Machine-Learning–Assisted Probabilistic Wind Assessment at Sechura, Peru

Ubaldo Yancachajlla Tito[1], Celso Antonio Sanga Quiroz[2], Edilberto Velarde Coaquira[3], Germán Belizario Quispe[4]
Escuela de Posgrado, Universidad Nacional de San Agustin de Arequipa, Arequipa, Peru[1]
Escuela Profesional de Ingeniería Mecánica, Universidad Nacional de San Agustin de Arequipa, Arequipa, Peru[2]
Escuela Profesional de Ingeniería Agrícola, Universidad Nacional del Altiplano, Puno, Peru[3, 4]

***Abstract*—Accurate characterization of wind resources is essential for reliable energy yield estimation and wind farm planning, particularly in regions with limited long-term measurements. This study presents a machine-learning–assisted probabilistic wind assessment in Sechura, Peru, based on multi-year hourly wind data obtained from the NASA POWER database. A representative Typical Meteorological Year (TMY) was constructed to preserve seasonal and diurnal variability while enabling standardized annual energy production (AEP) calculations. Wind speed distributions were modeled using empirical distributions, kernel density estimation (KDE), the Weibull distribution, and Gaussian mixture models (GMM). Statistical evaluation indicates that KDE and GMM reduce the annual RMSE by more than 50% compared to the Weibull model, achieving coefficients of determination above 0.98. Annual energy production is estimated at approximately 1.88 GWh, with differences below 0.3% among probabilistic models. The corresponding capacity factor is approximately 0.25 for a utility-scale wind turbine. The results demonstrate that advanced probabilistic models substantially improve wind speed representation while having a limited impact on integrated annual energy estimates, highlighting the importance of model selection for variability and seasonal analysis rather than for annual yield estimation.**

***Keywords*—Wind resource assessment; probabilistic modeling; machine learning; kernel density estimation; Gaussian mixture model; annual energy production; capacity factor**

I. Introduction

Reliable characterization of the wind resource is a fundamental prerequisite for accurate annual energy production (AEP) estimation, optimal site selection, and robust planning of wind-power projects. Traditionally, wind-speed variability has been modeled using parametric probability density functions, with the two-parameter Weibull distribution being the most commonly employed due to its mathematical simplicity and the availability of closed-form expressions for wind power density and related energy metrics. However, numerous studies have shown that the Weibull distribution can present significant shortcomings when representing wind regimes characterized by skewness, intermittency, strong seasonality, or multimodal behavior. Comparative studies at regional and global scales have demonstrated that no single parametric model can universally capture the complexity of wind-speed distributions across different climatic contexts, particularly in coastal areas and complex terrain [1], [2]. Such limitations may lead to biased estimates of wind power density and, consequently, of AEP [3]–[5].

In response to these limitations, more flexible statistical approaches have been developed and applied. Kernel density estimation (KDE) has emerged as a prominent nonparametric alternative, capable of reconstructing the probability density function directly from observed data without imposing a predefined functional form. Several studies report that KDE systematically outperforms Weibull models in goodness-of-fit metrics and distributional distances, especially when advanced bandwidth-selection strategies and bias-correction techniques are applied [6]–[8].

At the same time, mixture models offer an effective parametric compromise by representing complex distributions as weighted combinations of simpler component distributions. Weibull mixtures and Gaussian mixture models have shown superior capability to capture wind multimodality and to provide more realistic uncertainty estimates for energy production [9]–[15]. These approaches have been extensively used both for wind-resource assessment and for analyses of variability and energy-related risk.

In regions lacking long in-situ measurement records, atmospheric reanalysis products have become an established and reliable data source for wind-resource evaluation. In particular, the NASA MERRA-2 reanalysis has been extensively validated and demonstrated to provide surface and near-surface wind fields suitable for energy applications [6], [16]. Platforms such as NASA POWER facilitate access to these data and have been widely used in studies of wind potential and turbine selection [17]– [22].

This study combines a regional case study (Sechura, Peru) with a differentiated methodological approach. Sechura is a coastal area with high wind potential that has been little explored using advanced methods. Recent studies indicate that wind energy in Peru remains underdeveloped and its potential is not yet fully exploited [23]. In this context, our research provides original value by applying advanced probabilistic techniques, assisted by machine-learning approaches, to characterize the local wind resource. Specifically, we employ an unsupervised Gaussian mixture model (GMM) to identify wind regimes in Sechura, which, to the best of our knowledge, has not been widely documented in the local literature.

Within this context, the present work develops a machine-learning–assisted probabilistic assessment of the wind resource for Sechura, Piura (Peru), using hourly MERRA-2 reanalysis data. We systematically compare empirical distributions, kernel density estimation, the Weibull distribution, and mixture

models, evaluating their statistical performance and their impact on wind power density estimation, annual energy production, and capacity factor. The objective is to provide a rigorous methodological perspective to improve wind-resource characterization in arid coastal environments, where wind variability and multimodality are critical factors.

II. METHODOLOGY

A. Study Area and Data

The study area is located in Sechura, Province of Sechura, Piura, Peru, on the northern Pacific coast of South America (latitude −5.99°, longitude −81.08°). The region is characterized by arid climatic conditions, low surface roughness, and relatively flat terrain, making it suitable for wind resource assessment. The local wind regime is influenced by southeast trade winds and land–sea breeze circulations, which contribute to pronounced seasonal and diurnal variability.

Hourly wind speed data at 50 m above ground level were obtained from the NASA POWER database, based on the MERRA-2 reanalysis [16]. The dataset spans the period from December 2014 to January 2026 and includes wind speed, wind direction, and surface pressure. The average elevation of the corresponding grid cell is approximately 31.65 m above sea level.

Due to the absence of long-term in situ wind measurements in Sechura, MERRA-2 reanalysis data were used as the primary data source. Reanalysis products combine historical global observations with numerical atmospheric models through data assimilation techniques, providing temporally consistent, long-term wind datasets with global coverage. Such datasets are widely employed in preliminary wind resource assessments, particularly in regions lacking ground-based measurements.

