Alexandria Engineering Journal 87 (2024) 175–185

Contents lists available at ScienceDirect

Alexandria Engineering Journal

journal homepage: www.elsevier.com/locate/aej

ORIGINAL ARTICLE

Performance estimation technique for solar-wind hybrid systems: A machine learning approach

Fabian Salazar-Caceres [a], Harrynson Ramirez-Murillo [a], Carlos Andrés Torres-Pinzón [b,*], Martha Patricia Camargo-Martínez [a]

[a] Universidad de La Salle, Bogotá, Colombia
[b] Universidad Santo Tomás, Electronic Engineering Faculty, Bogotá D.C., Colombia

ARTICLE INFO

Dataset link: https://github.com/Ibrahim-Jayyam/La-Salle-Dataset-Energy-Estimation.git

MSC:
0000
1111

Keywords:
Bootstrap method
Clustering analysis
Hybrid power systems
Machine learning
Probability density function
Solar power generation
Wind power generation

ABSTRACT

This study introduces a novel method to assess solar-wind hybrid energy systems, utilizing data analysis for key metrics (Performance Yields Y_R, Capacity Factor *CF*, and Performance Ratio *PR*), focused on their uncertainties estimation. Employing a bootstrapped-based K-means algorithm on features from a local weather station in Universidad de La Salle, Bogotá, the approach evaluates the technical feasibility of a photovoltaic system on the campus. It discusses the clustering algorithm advantages, emphasizing its suitability for modeling continuous processes in renewable energy systems. The hypothesis assumes independence and identically distributed centroid coordinates, aligning with large-scale energy system characteristics. Methods include data normalization, feature selection, statistical calculations, bootstrapping, and Probability Density Functions (PDF) estimation through a Maximum Likelihood Estimator (MLE) to assess uncertainties. Results suggest the unfeasibility of a wind energy system due to technological constraints. The study evaluates system flexibility, and design considerations, providing valuable decision-making insights. The conclusion highlights the potential use for accurate estimations, enhancing the system reliability.

1. Introduction

Nowadays, hybrid renewable energy systems are widely studied, designed, and implemented exploiting abundant resources such as solar and wind energy. Most of the design approaches rely on third-party software that uses some collected data to set values for solar or wind energy as an input parameter in the design process, establishing a fixed scenario where there is no uncertainty or estimations of the potential energy available. In general, the performance indexes yielded under these assumptions are very conservative implying an oversized or undersized technological implementation of the energy system proposed. In this work, we proposed a data-driven method for decision-making about renewable energy system performance considering collected data, machine learning algorithms, and performance indexes based on uncertainty analysis.

K-means clustering is a widely used Data Science technique, and this method has advantages and disadvantages compared to other ones. Some of these advantages are simplicity, scalability, speed, and interpretability. On the other hand, some disadvantages are local optima convergence, spherical clusters, and the curse of dimensionality. When choosing a clustering technique, it is essential to consider the specific characteristics of the dataset and the goals of the analysis [1]. Other clustering methods like Hierarchical Clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Gaussian Mixture Models (GMM), and spectral clustering offer alternatives that overcome some K-means limitations [2]. Combining K-means bootstrapping and Monte Carlo simulation allows an understanding of the uncertainties in K-means centroids estimation [3]. On the other hand, K-means clustering has applications in the artificial intelligence field in unsupervised learning tasks such as customer segmentation and anomaly detection [4,5]. In telecommunications, K-means enables the identification of usage patterns and network congestion, thus optimizing resource allocation and enhancing network performance [6]. Some authors focused primarily on neural network stability [7–9], and they have been advanced in understanding neural network stability under stochastic and time-varying conditions, which is crucial for various real-world applica-

* Corresponding author.
E-mail address: carlostorresp@usta.edu.co (C.A. Torres-Pinzón).

https://doi.org/10.1016/j.aej.2023.12.029
Received 6 September 2023; Received in revised form 12 December 2023; Accepted 15 December 2023
Available online 27 December 2023

tions. However, the application of neural network models in clustering tasks, in conjunction with algorithms like K-means, has been used in high-dimensional data such as images [10].

Research works such as [11] evaluated the energy resources in Romania, specifically photovoltaic energy, wind, biomass, and water, clustering-based data mining method. This analysis suggests new techniques and potential resources should be considered. The study [12] in Colombia also applied cluster and Principal Component Analysis techniques to calculate the photovoltaic energy potential in four cities, providing a reference point for further research. On the other hand, [13] used a combination of Geographic Information Systems, Machine Learning, and Image Processing to determine the potential for large-scale photovoltaic. Clustering techniques help to estimate the available renewable energy to a group of similar data points. These techniques can identify patterns in renewable energy production data, being applied to forecast future energy production and decrease greenhouse gas emissions [14,15].

