML) techniques and are out of the domain of this report. We will focus on fitting the probability density of GHI values over a large data set such that we can reliably predict average power output. [5]

The determination of f(x) has a large impact on the determination of power output. The process of fitting wind speed or GHI data to a suitable distribution is three-fold. It involves –

i) Distributions
ii) Estimation of parameters
iii) Goodness-of-fit

We will delve more into this process in the methodology section of the report.

# 2. Literature Review

## 2.1 Wind Speed

The pre-existing literature landscape in the area of statistical distributions for wind speed analysis is quite vast. The main change over the years has been the distribution selected to fit the wind speed data. By far the most common combination of distribution and parameter estimation method used has been the two-parameter Weibull distribution, defined by its scale ($\alpha$) and shape (k), which are most frequently estimated using MLE. [6] [7] The MLE method is an optimizing process that estimates the parameters in such a way that the likelihood function (or log-likelihood function) is maximized. More recent research however uses several modifications and alternatives to the Weibull distribution to model wind speed data.

Wind speed data can contain multiple peaks and therefore, multi-modal distributions are also used to model wind speed data. These distributions consists of a linear combination of two distributions with an additional mixing parameter $\omega$. W-W is the most commonly used multi-modal distribution to estimate data. Other multi-modal distributions are B-GEV, N-N, GEV-W etc. [8]

## 2.2 GHI

Global Horizontal Irradiance (GHI) analysis involves examining a wide range of research outputs and methodologies that are pivotal for optimizing solar energy utilization. GHI, which measures the total solar radiation received per unit surface area by a surface horizontal to the ground, plays a critical role in the design and efficiency optimization of photovoltaic

(PV) systems. An important area of research in GHI analysis involves the statistical modelling of irradiance data. Researchers have applied various statistical distributions to fit GHI data, such as the beta distribution or the gamma distribution, which help in understanding the variability and predictability of solar irradiance. There has been research that has explored these distributions to model the clearness index, which is closely related to GHI. The fitting of these distributions aids in the development of probabilistic models that can forecast solar irradiance with greater accuracy. [9]

As ML evolved, more sophisticated models such as artificial neural networks (ANNs) and deep learning models began to dominate research. Deep learning models, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have been explored more recently, which demonstrated their superior capability in capturing temporal and spatial dependencies in solar radiation data. To further enhance the prediction accuracy, researchers have developed hybrid models that combine ML techniques with traditional statistical methods. However, there is little recent research relating to the application of statistical distributions to estimate average power output. In this report we expand the techniques used commonly in the domain of wind energy to solar GHI data to try and emulate similar results. [10] [4]

## 3. Methodology

Wind speed and GHI data were evaluated from four locations across India, namely, Bhopal, Hamirpur, Jafrabad and Thiruvananthapuram. The measurement of wind speed was done at 50 metres of height. The data consists of hourly wind speed readings from August 2015 to March 2023. This data was plotted into a histogram on Python and then fit into a few distributions available on the scipy.stats library. The GHI data has been collected from these locations between December 2012 and December 2022. Only GHI data between 8:00 AM and 4:00 PM was considered (GHI data apart from this is either 0 or nearing 0). The pre-processing of data also includes the exclusion of null values in the dataset. The primary difference between the treatment of GHI and wind speed data is the underlying distribution used. Wind speed data can often be accurately predicted with unimodal distributions. GHI data however often contains multiple peaks, therefore, we implement multimodal mixture distributions. These distributions consist of the sum of various unimodal distributions and are

each associated with a weight ($w_i$) such that the sum of these weights is equal to one. A similar approach is used in the treatment of both the datasets.

### 3.1 Distributions

The most commonly used distribution to fit wind speed data has been the two-parameter Weibull distribution. More recently, several more elaborate distributions have been put to use, many of which have more than 2 parameters. Many of these distributions arise as modifications to the exponential distribution, they are right-skewed and are found to model wind speed data quite well. The pdf of the Weibull the related exponentiated Weibull distributions respectively are [8] -

$$f_{W3}(x;\alpha,k,\mu) = \frac{k}{\alpha}\left(\frac{x-\mu}{\alpha}\right)^{k-1} e^{-\left(\frac{x-\mu}{\alpha}\right)^{k}}$$

$$f_{EW}(x;\alpha,k,\mu,o) = o\frac{k}{\alpha}\left(\frac{x-\mu}{\alpha}\right)^{k-1}\left(1-e^{-\left(\frac{x-\mu}{\alpha}\right)^{k}}\right)^{o-1} e^{-\left(\frac{x-\mu}{\alpha}\right)^{k}}$$

$\alpha$: first scale parameter

$\mu$: location parameter

k: first shape parameter

o: second shape parameter

Both these distributions are primarily used in the analysis of wind speed data. For the GHI data, we employ a mixture distribution. A mixture distribution has the general form –

$$f(x) = \sum_{i=1}^{n} w_i f_i$$

$w_i$: weights

$f_i$: probability density functions

such that $\sum_{i=1}^{n} w_i = 1$.