Nevertheless, previous studies have reported that reanalysis data may present biases in coastal environments, including potential underestimation of wind speeds due to spatial resolution and surface parameterization limitations [24]. Therefore, the results presented here should be interpreted as a probabilistic assessment based on large-scale atmospheric reanalysis data. Future work should incorporate local meteorological mast measurements to validate and calibrate the probabilistic models developed in this study.

The NASA POWER wind-speed data used in this study correspond to a nominal measurement height of 50 m above ground level. For consistency, no vertical extrapolation or wind-shear correction was applied to adjust the wind speeds to the turbine hub height. The selected Gamesa G58/850 turbine typically operates at hub heights slightly above this level (approximately 55–65 m), and the study area is characterized by flat terrain and low surface roughness. Under these conditions, the expected wind-speed difference between 50 m and the hub height is limited. Nevertheless, the use of 50 m wind speeds directly in the energy calculations represents a simplifying assumption. A more detailed site-specific assessment could apply logarithmic or power-law vertical wind profiles to explicitly account for wind shear effects.

Wind-speed and surface-pressure data were obtained from the NASA POWER reanalysis and updated through January 2026. Hourly series were downloaded manually from the NASA POWER portal and then processed using a Python pipeline that performs chronological concatenation of hourly records and basic quality control: nonphysical values (e.g., negative wind speeds) are clipped or removed, and remaining gaps are handled during aggregation. Derived quantities (e.g., air density) are computed from the cleaned time series, and a Typical Meteorological Year (TMY) was constructed by averaging all available observations for each calendar hour. This workflow supports straightforward manual updates as new data becomes available and thus facilitates frequent refreshes of the resource assessment.

B. Typical Meteorological Year Construction

To enable standardized AEP calculations while preserving long-term variability, a Typical Meteorological Year (TMY) was constructed from the multi-year dataset. For each calendar hour defined by month, day, and hour, the representative wind speed was computed as the arithmetic mean of all available observations corresponding to that time slot across the full record. Leap-day observations (29 February) were excluded to maintain a total of 8760 hourly values.

This procedure reduces the influence of anomalous years while retaining essential seasonal and diurnal variability, and is a widely accepted practice in wind and solar resource assessments [17], [20].

The Typical Meteorological Year (TMY) was constructed by averaging the multi-year hourly wind-speed series for each hour of the calendar year, rather than applying the Finkelstein–Schafer (FS) method, which selects representative months based on statistical goodness-of-fit criteria [25]. The hourly averaging approach produces a continuous and temporally homogeneous climatological year, avoiding potential discontinuities that may arise when concatenating months from different years.

Although the FS technique is widely applied in solar energy studies, the present wind-focused analysis prioritizes the temporal smoothness and long-term mean representativeness of the multi-year hourly average. This averaged TMY is appropriate for the comparative probabilistic assessment conducted in this study. Nevertheless, future research could evaluate a TMY constructed using the FS methodology to quantify potential differences in estimated energy production.

C. Probabilistic Wind Speed Models

1) Empirical distribution: The empirical probability density function (PDF) was obtained directly from the TMY wind speed data using normalized histograms and serves as the reference distribution for model comparison.

2) Kernel Density Estimation (KDE): KDE was employed as a nonparametric method to estimate the wind-speed PDF. Several bandwidth-selection strategies were considered, including normal-scale rules and cross-validation-based methods, which have demonstrated superior performance in wind-energy applications [6], [7], [8], [27].

The KDE of wind speed v is defined as

$$f_{KDE}(v) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{v - v_i}{h}\right) \quad (1)$$

where, n is the sample size, h is the bandwidth, and $K(\cdot)$ is a Gaussian kernel. Two bandwidth selection methods were considered: a normal-scale rule (KDE-NS) and least-squares cross-validation (KDE-LSCV), the latter being shown to provide superior performance in wind energy applications [4].

3) Weibull distribution: The two-parameter Weibull distribution was used as the parametric reference model. Shape and scale parameters were estimated via maximum likelihood following standard practice in the literature [1], [2], [5], [28].

The Weibull PDF is given by

$$f_W(v) = \frac{k}{c}\left(\frac{v}{c}\right)^{k-1} \exp\left[-\left(\frac{v}{c}\right)^{k}\right] \quad (2)$$

where, k is the shape parameter and c is the scale parameter. Maximum likelihood estimation was used to estimate the parameters. Despite its widespread use, the Weibull distribution may inadequately represent multimodal or highly variable wind regimes [1], [2].

4) Gaussian Mixture Model (GMM): To capture possible multimodal behaviour, mixture models were applied, including Weibull mixtures and Gaussian mixture models (GMMs). These models represent the overall PDF as a weighted sum of individual component distributions and have demonstrated superior capability to describe complex wind regimes and to quantify energetic uncertainty [9]–[15].

The GMM represents the wind speed PDF as a weighted sum of Gaussian components:

$$f_{GMM}(v) = \sum_{j=1}^{M} \pi_j \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left[-\frac{(v-\mu_j)^2}{2\sigma_j^2}\right] \quad (3)$$

where, M is the number of components, π_j are the mixing weights, and μ_j, σ_j are the mean and standard deviation of each component. A three-component GMM was selected based on model parsimony and interpretability, consistent with previous wind energy studies using mixture models [9].

Although the applied techniques correspond to classical statistical learning rather than deep learning, both kernel density estimation and Gaussian mixture models are formally categorized as unsupervised machine-learning methods.

The use of the term 'Machine-Learning–Assisted' is justified because the analysis incorporates an unsupervised machine-learning algorithm within the probabilistic modeling framework. Specifically, we employ a Gaussian Mixture Model (GMM) [26], which assumes that the wind-speed distribution arises from a weighted combination of multiple Gaussian components. Model parameters are estimated through the expectation–maximization (EM) algorithm, enabling the automatic identification of latent wind regimes without prior labeling. This unsupervised clustering approach captures multimodal characteristics of the wind-speed distribution that cannot be adequately represented by single-parameter models such as the Weibull distribution. Therefore, although the framework does not rely on supervised predictive learning, the integration of automated data-driven clustering techniques supports the use of the term 'machine-learning–assisted' in this study.