Some studies as [16] were carried out, where the features considered correspond to solar radiation, temperature, wind speed, and cost. This data exploration is done hourly, and it uses an autoregressive prediction model for each of the variables estimating the energy obtained through a real world model. The dataset obtained comes from satellites. The authors establish a correlation between the non-linear variables cost and solar radiation instead of the mutual information, which can capture nonlinear relationships and dependencies between variables [17,18]. In addition to the mentioned above, in-situ samples are necessary to provide a more realistic solution for the specific application and improve the resolution of the features. When working with an external source, a data adjustment is required to ensure the quality and suitability of the intended analysis, where the output of error calculation is typically a single value equal to the error or difference between the measured and real values, instead of a Probability Density Function (PDF) or a set of statistics that describe the range of expected output values and their likelihoods [19].

There exist analyses where the energy potential of renewable resources are evaluated, relying on extensive knowledge of geographical areas, elements and characteristics of natural surroundings or by interpreting complex spatial modeling maps [20]. The studies reveal a marked gap between data analysis methods, management information systems, and evaluation for creating mathematical models that depict the energy potential of specific geographical regions. Besides analyzing the energy resource itself, it is equally significant to evaluate both the technical and economic potential [21,22]. On the other hand, some authors have quantified the feasibility solar and wind generation systems by calculating annual and seasonal Performance Yields (Y_R), Capacity Factor (CF), and Performance Ratio (PR). These coefficients take into account the impact of the variability of the available energy resource. Solar and wind power profiles are first correlated with power demand to calculate these ratios, so the residual load curves are used to supply the available renewable energy resource on a time scale in hours [23–25]. However, when using a deterministic model, calculating uncertainties is essential for obtaining information to validate, improve, and make decisions based on model predictions [26–28].

The existing literature extensively discusses the dichotomy between worst-case scenarios and expected value design approaches in system engineering, highlighting their respective impacts on safety, reliability, and cost-effectiveness [29,30]. However, a notable gap emerges when considering these design paradigms alongside uncertainty quantification techniques like bootstrapping for unknown density probability functions. The lack of exploration concerning how these contrasting design philosophies might interact with or influence the interpretation and utilization of uncertainty estimations derived through bootstrapping techniques is a significant gap in the literature [31–33]. This gap inhibits a comprehensive understanding of the potential influences and intersections of differing design strategies on decision-making processes in scenarios where estimating unknown probability distributions is fundamental, indicating a need for investigation and synthesis of these concepts within practical applications.

This research introduces an innovative approach to analyze a specific dataset measuring photovoltaic and wind energy potentials at the Universidad de La Salle campus in Bogotá, Colombia. The study aims to provide comprehensive insights into practical implications and potential outcomes in the renewable energy context. The proposed methodology combines K-means bootstrapping and Monte Carlo simulation to quantify uncertainties in K-means centroids estimation. Fig. 1 shows the main steps of this approach, which will be widely explained in the next section. The motivation for choosing the bootstrapped-based K-means data clustering technique is based on considering the advantages and limitations of the K-means clustering technique mentioned above. The emphasis on addressing uncertainties through bootstrapping and Monte Carlo simulation reflects a thoughtful strategy to overcome K-means limitations and enhance the reliability of clustering results, particularly in the specific context of analyzing renewable energy potentials.

The work is organized as follows: first, the bootstrapped-based K-means data clustering technique is used for identifying the optimal clusters using validation indexes such as Silhouette (S), Partition (SC), Xie and Beni (XB), and Separation (SI); then the PDF of the centroids are determined for estimating the technical feasibility of implementing a solar-wind hybrid system, through the energy resource available and corresponding expected values of the metrics CF, Y_R, and PR. Finally, the paper concludes with average data visualization and PDF plots for each centroid feature and performance indexes, followed by an analysis, conclusions, and future work.

2. Materials and methods

The method proposed involves finding underlying patterns in a dataset that contains samples from a weather station denoted as NOVALYNX 110-WS-18. Initially, this data is normalized, cleaned, and prepared. Then, the most relevant features are selected, and the basic statistical properties, such as maximum and minimum values, standard deviation and mean are calculated. Afterward, the bootstrap technique is applied to the original dataset to implement the unsupervised learning algorithm, where the validation indexes are required to evaluate the performance as well as determine the optimal number of clusters. Finally, a density estimation model computes the uncertainties of the performance ratios CF and Y_R through a Maximum Likelihood Estimator (MLE) and their most representative Gaussian families.