One such example is the Weibull mixture model with the pdf –

$$f_{w-w} = w f_w(x;\alpha_1,k_1) + (1-w) f_w(x;\alpha_1,k_1)$$

## 3.2 Estimation of Parameters

Once a distribution is chosen, the parameters of the distribution must be estimated in such a way that the theoretical pdf provides a close fit to the empirical data. There are various methods to estimate these parameters. The most common method used to estimate parameters is the maximum likelihood estimator (MLE). Apart from this, the method of moments (MOM) estimator, the method of l-moments (LMOM), the modified method of moments (MMOM) and the least squares estimator (LSE) methods are used to estimate the parameters of certain distributions. Therefore whenever we describe a fit to the given data, we mention both the distribution and the method of estimation of parameters.

## 3.3 Goodness-of-fit metrics

Once a pdf has been fit to the empirical wind data, we need a way to evaluate and compare the fit we have obtained. Certain goodness-of-fit metrics are used to compare the theoretical estimate with the empirical data. Some of these metrics are listed below –

**Table I:** A few commonly used Goodness-of-fit metrics

| Goodness-of-fit metric | Equation |
|---|---|
| Kolmogorov-Smirnov statistic | $KS = \max\lvert F_i - \widehat{F}_i \rvert$ |
| AIC | $AIC = -2\ln\left(\prod_{i=1}^{n} f(x_i)\right) + 2NP$ |
| BIC | $BIC = -2\ln\left(\prod_{i=1}^{n} f(x_i)\right) + NPln(n)$ |

$F_i$: ith empirical cumulative distribution function (ecdf) value

$\widehat{F}_i$: estimated cdf of the ith ecdf value

$f$: estimated pdf

ds$NP$: number of parameters

$n$: sample size

A challenge in comparing studies is that they evaluate distributions based on different goodness-of-fit metrics. In most cases, the probability plot is used. The empirical cdf and the cdf from the probability plot are then compared using different statistics to evaluate the
goodness of fit. The Kolmogorov-Smirnov (KS) statistic calculates the absolute maximum difference between the ecdf and cdf in the probability plot. The Bayesian information criterion (BIC) and Akaike information criterion (AIC) are two other metrics used. The advantage of these metrics is that they prevent overfitting of data by adding penalty terms for the number of parameters in the model. Overfitting is the problem where the estimate fits the empirical data too closely and may fail to fit additional data or predict future observations reliably. Both BIC and AIC aim to resolve the trade-off between fit and complexity but do so in slightly different ways. The penalty for additional parameters is less stringent in AIC compared to BIC.

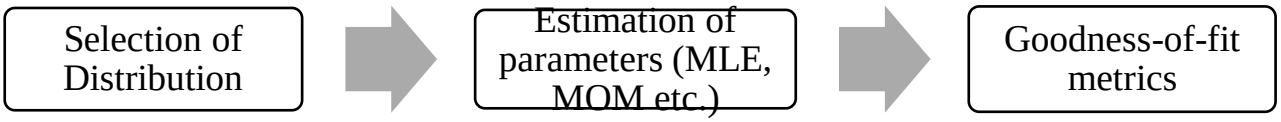


**Figure 1.** Summary of methodology

# 4. Results

## 4.1 Wind Speed

a. Bhopal

| $\text{Distribution}_{\text{method}}$ | AIC | BIC | KS-statistic |
|---|---|---|---|
| $W3_{MLE}$ | 297530 | 297548 | 0.020 |
| $W3_{MOM}$ | 297730 | 297745 | 0.014 |
| $L3_{MLE}$ | 298989 | 299007 | 0.025 |
| $L3_{MOM}$ | 299009 | 299027 | 0.023 |
| $B_{MLE}$ | 297456 | 297484 | 0.017 |
| $GEV_{MLE}$ | 298834 | 298861 | 0.024 |
| $GG_{MLE}$ | 297276 | 297303 | 0.013 |
| $GG_{MOM}$ | 297280 | 297307 | 0.012 |
| $EW_{MLE}$ | 297291 | 297318 | 0.013 |
| $EW_{MOM}$ | 297296 | 297324 | 0.012 |