The optimal number of mixture components was determined using statistical information criteria. Models with varying numbers of clusters were fitted, and both the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) were computed for each configuration. The criteria exhibited a minimum (or clear elbow point) around k=3, indicating that three components provide an adequate balance between model goodness-of-fit and complexity. In addition, the selection of three clusters is physically consistent with the local wind-resource characteristics, where it is plausible to distinguish low-, medium-, and high-wind regimes. This approach avoids overfitting while ensuring a parsimonious yet flexible probabilistic representation of the wind-speed distribution.

To avoid heuristic selection of the number of mixture components, models with different values of k were evaluated using the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Both criteria reached their minimum at k=3, indicating that three components provide the optimal trade-off between model fit and complexity. The large ΔAIC relative to the Weibull model further supports this selection.

D. Statistical Performance Metrics

Model performance was assessed using a combination of goodness-of-fit metrics and distributional distance measures commonly adopted in wind-speed probabilistic modeling. These include the root mean square error (RMSE), mean absolute error (MAE), the Kolmogorov–Smirnov (KS) statistic, the Kullback–Leibler (KL) divergence, the Hellinger distance, and the one-dimensional Wasserstein distance, which together provide a comprehensive evaluation of both pointwise errors and global distributional discrepancies [3], [4].

The R^2, RMSE and MAE were computed as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(f_{\text{emp},i} - f_{\text{model},i}\right)^2}{\sum_{i=1}^{n}\left(f_{\text{emp},i} - \overline{f_{\text{emp}}}\right)^2} \quad (4)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(f_{emp,i} - f_{model,i}\right)^2} \quad (5)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|f_{emp,i} - f_{model,i}\right| \quad (6)$$

The Kolmogorov–Smirnov statistic was used to quantify the maximum absolute deviation between the empirical and modeled cumulative distribution functions (CDFs):

$$KS = \sup_{v} \left| F_{emp}(v) - F_{model}(v) \right| \tag{7}$$

Kullback–Leibler divergence

$$D_{KL}(f_{emp} \| f_{model}) = \sum_{i} f_{emp,i} \ln \left(\frac{f_{emp,i}}{f_{model,i}} \right) \tag{8}$$

Hellinger distance

$$H = \frac{1}{\sqrt{2}} \sqrt{\sum_{i} \left(\sqrt{f_{emp,i}} - \sqrt{f_{model,i}} \right)^2} \tag{9}$$

Wasserstein distance (1D, discrete form)

$$W = \sum_{i} \left| F_{emp}(v_i) - F_{model}(v_i) \right| \Delta v \tag{10}$$

E. Wind Power Density Estimation

Wind power density (WPD) was computed from each modeled PDF, incorporating air density estimated from surface pressure data. Annual energy production and capacity factor were calculated by mapping wind speeds to the power curve of a representative commercial wind turbine and integrating the resulting hourly power time series according to standard procedures [17]– [22].

Wind power density (WPD) was calculated from each modeled PDF as:

$$WPD = \frac{1}{2} \rho \int_{0}^{\infty} v^3 f(v) dv \tag{11}$$

where, ρ is air density, computed from surface pressure data under standard atmospheric assumptions.

F. Energy Production and Capacity Factor

Hourly electrical power output was obtained by mapping wind speeds to a commercial-scale wind turbine power curve. The annual energy production (AEP) was computed as:

$$E_{annual} = \sum_{i=1}^{8760} P_i \Delta t, \qquad \Delta t = 1\,\text{h} \tag{12}$$

The mean annual power was calculated as:

$$\bar{P} = \frac{E_{annual}}{8760} \tag{13}$$

The capacity factor (CF) was defined as:

$$CF = \frac{\bar{P}}{P_{rated}} \tag{14}$$

Seasonal energy production and capacity factors were computed analogously by integrating over the corresponding seasonal subsets of the TMY dataset.

Reference wind turbine. A commercial horizontal-axis wind turbine, Gamesa G58/850, was selected as the reference machine for energy production and capacity factor calculations, with technical specifications obtained from TheWindPower.net [29]. The turbine has a rated power of 850 kW and a three-bladed rotor with a diameter of 58 m (swept area = 2,642 m²). Cut-in, rated, and cut-out wind speeds are 3.0 m/s, 12.5 m/s, and 25.0 m/s, respectively. Maximum rotor speed is 30.8 rpm (tip speed ≈ 94 m/s). Rotor blades are fiberglass/epoxy composites. This turbine model was used to map probabilistic wind speed distributions to electrical output via the manufacturer's power curve for AEP and CF estimation [29].

G. Computational Implementation and Cost

To assess practical computational costs, bandwidth selection via leave-one-out least-squares cross-validation (LSCV) was compared against a simple rule-of-thumb (Silverman's rule). Using our Python implementation (vectorized operations and peak memory tracking via tracemalloc), the following empirical results were obtained: the LSCV bandwidth search required ≈ 1.06 s with a peak Python allocation of ≈ 32.03 MB, whereas KDE evaluation on the fine integration grid (using either the Silverman bandwidth or the LSCV-selected bandwidth) completed in ≈ 0.33 – 0.35 s with negligible allocation peaks (≈ 0.11 MB). All benchmarks were executed on the author's workstation: Intel Corporation Core(TM) i5-7200U @ 2.50 GHz (2 physical cores, 4 logical processors) with 17,060,569,088 bytes (≈ 15.89 GB) RAM. These results indicate that the dominant computational overhead arises from the repeated evaluations inherent to LSCV bandwidth selection rather than from KDE evaluation once the bandwidth is fixed. For practical applications, we therefore recommend: 1) subsampling or parallelization when applying LSCV to large datasets; 2) using rule-of-thumb bandwidths (Silverman/Scott) or approximate KDE implementations (KD-tree, FFT/binning) for very large n or latency-constrained settings; and 3) complementing tracemalloc with system-level monitoring tools (e.g., psutil) when total process memory usage is required. Table I summarizes the measured execution times and peak Python allocation on the reference workstation.

TABLE I. BENCHMARK RESULTS (EXECUTION TIME AND PEAK PYTHON ALLOCATION

Metric	Time (s)	Peak allocation (bytes)	Peak (MB)
LSCV bandwidth search	1.0621	32033025	30.54
KDE evaluation (Silverman)	0.3499	119464	0.11
KDE evaluation (LSCV bandwidth)	0.3341	119424	0.11

Note: Reference workstation: Intel Corporation Core(TM) i5-7200U CPU @ 2.50 GHz — 2 physical cores (4 logical processors); 17,060,569,088 bytes (≈ 15.89 GB) RAM. Measurements performed on the author's machine; memory values report Python allocation peaks captured by tracemalloc.