2.1. Data analysis

Our objective is to develop a data-driven approach for computing performance ratios. We assume sampled data independent and identically distributed, were generated by an unknown PDF. This PDF was estimated using a validation index to calculate the optimal centroids and represent it as a product of densities. It considered the scale and displacement parameters in a MLE framework. More detailed explanations related to these concepts can be reviewed in [34]. The following sections outline the general process for implementing an unsupervised machine learning algorithm that utilizes a data-driven model to determine the available energy potential.

2.1.1. K-means algorithm

The K-means algorithm is an unsupervised and non-hierarchical clustering method that does not require prior knowledge of the data [35]. Given a normalized set of samples $\mathbf{X}^t = \{\mathbf{x}_1^t, \dots, \mathbf{x}_n^t\}\ \forall\ t, \quad \mathbf{x}_i^t \in \mathbb{R}^d$, a specified number of clusters K, and a set of centroids $\mathbf{C}^t = \{\mathbf{c}_1^t, \dots, \mathbf{c}_K^t\}, \quad \mathbf{c}_k^t \in \mathbb{R}^d$, the algorithm assigns each data point in $\mathbb{R}^d$ to a cluster using an assignment rule based on a similarity measure. The Euclidean metric in a L_2 space is commonly used as the reference distance measure and is expressed in equation (1).

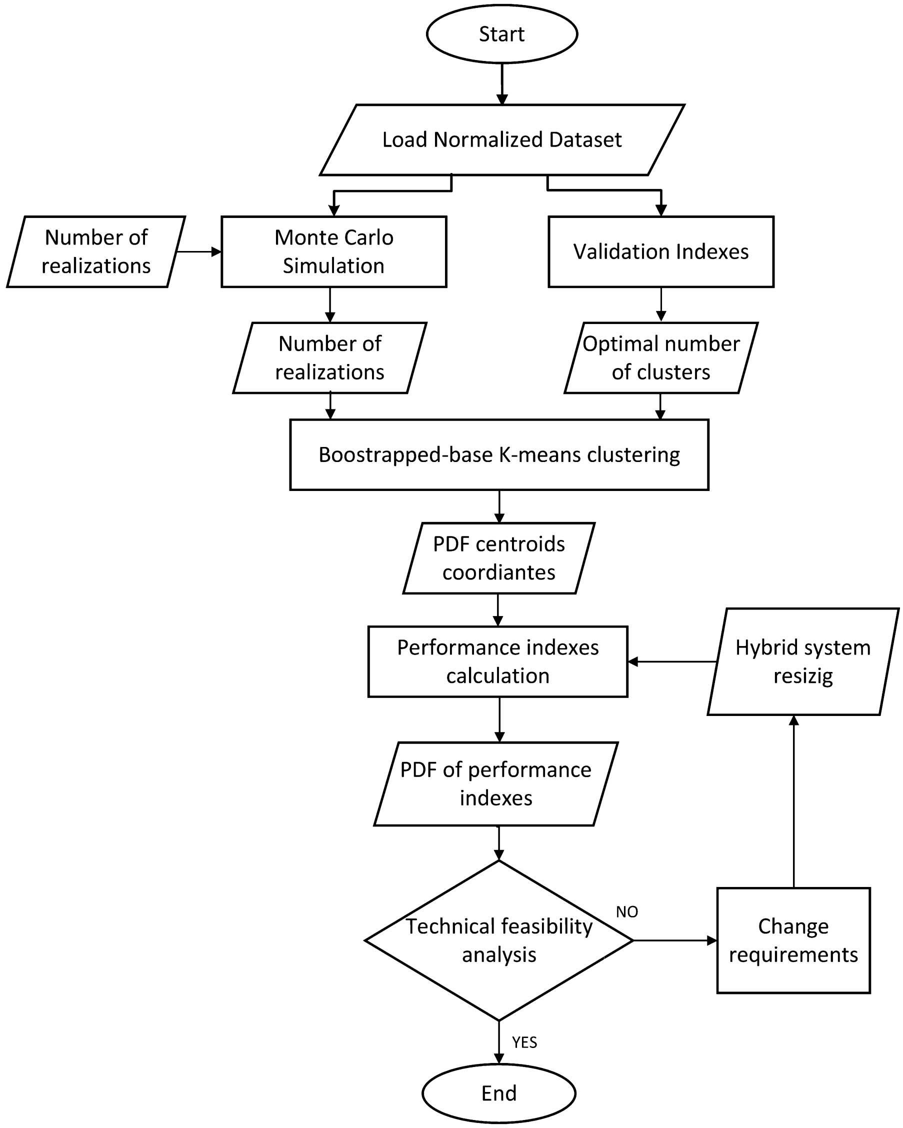


Fig. 1. Bootstrapped-based K-means data clustering method Flowchart.