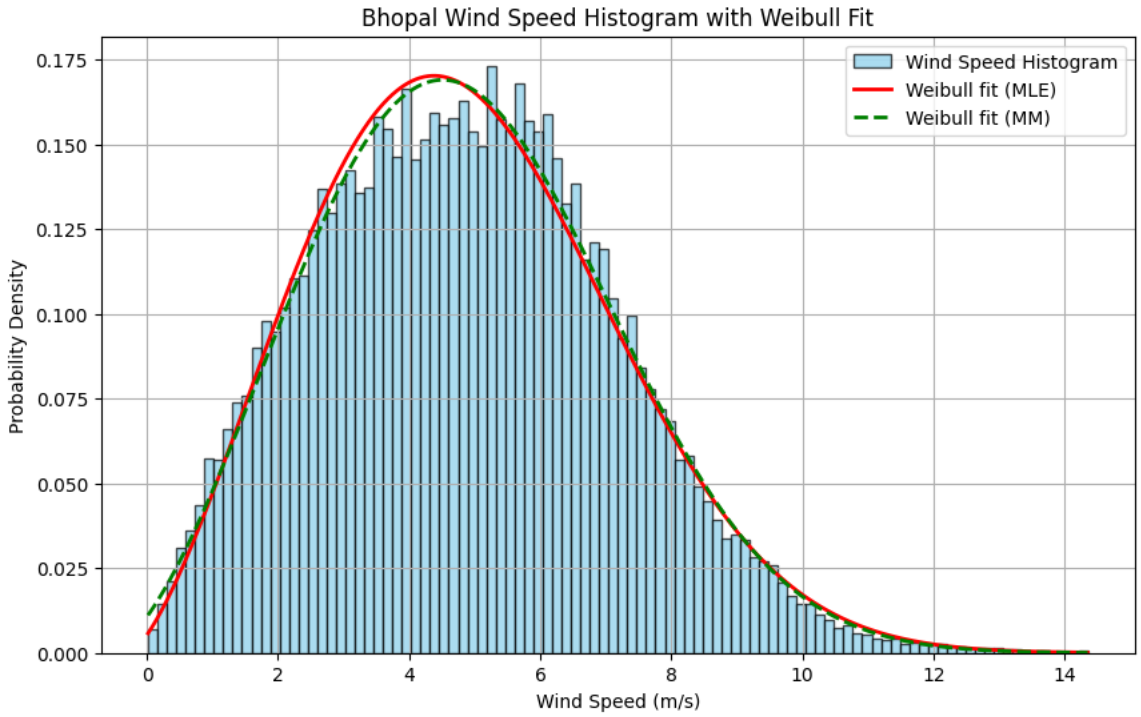


**Figure 2.** Bhopal wind speed data with Weibull fit using MOM and MLE

b. Hamirpur

| **Distribution$_{method}$** | **AIC** | **BIC** | **KS-statistic** |
|---|---|---|---|
| $W3_{MLE}$ | 256299 | 256317 | 0.017 |
| $W3_{MOM}$ | 255738 | 255812 | 0.021 |
| $L3_{MLE}$ | 255797 | 255815 | 0.009 |
| $L3_{MOM}$ | 255810 | 255819 | 0.007 |
| $B_{MLE}$ | 256062 | 256090 | 0.016 |
| $GEV_{MLE}$ | 255956 | 255983 | 0.010 |
| $GG_{MLE}$ | 255594 | 255622 | 0.008 |
| $GG_{MOM}$ | 258433 | 258461 | 0.025 |
| $EW_{MLE}$ | 255578 | 255605 | 0.007 |
| $EW_{MOM}$ | 256007 | 256034 | 0.011 |