Execution times and memory usage depend on hardware and system configuration (e.g., multithreaded BLAS).

III. Results and Discussion

The performance of probabilistic wind speed models was evaluated using both statistical and distributional metrics on the Typical Meteorological Year (TMY) dataset for Sechura, Peru. The models analyzed include empirical histograms, kernel density estimation with normal and least-squares cross-validation bandwidths (KDE-NS and KDE-LSCV), a two-parameter Weibull distribution, and a three-component Gaussian mixture model (GMM(3)). The main objective of this section is to compare these models in terms of fit accuracy and to interpret their implications for wind resource assessment.

A. Annual PDF Fitting Accuracy

Fig. 1 compares the annual empirical wind speed distribution with the fitted PDFs and CDFs obtained using KDE, Weibull, and GMM models. The KDE-LSCV closely follows the empirical distribution across the entire wind speed range, while the Weibull model exhibits noticeable deviations, particularly in the tails, explaining its poorer goodness-of-fit metrics reported in Table II.

Table II summarizes the annual goodness-of-fit metrics for each model. The empirical distribution serves as a perfect reference, as expected, with zero error across all metrics.

Among the fitted models, KDE-LSCV achieves the best overall performance, with the highest coefficient of determination (R^2 = 0.98) and the lowest values in the Kolmogorov–Smirnov statistic (KS = 0.02), RMSE, MAE, and distributional distance measures (Hellinger, Kullback–Leibler, and Wasserstein). This confirms that KDE-LSCV accurately represents both the central tendency and the tails of the wind speed distribution, which is critical for wind energy applications where tail behavior influences extreme events and energy yield predictions [6]– [8], [23].

Similar conclusions regarding the superiority of kernel-based approaches have been reported in recent studies employing advanced and adaptive bandwidth KDE techniques. For example, Chau et al. [30] demonstrated that hybrid adaptive KDE formulations significantly improve wind speed representation across diverse climatic conditions, while Ning et al. [31] showed that KDE-based probabilistic modeling provides robust performance even in highly variable and extreme wind regimes. Although the present study employs classical LSCV bandwidth selection, the results are consistent with these findings and confirm the strong suitability of KDE for wind resource assessment.

In contrast, the Weibull distribution exhibits substantially poorer performance, with R^2 = 0.62 and the largest error and distance metrics (e.g., Wasserstein = 0.37). This indicates a significant mismatch with the empirical distribution, especially in the tails. These findings align with previous comparisons showing that fixed parametric models such as Weibull can be inadequate in complex wind regimes [1]– [5].

The GMM(3) model improves significantly upon the Weibull distribution and yields results comparable to KDE-NS, demonstrating that mixture distributions are effective in capturing multimodal wind speed structures. Previous studies likewise highlight the advantages of mixture models for improved uncertainty representation in wind energy modeling [9]– [15]. Nevertheless, GMM(3) remains slightly inferior to KDE-LSCV across most evaluation metrics, reflecting the superior flexibility of non-parametric density estimation methods.

TABLE II. Annual PDF Fit Metrics

Model	R^2	KS	RMSE	MAE	Hellinger	KL	Wasserstein
Empirical	1.00	0.00	0.00	0.00	0.00	0.00	0.00
KDE-NS	0.95	0.03	56.56	31.43	0.06	0.01	0.05
KDE-LSCV	0.98	0.02	39.35	23.07	0.04	0.01	0.02
Weibull	0.62	0.11	154.00	95.00	0.18	0.12	0.37
GMM (3)	0.94	0.03	59.70	36.49	0.07	0.02	0.05

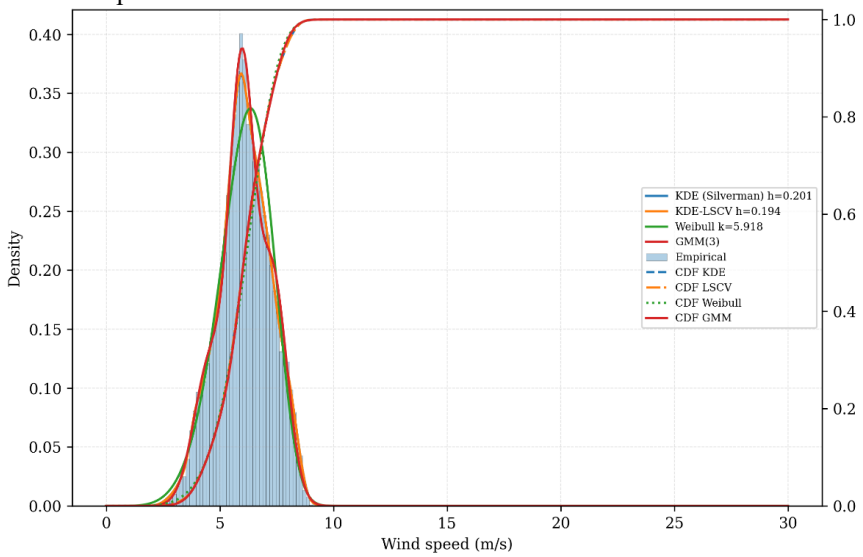

Fig. 1. Comparison of annual wind speed probability density functions (PDFs) and cumulative distribution functions (CDFs) for Sechura using empirical data, KDE, Weibull, and GMM models.

B. Seasonal Probabilistic Fitting Performance

Fig. 2 presents the seasonal empirical wind speed distributions together with the fitted probability density functions (PDFs) and cumulative distribution functions (CDFs) for Summer, Autumn, Winter, and Spring. Across all seasons, the KDE-LSCV model closely follows the empirical distribution, accurately capturing both central tendencies and tail behavior. In contrast, the Weibull distribution exhibits noticeable deviations, particularly during Winter and Summer, where multimodal features are evident in the empirical data. The GMM(3) model provides an improved representation of multimodality compared to Weibull, although it remains slightly less accurate than KDE-LSCV in terms of overall distributional agreement.