$$dist(\mathbf{x}_i^t, \mathbf{c}_k^t) = \left|\left|\mathbf{x}_i^t - \mathbf{c}_k^t\right|\right| = \sqrt{\left(\mathbf{x}_i^t - \mathbf{c}_k^t\right)^2} \quad \forall \;\; i, k \tag{1}$$

The objective function J and the centroids $\mathbf{c}_k^t$ are optimized iteratively in the partition process, as shown in equations (2) and (3). The rule for assigning a data point $\mathbf{x}_i^t$ to a set $\mathcal{R}_k^t$ is based on minimizing the distance between the point and the centroid $dist(\mathbf{x}_i^t, \mathbf{c}_k^t)$. The iterative optimization process is an essential part of the K-means algorithm, as it minimizes the within-cluster sum of squares, which measures the distance between the data points and their assigned centroids. By minimizing this objective function, the algorithm aims to group similar data points in the same cluster and maximize the distances between the clusters.

$$J = \sum_{i=1}^{n} \sum_{k=1}^{K} \left\|\mathbf{x}_i^t - \mathbf{c}_k^t\right\|^2 \tag{2}$$

$$c_k^t = \frac{\sum_{x \in \mathcal{R}_k^t} x^t}{\|\mathcal{R}_k^t\|} \tag{3}$$

2.1.2. Validation indexes

To determine the best number of clusters for a dataset, is essential to consider how validation indexes are interpreted. These indexes are computed based on the average distance between data points and the centroid of their respective cluster, as well as the centroid of the nearest cluster, which is known as cohesion and separation. The decision on the right number of clusters ultimately rests with the researcher, as their ability to interpret these indexes correctly is vital. This research encompasses various indexes for cluster validation, each with its interpretation and graphical criteria to identify the optimal number of clusters K [36]. It's essential to recognize that the ideal cluster count isn't always straightforward and might differ based on the dataset and the problem. Thus, using multiple validation methods and expert insight is crucial for pinpointing the most suitable cluster count. Moreover, it's worth noting that the choice of clustering algorithm also significantly influences the optimal cluster count determination, given that different algorithms can yield distinct outcomes [12].

Multiple indicators exist to figure out the best cluster count for a dataset. Among them is the Silhouette Index S, which quantifies the difference between separation and cohesion values by dividing it by the larger of those two values. The optimal number of clusters corresponds to the first local maximum on the graph. Another index is the Partition Index, or SC, defined as the ratio between the sum of all cohesions and separations. In this case, the optimal number of clusters is the lowest slope on a graph. We use the Separation Index, or SI, which corresponds to the minimum separation distance to validate each partition. The optimal number of centroids can be established using a similar graphical interpretation used for the SC index. Finally, the Xie and Beni index, or XB, is obtained by dividing the total variance by the minimum separation of the clusters. The optimal number of centroids corresponds to the first local minimum on the graph.

It's worth noting that each of these indexes has its strengths and weaknesses, and the choice of which one to use may depend on the specific application. Additionally, other validation techniques could help to provide a more comprehensive understanding of the optimal number of clusters for $\mathcal{X}$ datasets.

2.2. Randomized K-means algorithm

When unsupervised machine learning models are deployed there exist many ideas and approaches which use data to reveal information and structural patterns that are unknown in principle. In general, a modeling technique is proposed without labeling the data output, afterward, an optimization is carried out to find the best parametric fit minimizing a metric, and finally, a validation is implemented including, hyperparameter optimization, cross-validation, residue analysis, and hypothesis testing to verify convergence, correlation, and parsimony. The main drawback of the method mentioned is the lack of uncertainty quantification when the method is performed without any knowledge about probability densities of the underlying phenomena, it implies that there are no, in principle, uncertainties measures in the results computed by the algorithms, to solve this situation we propose a randomized algorithm so-called bootstrapping which is a Monte Carlo-type method where the original data is treated as a pseudo-population, where sampling with replacement procedure is implemented obtaining new populations with equal probability where the clustering technique is deployed. On each of these realizations a K-means algorithm is performed, not getting a singleton set of centroids in a d-dimensional Euclidean space, but empirical probability distributions of centroids are achieved through this analysis, extracting new information as sampling means or variances about the centroids, also, confidence intervals leading to an uncertainty quantification about the energy resources for Solar-Wind Hybrid Systems