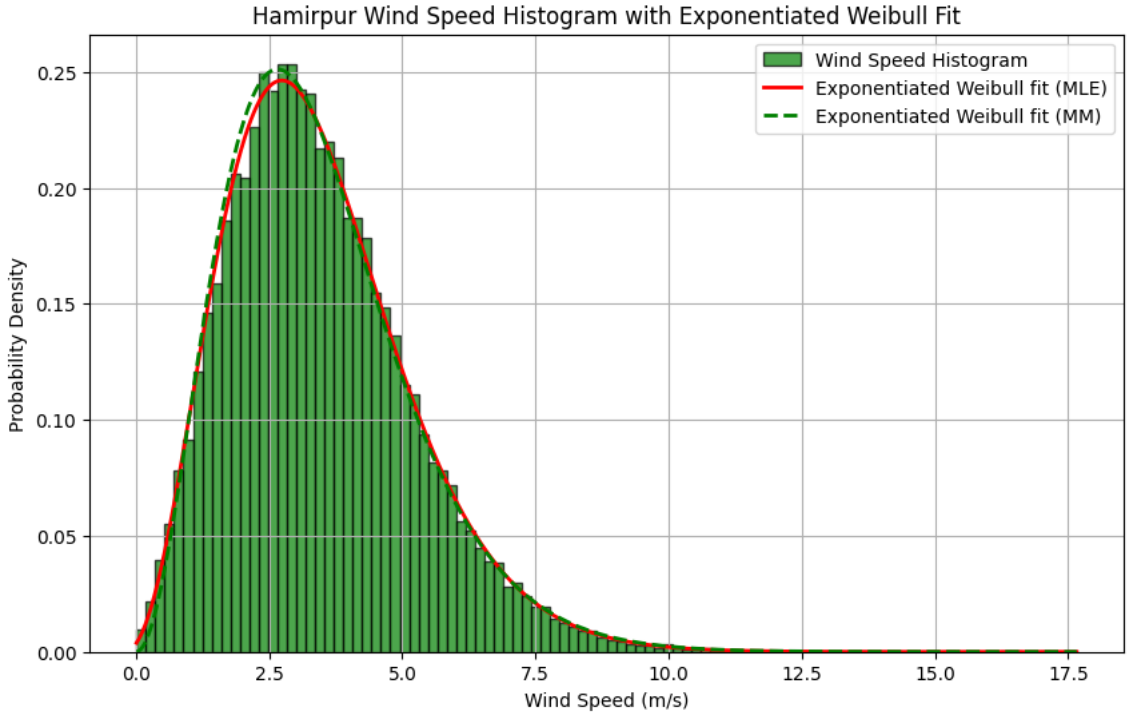


**Figure 3.** Hamiprur wind speed data with Exponentiated Weibull Fit using MOM and MLE

c. Jafrabad

| **Distribution$_{method}$** | **AIC** | **BIC** | **KS-statistic** |
|---|---|---|---|
| $W3_{MLE}$ | 308629 | 308647 | 0.031 |
| $W3_{MOM}$ | 308650 | 308668 | 0.027 |
| $L3_{MLE}$ | 308917 | 308936 | 0.028 |

| | | | |
|---|---|---|---|
| $L3_{MOM}$ | 310965 | 310983 | 0.048 |
| $B_{MLE}$ | 306538 | 306566 | 0.012 |
| $GEV_{MLE}$ | 311612 | 311639 | 0.055 |
| $GG_{MLE}$ | 308689 | 308716 | 0.034 |
| $GG_{MOM}$ | 308906 | 308933 | 0.031 |
| $EW_{MLE}$ | 312322 | 312350 | 0.062 |
| $EW_{MOM}$ | 326616 | 326644 | 0.071 |

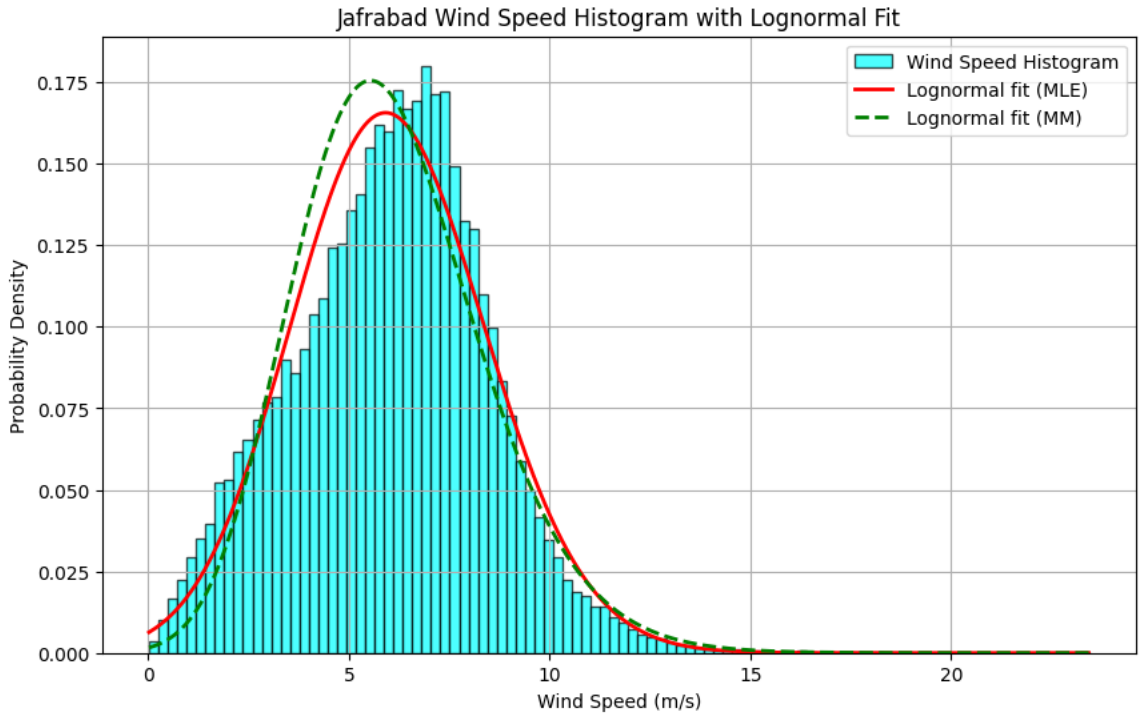


**Figure 4.** Hamirpur wind speed with MOM and MLE fit

d. Thiruvananthapuram

| **$Distribution_{method}$** | **AIC** | **BIC** | **KS-statistic** |
|---|---|---|---|
| $W3_{MLE}$ | 294623 | 294641 | 0.026 |
| $W3_{MOM}$ | 294849 | 294868 | 0.018 |
| $L3_{MLE}$ | 295882 | 295900 | 0.024 |
| $L3_{MOM}$ | 296413 | 296431 | 0.037 |
| $B_{MLE}$ | 293686 | 293713 | 0.007 |
| $GEV_{MLE}$ | 296449 | 296476 | 0.039 |
| $GG_{MLE}$ | 294056 | 294083 | 0.021 |

| | | | |
|---|---|---|---|
| $GG_{MOM}$ | 306974 | 307001 | 0.061 |
| $EW_{MLE}$ | 298162 | 298189 | 0.059 |
| $EW_{MOM}$ | 294106 | 294134 | 0.017 |