Seasonal performance for Summer, Autumn, Winter and Spring is shown in Table II. KDE-LSCV consistently achieves the highest R^2 and the lowest KS, RMSE, MAE, and distance metrics across all seasons, indicating robustness in capturing seasonal wind speed variations.

The GMM(3) model exhibits intermediate performance, consistently outperforming the Weibull distribution across most evaluation metrics. In particular, the Weibull model performs poorly during extreme seasonal regimes (e.g., summer and winter), as evidenced by elevated Kolmogorov–Smirnov statistics and large error values. This behavior indicates a limited ability to adapt to seasonally varying distribution shapes and tail characteristics. These observations are consistent with findings in the literature, which report that simple parametric assumptions frequently fail in environments characterized by complex and highly variable climatic patterns [3]– [5].

Table III lists the seasonal PDF fit metrics for the selected wind-speed models (KDE-LSCV, GMM(3) and Weibull). The table reports R^2, Kolmogorov–Smirnov (KS) statistic, RMSE, MAE and three distributional distances (Hellinger, KL and Wasserstein) for Summer, Autumn, Winter and Spring.

As shown in Table III, KDE-LSCV consistently delivers the best seasonal goodness-of-fit, with $R^2 \approx 0.97$–0.98 and the lowest KS, RMSE and distributional distances across all seasons. GMM(3) produces competitive, intermediate results (typically $R^2 \approx 0.95$), while the Weibull model performs poorly ($R^2 \approx 0.60$–0.65) and exhibits the largest errors and distributional discrepancies, especially in Winter and Summer. These seasonal patterns confirm the robustness of non-parametric and mixture approaches for capturing changing wind regimes (Table III).

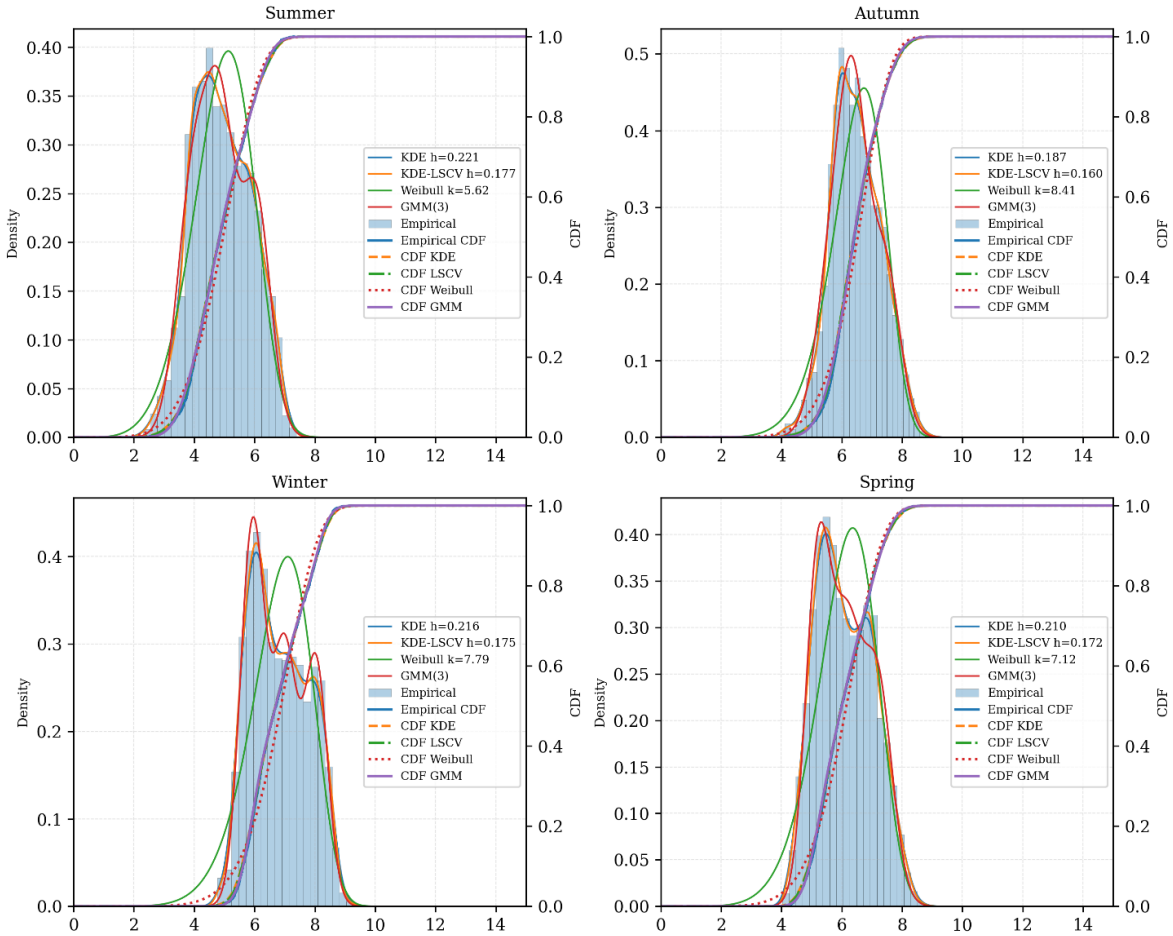

Fig. 2. Seasonal empirical wind speed distributions and fitted PDFs and CDFs for Summer, Autumn, Winter, and Spring using KDE, Weibull, and GMM models.