In this section, some theoretical tools and mathematical notation are depicted that formalize the practical results of this work. Let x_l^t be the l-th vector feature constructing a dataset $\mathbf{X}^t = \{\mathbf{x}_1^t, \dots, \mathbf{x}_n^t\}, \quad \mathbf{x}_i^t \in \mathbb{R}^d$, being $t = \{1, \dots, nb\}$ defining nb as the maximum number of bootstraps, $\mathbf{X}_0$ is the original dataset, taking $\mathbf{p} = \{x_1, \dots, x_d\}$ corresponds to an arbitrary d-dimensional point which is prone to be grouped into a certain set $\mathcal{R}_k,\ k = \{1, \dots, K\}$. The centroids of each dataset in the t-th bootstrap are defined as $\mathbf{c}_k^{t(\star)} \in \mathbb{R}^d$, with the l-th centroid coordinate $c_l^{t(\star)}$. Our approach considers each centroid component as a random variable empirically built by the bootstrap method, where the process estimation is performed considering a parametrized model adjusted by the data. This analysis relies on the assumption about independence and identically distributed centroid coordinates sampled of a Gaussian distribution, this premise is determined by its suitability for modeling continuous and variable processes [37]. In this analysis, we claim, based on a theoretical result named as Central Limit Theorem that if we construct a large enough sequence of nb bootstrapped datasets and each sample is i.i.d. (independent and identically distributed), the stable density achieved is a Gaussian probability density function with parameters μ and σ. Moreover, by the Weak Law of Large Numbers and the Strong Law of Large Numbers, we argue that by increasing the number of samples we converge in probability to μ for every ϵ following $\lim_{nb\to\infty} \mathbb{P}[|\mathbf{X}_{nb} - \mu| \geq \epsilon] = 0$, also, $\mathbb{P}[\lim_{nb\to\infty} \overline{\mathbf{X}}_{nb} = \mu] = 1$, where μ is the expected value of the random variable. Through this idea, as the number of samples is increased the tail probability tends to zero, therefore, this leads to statistically describe the dataset through its expected value, which is a necessary condition to approximate the empirical density distribution to Gaussian distribution [38,39]. Therefore, we define the following coordinate density model as follows,

$$g(c_l^{t(\star)} \mid \mu_l, \sigma_l) = \frac{1}{2\pi\sqrt{\sigma_l}} e^{-\frac{1}{2}\left(\frac{c_l^{t(\star)} - \mu_l}{\sigma_l}\right)^2}, \ \forall l = \{1, \dots, d\},\ \forall t = \{1, \dots, nb\} \tag{4}$$

which is read as the probability density function of the l-th coordinate given by mean and standard deviation assuming a likelihood function as a Gaussian density function. Moreover, an optimization framework is solved to find the best values that explain the centroids sample according to the chosen distribution family

$$\theta_k^{(\star)} = [\boldsymbol{\mu}_k^{(\star)}, \boldsymbol{\sigma}_k^{(\star)}] = argmax_{\theta_k} L(\theta_k \mid \boldsymbol{c}_k^{t(\star)}), \quad k = \{1, \dots, K\} \tag{5}$$

explicitly, each centroids sample is yielded by the bootstrap method where the likelihood function takes the form of,

$$\nabla_{\theta_k} \prod_{t=1}^{n_b} g(\mathbf{c}_k^{t(\star)} \mid \theta_k) = 0, \tag{6}$$

that is the analytic condition to obtain the maximum of this function, so-called maximum likelihood estimator, where ∇ stands for gradient operator which is a necessary condition for the parameters to achieve the maximum, it implies that a probability density function models the centroid distribution being the main goal to compute the parameters through optimization numerical algorithms. Since the estimation algorithm maximizes the likelihood function, the optimal parameters computed are interpreted as the best fit of the observed sample therefore they quantify the uncertainty of each component establishing a randomized interpretation of the K-means obtaining a variance parameter of each centroid coordinate. This density parameter is interpreted as an empirical confidence interval which is computed through a sequence of datasets artificially sampled without replacement yielding a randomized clustering process that is summarized in the Algorithm 1.