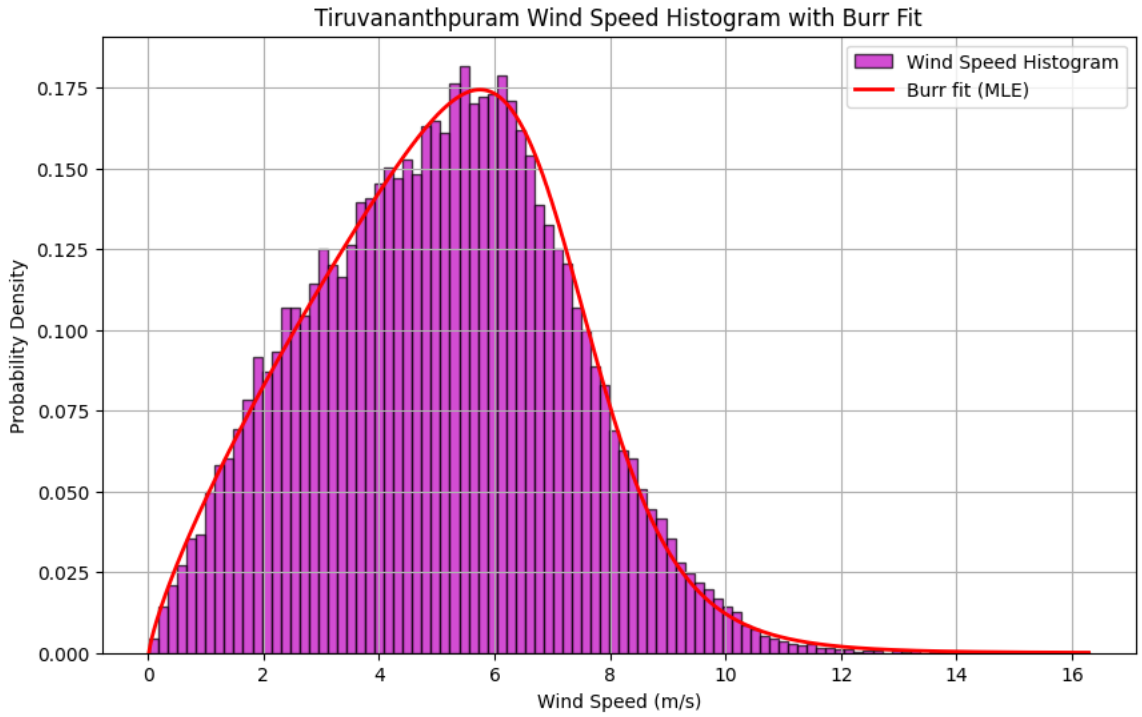


**Figure 5.** Thiruvananthapuram wind speed with Burr fit using MLE

## 4.2 GHI

a. Bhopal

| **$Distribution_{method}$** | **AIC** | **BIC** | **KS-statistic** |
|---|---|---|---|
| Weibull mixture | 450685 | 450718 | 0.012 |
| Weibull 2 | 454582 | 454589 | 0.061 |
| Weibull 3 | 454584 | 454609 | 0.061 |
| Gumbel 2 | 456392 | 456408 | 0.060 |
| Gamma 3 | 458894 | 458919 | 0.095 |
| Lognormal 3 | 466581 | 466607 | 0.124 |