TABLE III. SEASONAL PDF FIT METRICS FOR SELECTED WIND SPEED MODELS

Season	Model	R^2	KS	RMSE	MAE	Hellinger	KL	Wasserstein
Summer	KDE-LSCV	0.97	0.02	45.12	25.99	0.04	0.01	0.02
	GMM (3)	0.95	0.03	52.33	30.44	0.05	0.01	0.05
	Weibull	0.61	0.11	160.23	98.11	0.19	0.13	0.39
Autumn	KDE-LSCV	0.98	0.02	35.99	21.46	0.04	0.00	0.02
	GMM (3)	0.96	0.03	49.12	29.25	0.05	0.01	0.04
	Weibull	0.65	0.10	140.99	85.12	0.17	0.11	0.36
Winter	KDE-LSCV	0.97	0.02	40.46	23.35	0.04	0.01	0.02
	GMM (3)	0.95	0.03	55.68	34.12	0.06	0.01	0.06
	Weibull	0.60	0.12	170.46	105.68	0.19	0.14	0.41
Spring	KDE-LSCV	0.97	0.02	42.33	24.57	0.04	0.01	0.02
	GMM (3)	0.95	0.03	57.00	36.79	0.06	0.01	0.06
	Weibull	0.61	0.12	165.12	100.23	0.19	0.13	0.40

C. Impact on Energy Production and Capacity Factors

Table IV presents the seasonal mean power output, annual energy production (AEP), and capacity factor estimated using the different probabilistic models. Although substantial differences are observed in the quality of the wind speed distribution fitting, the resulting integrated energy metrics remain remarkably consistent across models. Specifically, deviations in AEP and capacity factor are below 0.3%, indicating that long-term energy estimates are relatively insensitive to moderate discrepancies in probability density representation.

This behavior can be attributed to the smoothing effect of turbine power curves and the dominance of mid-range wind speeds in the energy contribution, which reduces sensitivity to distributional differences in the tails. Similar findings have been reported in previous wind energy studies, where parametric and non-parametric models yield comparable long-term energy outputs despite notable differences in statistical performance metrics [17]– [22].

Nevertheless, accurate probabilistic modeling remains critical when the scope extends beyond aggregated annual indicators. Applications such as uncertainty quantification, risk assessment, seasonal forecasting, and turbine selection or optimization require faithful representation of the full wind speed distribution, particularly in the tails, where extreme events and variability play a decisive role.

TABLE IV. SEASONAL ENERGY PRODUCTION AND CAPACITY FACTOR

Season	Model	Mean power (kW)	AEP (MWh)	CF
Summer	Empirical	115.87	250.28	0.14
	KDE-NS	116.39	251.40	0.14
	KDE-LSCV	116.20	251.00	0.14
	Weibull	116.62	251.91	0.14
	GMM (3)	115.87	250.28	0.14
Autumn	Empirical	246.28	537.88	0.29
	KDE-NS	246.60	538.58	0.29
	KDE-LSCV	246.52	538.39	0.29
	Weibull	246.27	537.86	0.29
	GMM (3)	246.29	537.91	0.29
Winter	Empirical	282.67	624.13	0.33
	KDE-NS	283.06	625.00	0.33
	KDE-LSCV	282.93	624.70	0.33
	Weibull	283.65	626.31	0.33
	GMM (3)	282.67	624.13	0.33
Spring	Empirical	211.47	466.92	0.25
	KDE-NS	211.88	467.84	0.25
	KDE-LSCV	211.75	467.54	0.25
	Weibull	212.11	468.35	0.25
	GMM (3)	211.49	466.96	0.25

Note: AEP and CF were computed using the power curve of the Gamesa G58/850 wind turbine (rated power: 850 kW; rotor diameter: 58 m; cut-in / rated / cut-out wind speeds: 3.0 / 12.5 / 25.0 $m \cdot s^{-1}$). Technical specifications were obtained from TheWindPower.net [29].

Fig. 3 compares the seasonal energy production estimated using the different probabilistic wind speed models. Despite the substantial differences observed in the statistical fitting accuracy of the distributions, the resulting seasonal energy estimates are remarkably consistent across all models. For all seasons, deviations in AEP remain very small, confirming that integrated energy metrics are weakly sensitive to moderate discrepancies in wind speed probability density modeling. This behavior is mainly attributed to the smoothing effect of turbine power curves and the dominance of mid-range wind speeds in energy production.

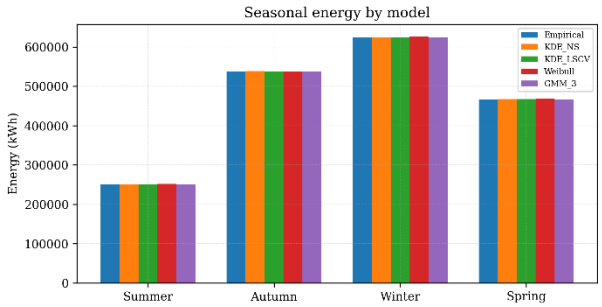

Fig. 3. Seasonal energy production estimated using different probabilistic wind speed models.

TABLE V. ANNUAL GENERATED ENERGY AND CAPACITY FACTOR FOR DIFFERENT PROBABILISTIC WIND MODELS

Model	Generated Energy (Wh)	Capacity Factor
Empirical	1,879,199,060	0.25
KDE-NS	1,882,521,226	0.25
KDE-LSCV	1,882,318,322	0.25
Weibull	1,883,887,128	0.25
GMM(3)	1,879,320,489	0.25

As shown in Table V, the estimated annual energy production varies by less than 0.3% among the considered probabilistic models, resulting in capacity factors close to 0.25 for all cases. Although advanced probabilistic models significantly improve the representation of wind speed distributions, their impact on integrated annual energy production is limited, as evidenced by the results in Table V. This confirms that annual energy yield is primarily driven by the mean wind regime, whereas model selection becomes more relevant for variability, extremes, and seasonal analysis.

From a methodological perspective, the results highlight the effectiveness of data-driven and machine-learning-based probabilistic models for wind resource characterization. Although the applied techniques belong to classical statistical learning rather than deep learning, both kernel density estimation with data-adaptive bandwidth selection and Gaussian mixture models can be formally classified as unsupervised machine learning approaches. Their superior performance relative to classical parametric models confirms the value of machine learning tools for capturing complex wind speed distributions.

These statistical indicators quantify the degree of divergence between wind-speed probability distributions, which directly translates into differences in the estimated annual energy production (AEP). From an energy perspective, low values of KS, KL divergence, or Hellinger distance (approaching zero) indicate highly similar distributions, implying that AEP estimates derived from alternative models would differ only marginally. Conversely, higher metric values reveal substantial discrepancies in the representation of wind regimes, particularly in the tails of the distribution, which disproportionately influence power output due to the cubic dependence of wind power on velocity.