Algorithm 1 Randomized K-means.

procedure BOOTSTRAP K-MEANS FOR EMPIRICAL DISTRIBUTIONS OF CENTROIDS($\mathcal{X}$)
Collection of sampled datasets $\mathcal{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_{nb}\}, \quad \mathbf{X}^t \in \mathbb{R}^{d\times n}$.
Collection of points $\mathbf{X}^t = \{\mathbf{x}_1^t, \dots, \mathbf{x}_n^t\}, \quad \mathbf{x}_i^t \in \mathbb{R}^d$.
Number of clusters K.
Initial centers $\mathbf{C}^t = \{\mathbf{c}_1^t, \dots, \mathbf{c}_K^t\}, \quad \mathbf{c}_k^t \in \mathbb{R}^d$.
Initial cluster set $\{\mathcal{R}_1, \dots, \mathcal{R}_K\} \leftarrow \{\emptyset\}$.
  for $t = 1 : nb$ **do**
    while a stopping criteria is not met **do**
      for $i = 1 : n$ **do** ▷ K-means stage
        $\mathbf{d} \leftarrow [\mathrm{dist}(\mathbf{x}_i^t, \mathbf{c}_1^t), \dots, \mathrm{dist}(\mathbf{x}_i^t, \mathbf{c}_K^t)]$
        $k \leftarrow \mathrm{argmin}_j d_j$
        $\mathcal{R}_k^t \leftarrow \mathcal{R}_k \cup \mathbf{x}_i^t$
      for $k = 1 : K$ **do**
        $\mathbf{c}_k^t = \frac{\sum_{x\in\mathcal{R}_k^t} \boldsymbol{x}^t}{\|\mathcal{R}_k^t\|}$
  $\{\mathbf{c}_k^{t(\star)}\}, \{\mathcal{R}_k^{t(\star)}\}$. ▷ K-means output
  Select the optimal centroid coordinate ▷ MLE Stage
  $\{\mathbf{c}_{kl}^{t(\star)}\}_{l=1}^d = \{c_{k1}^{t(\star)}, \dots, c_{kd}^{t(\star)}\}$
  Gaussian paramaters $\{\theta_l\}_{l=1}^d = \{\mu_l, \sigma_l\}_{l=1}^d$
  while a stopping criteria is not met **do**
    for $k = 1 : K$ **do**
      for $l = 1 : d$ **do**
        $L(\theta_l \mid c_{kl}^{t(\star)}) = \prod_{t=1}^{nb} g(c_{kl}^{t(\star)} \mid \theta_l)$
        $\theta_l^\star = \mathrm{argmax}_\theta L(\theta_l, c_{kl}^{t(\star)})$
  Return $\{\theta_k^{(\star)}\}_{k=1}^K$

The algorithm shown briefly depicts the method proposed in this article, offering a procedure to compute a measure of the dispersion of the outputs of the K-means algorithm, leading, with this new information, an improvement of the planning and design process of hybrid generation of electrical systems as microgrids, distributed generation systems, and renewable energy systems.

2.3. Performance metrics for hybrid systems

This subsection shows the different steps considered for each of the $\mathbf{X}^t \in \mathbb{R}^{d\times n}$ datasets, which allow the calculation of the performance ratios for the available energy resources by means of a K-means-based algorithm.

2.3.1. Available solar energy

The emission of energy from the sun in the form of electromagnetic radiation is a physical phenomenon that can be measured using units of irradiance [Wh/m^2] [12]. This type of energy travels through space and reaches the Earth's surface at different angles and directions depending on the location's latitude and longitude. Solar radiation is the amount of energy received per unit area on a surface [Wh/m^2], and it can vary around 1000 [Wh/m^2] [40] over the Earth's surface. The equation that describes the relationship between irradiance I and daily radiation $Rad(t)$ is provided by equation (7):

$$I = \int_{t_1}^{t_2} Rad(t)dt \quad \left[\frac{\text{Wh}}{\text{m}^2\text{day}}\right] \tag{7}$$

The given equation is a definite integral of power density over some time between t_1 and t_2, which provides the total energy generated by the solar resource. Likewise, $Rad(t)$ corresponds to a second-degree polynomial equation (8) [41], where t is in $[h]$. The calculation also takes into account the dimensions of the solar panels (width a_{panel} and length b_{panel}) as well as the number of solar panels N_{PV}, which altogether determine the daily photovoltaic energy E_{PV} as presented in the equation (9):

$$Rad(t) = a_{PV}t^2 + b_{PV}t + c_{PV} \quad \left[\frac{\text{W}}{\text{m}^2\text{day}}\right] \tag{8}$$

$$E_{PV} = A_{PV} \cdot I = N_{PV} \cdot a_{panel} \cdot b_{panel} \cdot I \quad \left[\frac{\text{Wh}}{\text{m}^2\text{day}}\right] \tag{9}$$