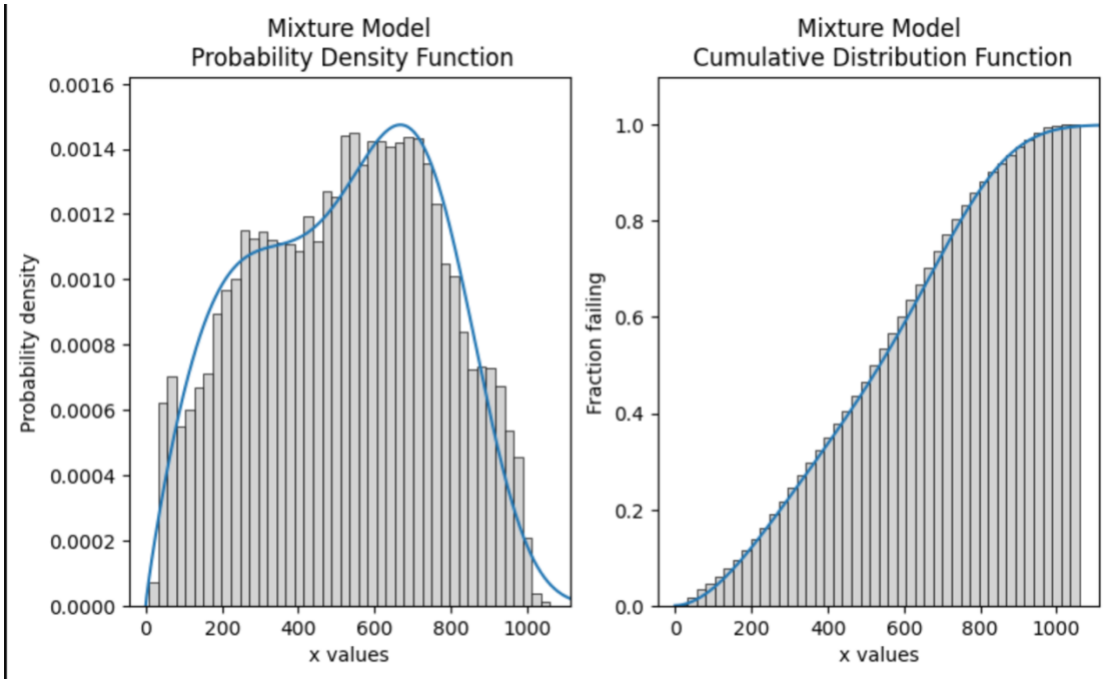


**Figure 6.** Bhopal GHI fit using Weibull mixture model

b. Hamirpur

| **Distribution$_{method}$** | **AIC** | **BIC** | **KS-statistic** |
|---|---|---|---|
| Weibull mixture | 451013 | 451055 | 0.011 |
| Weibull 2 | 454732 | 454749 | 0.061 |
| Weibull 3 | 454734 | 454760 | 0.061 |
| Gumbel 2 | 456294 | 456311 | 0.061 |
| Gamma 3 | 459375 | 459401 | 0.095 |
| Lognormal 3 | 468598 | 468623 | 0.126 |

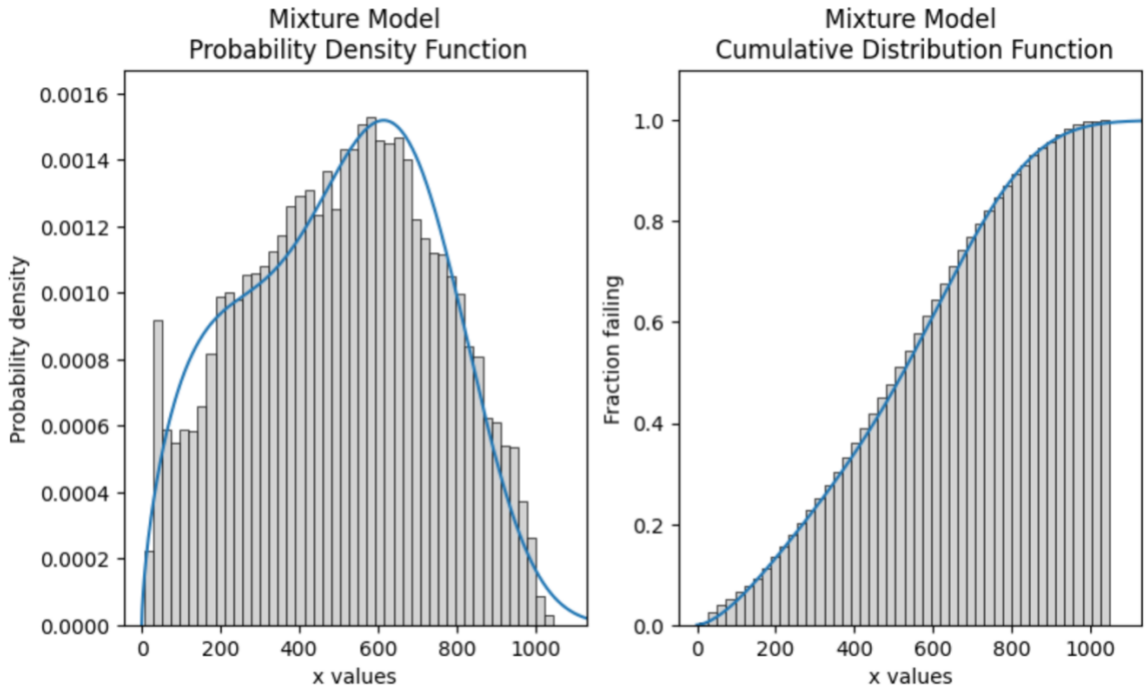


**Figure 7.** Hamirpur GHI fit using Weibull mixture model

c. Jafrabad

| **Distribution**$_{\text{method}}$ | **AIC** | **BIC** | **KS-statistic** |
|---|---|---|---|
| Weibull mixture | 444875 | 444917 | 0.012 |
| Weibull 2 | 448632 | 448649 | 0.061 |
| Weibull 3 | 448634 | 448660 | 0.061 |
| Gumbel 2 | 449491 | 449507 | 0.060 |
| Gamma 3 | 453179 | 453204 | 0.092 |
| Lognormal 3 | 458979 | 459004 | 0.113 |