In particular, divergence-based measures such as KL and Hellinger are sensitive to distributional shape differences, including multimodality and tail behavior, which can lead to meaningful variations in AEP when inadequately modeled. The Wasserstein distance provides an intuitive interpretation, as it measures the average displacement required to transform one distribution into another in the wind-speed domain; larger values therefore indicate greater shifts in effective wind regimes and potentially greater deviations in energy yield. Overall, these metrics provide a quantitative bridge between statistical goodness-of-fit and the reliability of energy production estimates.

As a potential extension of this work, the probabilistic framework developed here could be used to estimate P50 and P90 energy production levels. This would involve generating multiple synthetic wind-speed realizations from the fitted probabilistic model (e.g., via Monte Carlo sampling) and computing the corresponding annual energy production (AEP) for each realization. The resulting AEP distribution would allow direct extraction of production percentiles such as P50 and P90. Alternatively, analytical integration of the estimated joint probability density could be explored to derive production quantiles. Although beyond the scope of the present study, this extension would enable explicit quantification of production uncertainty and improve risk assessment in wind-energy project evaluation.

IV. CONCLUSIONS

This study presented a comprehensive probabilistic assessment of the wind resource at Sechura, Peru, based on a Typical Meteorological Year derived from long-term reanalysis data. Empirical distributions, kernel density estimation, Weibull fitting, and Gaussian mixture models were systematically compared using classical goodness-of-fit metrics, distributional distance measures, and energy-based indicators at both annual and seasonal scales.

The results demonstrate that non-parametric kernel density estimation with least-squares cross-validation (KDE–LSCV) provides the most accurate representation of the wind speed distribution. At the annual scale, KDE–LSCV consistently achieved the highest coefficient of determination and the lowest Kolmogorov–Smirnov statistic, RMSE, and distributional distances (Hellinger, Kullback–Leibler, and Wasserstein). These findings confirm its superior ability to reproduce both the central tendency and the tails of the empirical distribution.

Seasonal analyses further reinforced these conclusions. KDE–LSCV outperformed all alternative models in Summer, Autumn, Winter, and Spring, exhibiting stable performance across contrasting wind regimes. In particular, the Winter season revealed pronounced limitations of unimodal parametric models, where the Weibull distribution failed to capture the observed complexity of the wind regime, resulting in very poor goodness-of-fit metrics. In contrast, Gaussian mixture models demonstrated strong seasonal robustness, especially during Winter, indicating their suitability for representing multimodal wind patterns.

Despite the substantial differences observed in statistical goodness-of-fit metrics, annual and seasonal energy production estimates remained remarkably consistent across models, with deviations generally below 0.3 %. Electrical efficiency and capacity factor values were virtually unchanged, indicating that energy-based indicators are less sensitive to moderate distributional discrepancies under the conditions examined. Nevertheless, accurate probabilistic modeling remains essential for applications involving risk analysis, extreme wind assessment, and turbine selection, where misrepresentation of distribution tails can have significant consequences.

Overall, the results show that reliance on a single Weibull distribution may lead to systematic mischaracterization of wind regimes in coastal sites with complex seasonal variability. KDE–LSCV is recommended as the primary modeling approach for wind resource characterization at Sechura, while Gaussian mixture models provide a valuable parametric alternative when

compact analytical representations are required. The combined use of distributional distance metrics and energy-based indicators is strongly recommended for robust wind energy assessments.

Future work should focus on validating these findings using in situ measurements, extending the analysis to hub-height extrapolation and uncertainty quantification, and evaluating the implications of probabilistic model choice on long-term economic performance and grid-integration studies.

REFERENCES

[1] P. Wais, "A review of Weibull functions in wind sector," Renewable and Sustainable Energy Reviews, vol. 70, pp. 1099–1107, Apr. 2017, doi: 10.1016/J.RSER.2016.12.014.

[2] J. A. Carta, P. Ramírez, and S. Velázquez, "A review of wind speed probability distributions used in wind energy analysis: Case studies in the Canary Islands," Renewable and Sustainable Energy Reviews, vol. 13, no. 5, pp. 933–955, 2009, doi: https://doi.org/10.1016/j.rser.2008.05.005.

[3] C. Jung and D. Schindler, "Global comparison of the goodness-of-fit of wind speed distributions," Energy Convers. Manag., vol. 133, pp. 216–234, 2017, doi: https://doi.org/10.1016/j.enconman.2016.12.006.

[4] N. Masseran, "Evaluating wind power density models and their statistical properties," Energy, vol. 84, pp. 533–541, 2015, doi: https://doi.org/10.1016/j.energy.2015.03.018.

[5] T. P. Chang, "Estimation of wind energy potential using different probability density functions," Appl. Energy, vol. 88, no. 5, pp. 1848–1856, 2011, doi: https://doi.org/10.1016/j.apenergy.2010.11.010.

[6] Q. Han, S. Ma, T. Wang, and F. Chu, "Kernel density estimation model for wind speed probability distribution with applicability to wind energy assessment in China," Renewable and Sustainable Energy Reviews, vol. 115, p. 109387, 2019, doi: https://doi.org/10.1016/j.rser.2019.109387.

[7] M. Wahbah, B. Mohandes, T. H. M. EL-Fouly, and M. S. El Moursi, "Unbiased cross-validation kernel density estimation for wind and PV probabilistic modelling," Energy Convers. Manag., vol. 266, p. 115811, 2022, doi: https://doi.org/10.1016/j.enconman.2022.115811.

[8] X. Xu, Z. Yan, and S. Xu, "Estimating wind speed probability distribution by diffusion-based kernel density method," Electric Power Systems Research, vol. 121, pp. 28–37, 2015, doi: https://doi.org/10.1016/j.epsr.2014.11.029.

[9] S. A. Akdağ, H. S. Bagiorgas, and G. Mihalakakou, "Use of two-component Weibull mixtures in the analysis of wind speed in the Eastern Mediterranean," Appl. Energy, vol. 87, no. 8, pp. 2566–2573, 2010, doi: https://doi.org/10.1016/j.apenergy.2010.02.033.

[10] J. A. Carta and P. Ramírez, "Analysis of two-component mixture Weibull statistics for estimation of wind speed distributions," Renew. Energy, vol. 32, no. 3, pp. 518–531, 2007, doi: https://doi.org/10.1016/j.renene.2006.05.005.