2.3.2. Available wind energy

The air density ρ is obtained directly through the air mass and volume or indirectly through factors such as temperature, atmospheric pressure, and altitude, which play a crucial role in assessing the wind potential of a given location [42]. In addition, the wind power generated by a turbine at any given moment $P_w(t)$ is directly proportional to the cubed wind speed $v^3(t)$ and the surface area of the turbine blades A_w, as demonstrated in equation (10).

$$P_w(t) = \frac{1}{2} \cdot \rho \cdot A_w \cdot v^3 \cdot C_p \quad [\text{W}] \tag{10}$$

Wind power density (*WPD*) is a key indicator used to classify regions based on their wind power potential, calculated by dividing wind power by surface area A_w, as shown in equation (11) [43]. This is a fundamental starting point in evaluating the amount of wind energy available daily, denoted by E_W, and it is calculated using equation (12), which incorporates the radius of the turbine blades r_w and the number of turbines N_w. The maximum amount of theoretical energy obtained from the wind requires the Power Coefficient C_p, and it is limited to the Betz limit [44].

$$WPD(t) = \frac{P_w}{A_w} = \frac{1}{2} \cdot \rho \cdot v^3(t) \cdot C_p \quad \left[\frac{\text{W}}{\text{m}^2}\right] \tag{11}$$

$$E_w = A_w \int_{t_1}^{t_2} WPD(t)dt = N_w \cdot \pi \cdot r_w^2 \cdot \int_{t_1}^{t_2} WPD(t)dt \quad \left[\frac{\text{Wh}}{\text{day}}\right] \tag{12}$$

2.3.3. Capacity factor (*CF*):

This calculated by dividing the energy generated during a daily period by the maximum amount of energy $E_N = P_N(t_2 - t_1)$ [Wh/day]. The specific time is determined by two independent variables, t_1 and t_2, with the overall time interval denoted as t. A feasible range of values for the *CF* depends on the available energy resource or the technology considered. For solar and wind energy, the typical range interval for *CF* is [25, 35] [%] and [25, 45] [%], respectively, denoted as CF_{PV} and CF_w in equations (13) and (14) [25]. The values for E_{PV} and E_W used in these equations are obtained from equations (9) and (12).

$$CF_{PV}[\%] = \left[\frac{E_{PV}}{E_{NPV}}\right] \times 100[\%] \tag{13}$$

$$CF_w[\%] = \left[\frac{E_W}{E_{Nw}}\right] \times 100[\%] \tag{14}$$

2.3.4. Performance yields (Y_R)

The index mentioned here is the ratio of the total energy density H_t [Wh/(m^{2}day)] (7) and (12), considering a reference power density G_0 [Wh/(m^{2}day)] that varies depending on the available energy resource. This ratio corresponds to equation (15), with $G_0 = 1000$ [Wh/(m^{2}day)] for a photovoltaic system, and $G_0 = (1/2) \cdot \rho \cdot {v_{nom}}^3$ [Wh/(m^{2}day)] for a wind system, where v_{nom} represents the nominal speed of the turbine in [m/s] [45,46].

$$Y_R = \frac{H_t}{G_0} \quad [\text{h}] \tag{15}$$

2.4. Performance ratio (PR)

This factor assesses the actual performance of an energy system relative to its ideal operation. In other words, it indicates how realistic the energy system is at converting input energy into effective output energy. This factor corresponds to the ratio between the actual performance, represented by Y_f, and a reference value, denoted as Y_r, which is the ideal performance. The units in numerator and denominator are energy density [Wh/(m^{2}day)], so this metric is given in the equation (16) [47]:

$$PR[\%] = \frac{Y_f}{Y_r} \times 100[\%] \tag{16}$$

This section outlines a suggested approach for examining formulae and ratios used to measure the amount of energy generated by wind and solar resources. By utilizing these equations, we can assess the available renewable energy and determine the most effective engineering strategies to implement in the specific study case.

2.5. Uncertainty quantification and deviation

An important aspect to be emphasized in this article is the definition of a procedure to estimate centroid uncertainties computed by the randomized K-means algorithm. This measures performance index uncertainties in the estimation scheme for hybrid energy systems in any selected geographic location. In this work, we assume that computed centroids are sampled from a continuous Gaussian distribution, which is reasoned through the Law of Large Numbers, so, an empirical probability distribution is defined through the bootstrap and MLE techniques, estimating the mean and variance of each lth-coordinate for the kth-centroid. Subsequently, with these expressions, a fully probabilistic approach to studying the energy performance of hybrid systems is completed leading to more detailed studies like large deviation bounds, tail distributions, and algorithm convergence.