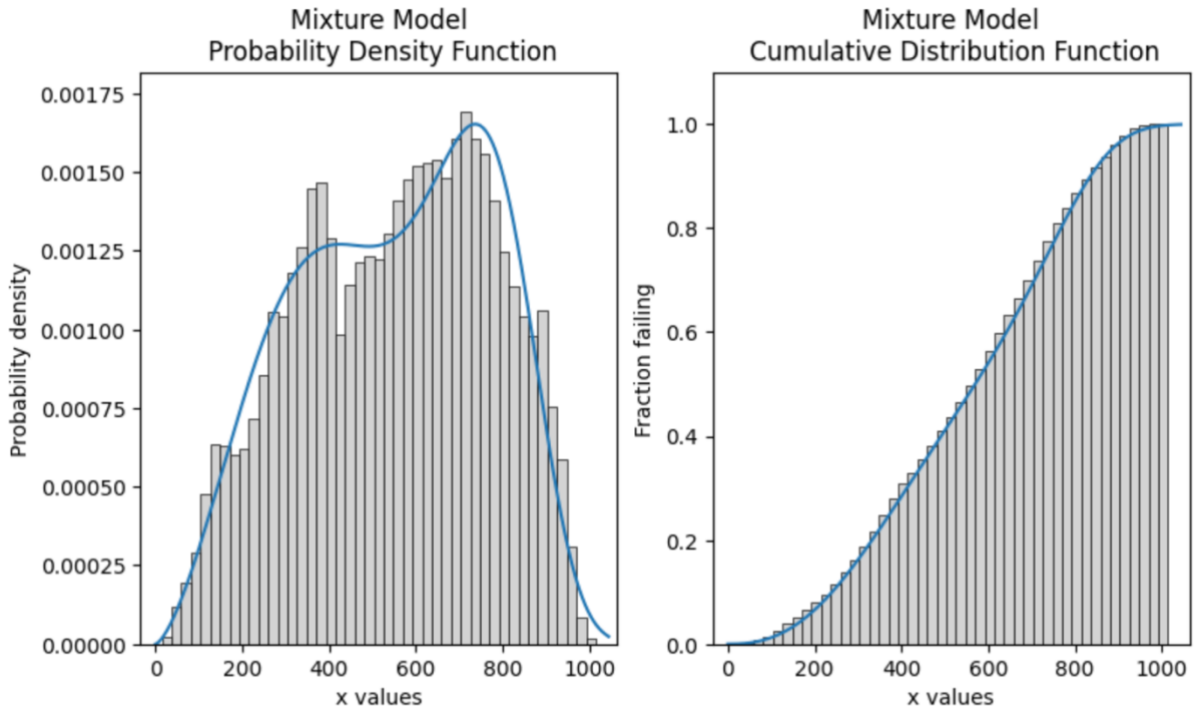


**Figure 8.** Jafrabad GHI fit using Weibull mixture model

d. Thiruvananthapuram

| **Distribution**$_{\text{method}}$ | **AIC** | **BIC** | **KS-statistic** |
|---|---|---|---|
| Weibull mixture | 445438 | 445480 | 0.022 |
| Weibull 2 | 450509 | 450526 | 0.071 |
| Weibull 3 | 450511 | 450536 | 0.071 |
| Gumbel 2 | 450182 | 450199 | 0.057 |
| Gamma 3 | 455526 | 455551 | 0.094 |
| Lognormal 3 | 461875 | 459004 | 0.112 |