[11] X. Qin, J. Zhang, and X. Yan, "Two Improved Mixture Weibull Models for the Analysis of Wind Speed Data," J. Appl. Meteorol. Climatol., vol. 51, no. 7, pp. 1321–1332, 2012, doi: https://doi.org/10.1175/JAMC-D-11-0231.1.

[12] Q. Hu, Y. Wang, Z. Xie, P. Zhu, and D. Yu, "On estimating uncertainty of wind energy with mixture of distributions," Energy, vol. 112, pp. 935–962, 2016, doi: https://doi.org/10.1016/j.energy.2016.06.112.

[13] F. S. dos Santos, K. K. F. do Nascimento, J. da Silva Jale, S. F. A. Xavier, and T. A. E. Ferreira, "Brazilian wind energy generation potential using mixtures of Weibull distributions," Renewable and Sustainable Energy Reviews, vol. 189, p. 113990, 2024, doi: https://doi.org/10.1016/j.rser.2023.113990.

[14] M. Franco and J.-M. Vivo, "Constraints for generalized mixtures of Weibull distributions with a common shape parameter," Stat. Probab. Lett., vol. 79, no. 15, pp. 1724–1730, 2009, doi: https://doi.org/10.1016/j.spl.2009.05.005.

[15] W. Wang, Y. Gao, and N. Ikegaya, "Approximating wind speed probability distributions around a building by mixture weibull distribution with the methods of moments and L-moments," Journal of Wind Engineering and Industrial Aerodynamics, vol. 257, p. 106001, 2025, doi: https://doi.org/10.1016/j.jweia.2024.106001.

[16] D. Carvalho, "An assessment of NASA's GMAO MERRA-2 reanalysis surface winds," J. Clim., vol. 32, no. 23, pp. 8261–8281, 2019, doi: 10.1175/JCLI-D-19-0199.1.

[17] Z. Wang and W. Liu, "Wind energy potential assessment based on wind speed, its direction and power data," Sci. Rep., vol. 11, no. 1, p. 16879, 2021, doi: 10.1038/s41598-021-96376-7.

[18] B. Belabes, A. Youcefi, O. Guerri, M. Djamai, and A. Kaabeche, "Evaluation of wind energy potential and estimation of cost using wind energy turbines for electricity generation in north of Algeria," Renewable and Sustainable Energy Reviews, vol. 51, pp. 1245–1255, 2015, doi: https://doi.org/10.1016/j.rser.2015.07.043.

[19] M. S. Adaramola, S. S. Paul, and S. O. Oyedepo, "Assessment of electricity generation and energy cost of wind energy conversion systems in north-central Nigeria," Energy Convers. Manag., vol. 52, no. 12, pp. 3363–3368, 2011, doi: https://doi.org/10.1016/j.enconman.2011.07.007.

[20] P. Spiru and P. L. Simona, "Wind energy resource assessment and wind turbine selection analysis for sustainable energy production," Sci. Rep., vol. 14, no. 1, p. 10708, 2024, doi: 10.1038/s41598-024-61350-6.

[21] M. M. Ahmed, Md. K. Islam, and M. H. Masud, "Evaluating wind energy feasibility in different locations of Bangladesh: A techno-economic and environmental perspective," Energy Conversion and Management: X, vol. 28, p. 101352, 2025, doi: https://doi.org/10.1016/j.ecmx.2025.101352.

[22] F. H. Ouerghi, M. Omri, A. A. Menaem, A. I. Taloba, and R. M. Abd El-Aziz, "Feasibility evaluation of wind energy as a sustainable energy resource," Alexandria Engineering Journal, vol. 106, pp. 227–239, 2024, doi: https://doi.org/10.1016/j.aej.2024.06.055.

[23] A. Gordillo Valdez, M. Montoya Granda, and P. A. Salinas Pedemonte, "Análisis del desarrollo y potencial de la energía eólica en el Perú," Ingeniería Industrial, vol. 43, pp. 177–198, Oct. 2022, doi: 10.26439/ing.ind2022.n43.6114.

[24] Gualtieri, G. (2022). Analysing the uncertainties of reanalysis data used for wind resource assessment: A critical review. Renewable and Sustainable Energy Reviews, 167, 112741. https://doi.org/https://doi.org/10.1016/j.rser.2022.112741

[25] J. M. Finkelstein and R. E. Schafer,"Improved goodness-of-fit tests," Biometrika, vol. 58, no. 3, pp. 641–645, 1971. [Online]. Available: https://academic.oup.com/biomet/article-abstract/58/3/641/233677

[26] C. M. Bishop, Pattern Recognition and Machine Learning. New York, NY, USA: Springer, 2006. doi: 10.1007/978-0-387-45528-0.

[27] M. Wahbah, S. F. Feng, T. H. M. EL-Fouly, and B. Zahawi, "Wind speed probability density estimation using root-transformed local linear regression," Energy Convers. Manag., vol. 199, p. 111889, 2019, doi: https://doi.org/10.1016/j.enconman.2019.111889.

[28] J. Liu, G. Xiong, X. Fu, and A. W. Mohamed, "Estimating the best-fit parameters of Weibull distribution with numerical methods for wind energy assessment: A case study in China," Energy Strategy Reviews, vol. 63, p. 102017, 2026, doi: https://doi.org/10.1016/j.esr.2025.102017.

[29] The Wind Power, "Gamesa G58–850 Wind Turbine," TheWindPower.net. [Online]. Available: https://www.thewindpower.net/turbine_en_43_gamesa_g58-850.php. Accessed: Jan. 17, 2026.

[30] T. T. Chau, A. B. Alhassan, M. Shaltayev, K. Talapiden, M. A. Shehu, and T. D. Do, "Assessment of wind energy potential using hybrid adaptive bandwidth kernel density technique: A case study of major cities in Kazakhstan," Energy Conversion and Management: X, vol. 29, p. 101439, 2026, doi: https://doi.org/10.1016/j.ecmx.2025.101439.

[31] J. Ning, R. Shi, S. Xuan, C. Jiang, and L. Jia, "Assessment of offshore island wind energy potential in typhoon-prone regions with a KDE-based probabilistic modeling approach," Energy, vol. 339, p. 139092, 2025, doi: https://doi.org/10.1016/j.energy.2025.139092.