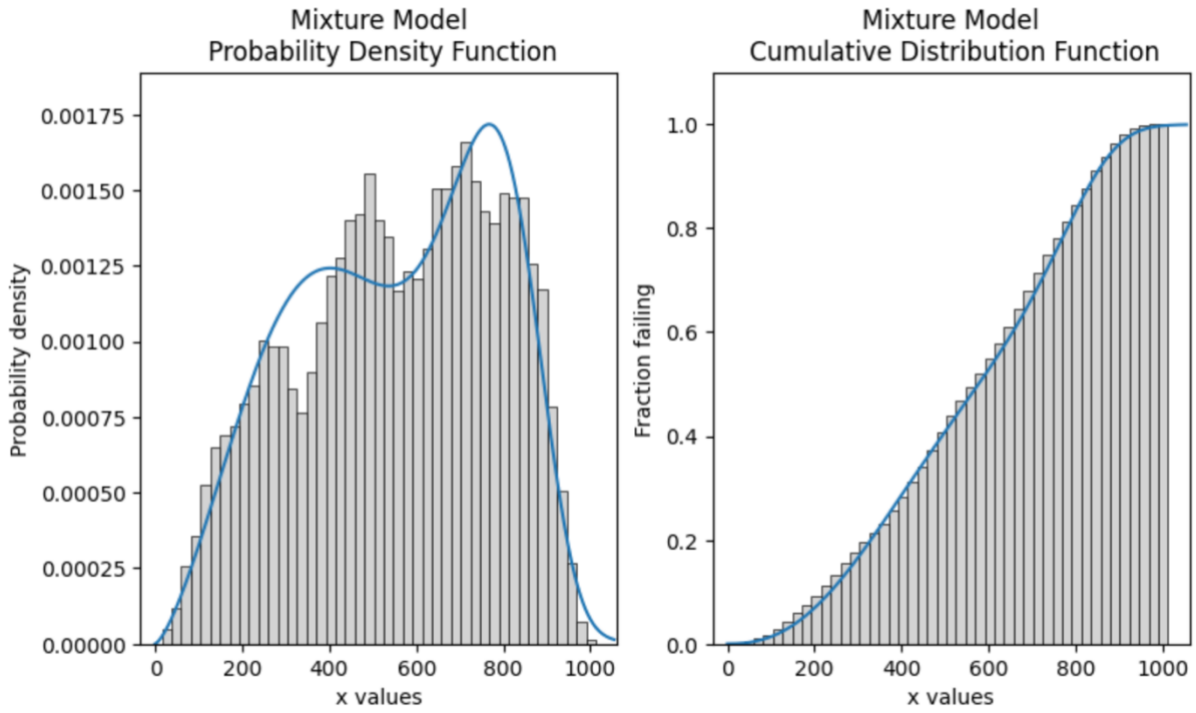


**Figure 9.** Thiruvananthapuram GHI fit using Weibull mixture model

## 5. Discussion and Conclusion

The comprehensive statistical analysis undertaken to model wind speed and Global Horizontal Irradiance (GHI) data for various locations across India has provided significant insights into the suitability of different distributions in capturing the environmental variables relevant to renewable energy projects. The fitting of different distributions and the assessment of their effectiveness using goodness-of-fit metrics such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Kolmogorov-Smirnov (KS) statistic has facilitated a deeper understanding of regional variability in wind and solar energy potential.

The study explored several distributions for modelling wind speeds, such as the Weibull, Lognormal, Generalized Extreme Value (GEV), and Exponentiated Weibull (EW). These distributions were chosen based on their ability to capture the unique characteristics of wind speed data, which is critical for predicting the performance of wind turbines. The analysis revealed that the Generalized Gamma Model provides the best fit for Bhopal's wind speed data, as indicated by the lowest AIC and a relatively modest KS-statistic. This suggests that

the Generalized Gamma distribution is capable of capturing the varied wind speed dynamics specific to Bhopal, potentially due to its flexibility in accommodating different wind speed regimes.

Given the complexity of solar irradiance data, which often exhibits multiple peaks due to diurnal and seasonal changes, mixture models were primarily utilized. These models are better equipped to handle the multimodal nature of GHI data. In the case of GHI data, the Weibull mixture model provided the best results across all locations.

The methodology employed, involving fitting various probability distributions to the wind speed and GHI data and evaluating these fits using statistical metrics, has provided insightful results. In terms of wind speed, the Generalized Gamma and Burr models generally offer better fits depending on the location, which suggests variability in wind patterns across different regions. For GHI, mixture models, particularly Weibull mixtures, prove to be highly effective, indicating the complex nature of solar irradiance data that requires models capable of capturing multiple modes of variability. In Hamirpur, the Exponentiated Weibull and Gamma Generalized models showcased superior performance. For Jafrabad, the Burr Model demonstrated notably better fit. he results for Thiruvananthapuram again highlighted the effectiveness of the Burr model. It managed to provide a robust fit across the wind speed data spectrum, balancing the trade-off between model complexity and fit quality.

Future work could focus on refining these models further or exploring additional multi-modal distributions that might better capture the underlying physical processes affecting both wind assist in optimizing the design and location of new renewable energy projects.

# References


[1] Ember, “Global Electricity Review 2023,” 2023.

[2] Ministry of New and Renewable Energy, “Ministry of New and Renewable Energy,” [Online]. Available: https://mnre.gov.in/physical-progress/. [Accessed 6 March 2024].

[3] Ember, “Yearly electricity data,” 2023.

[4] C. Gueymard, “Solar irradiance variations related to measured cloud behaviour,” *Journal of Applied Meteorology ,* vol. 31, pp. 247-254, 1992.

[5] A. K. Dube, Artificial Intelligence for Renewable Energy systems, Woodhead Publishing, 2022.

[6] C. Jung and D. Schindler, “Wind speed distribution selection – A review of recent development and progress,” *Renewable and Sustainable Energy Reviews,* 2019.

[7] C. Jung and D. Schindler, “National and global wind resource assessment under six wind turbine installation scenarios,” *Energy Conversion and Management,* 2018.

[8] Q. Hu a, Y. Wang, Z. X. P. Zhu and D. Yu, “On estimating uncertainty of wind energy with mixture of distributions,” *Energy,* vol. 112, pp. 935-962, 2016.

[9] C. Jacovides, “Statistical procedures for the evaluation of solar radiation estimation models,” *Atmospheric Research,* vol. 66, pp. 187-200, 2003.

[10] M. Diagne and P. Blanc, “Solar forecasting: a review,” *Renewable and Sustainable Energy Reviews,* vol. 31, pp. 272-283, 2014.

[11] S. Akdag, H. Bagiorgas and G. Mihalakakou, “Use of two-component Weibull mixtures in the analysis of wind speed in the Eastern Mediterranean,” *Applied Energy,* vol. 87, no. 8, pp. 2566-2573, 2010.