At this point, a question arises, What is the probability that any coordinate centroid deviates about its mean? To analyze this problem some basic elements of probability theory are defined in the following lines, easing the understanding of the procedure proposed. First, we use an important result known as *moment generating function* of a continuous random variable [48], which corresponds to the coordinate centroid c presented here as an event of this variable Y. That is, $\{\mathbf{c} \in \Omega | Y(\mathbf{c}) = c\}$, where Ω is the sample space. Therefore,

$$M_Y(s) = \mathbb{E}\{e^{sY}\},$$

where $\mathbb{E}$ is the expected operator, yielding a function that depends on the distribution parameters and a new variable s that, in general, is well-chosen. The importance of this expression lies in the ability to compute expected values, variances, or tail bounds of generalized random variables using Markov's, Chebyshev's, or Chernoff inequalities, which measure the probability of deviation about the expected value of the random variable.

Let the density $Y \in \mathcal{N}(\mu, \sigma)$, be a Gaussian random variable, then, the moment generating function follows,

$$M_Y(s) = e^{\left(\mu s + \frac{\sigma^2 s^2}{2}\right)}$$

Therefore, by establishing a Chernoff probability bound we achieve

$$\forall \delta > 0 \quad P\{Y \geq ((1+\delta)\mu)\} \leq \min_{s\in\mathbb{R}} \frac{\mathbb{E}\{e^{sY}\}}{e^{s((1+\delta)\mu)}}$$

$$\Rightarrow P\{Y \geq ((1+\delta)\mu)\} \leq \min_{s\in\mathbb{R}} \frac{e^{\left(\mu s + \frac{\sigma^2 s^2}{2}\right)}}{e^{(1+\delta)\mu s}}$$

The right-hand side minimizes an expression with decision variable s, $\min_{s\in\mathbb{R}} f(s)$, likewise, δ value determines the precise deviation to be analyzed, therefore,

$$\frac{df}{ds} = f(s)(\sigma^2 s - \delta\mu)$$

obtaining the critical point and solving for s

$$\frac{df}{ds} = 0 \Rightarrow s = \frac{\delta\mu}{\sigma^2}.$$

Finally,

$$\forall \delta > 0 \quad P\{Y \geq ((1+\delta)\mu)\} \leq e^{\left(\frac{\delta^2\mu^2}{2\sigma^2} - \frac{\delta^2\mu^2}{\sigma^2}\right)} = e^{-\left(\frac{\delta^2\mu^2}{\sigma^2}\right)}$$

which is the tight bound for a Gaussian random variable obtained through the Chernoff bound for each coordinate centroid in the Randomized K-means. It means, that analysis beyond the mean values of the performance indexes could potentially be proposed, considering probabilities deviations about the density mean driven by the δ value. For instance, a worst-case or very optimistic scenario with high or low probability can represent extreme climate phenomena in the data analysis proposed.

3. Results

Employing solar-wind hybrid systems with strategies like K-means bootstrapping and Monte Carlo simulation aims to quantify the uncertainties in centroids estimation. The engineering background of this application is focused on renewable energy forecasting, statistical analysis in energy harnessing, and the implementation of data-driven methodologies in the renewable energy research field [49–51]. The main aspects of this approach are listed below:

- The dataset, denoted as $\mathbf{X}_0$, was collected at La Salle University Campus in Bogotá, Colombia. This information is crucial for understanding the geographical context and the specific environmental conditions that it represents. This dataset links the research to a real-world location, making it relevant for regional applications.
- There are four distinct features included in the dataset $\mathbf{X}_0$, namely, Time t [h], Solar Radiation Rad [W/m^2], Cubed Wind Speed v^3 [m^3/s^3], and Temperature T [°C]. These features represent important environmental variables that can be relevant for various engineering applications, such as renewable energy, weather forecasting, and building design.
- The data samples were collected at two different times, from 2:30 to 6:40, between December 13th, 2018, and March 9th, 2020, and they are available in [52]. This information provides insights into the temporal scope of the dataset, which is critical for understanding the seasonality and time-dependent characteristics of the data.
- Cubed wind speed feature, denoted as v^3, is a modified variable, and its value is directly proportional to the WPD. This modification highlights the engineering relevance of the dataset, as wind energy is a crucial aspect of renewable energy systems and environmental monitoring. The relationship between wind power and energy appears in equations (10), (11), and (12).
- The dataset with 56807 observations is transformed using the Minmax normalization, ensuring that data from different sources or periods can be compared and analyzed consistently.

Table 1
Daily normalized and denormalized averaged centroids coordinates obtained by K-means.

Normalized Features	t [-]	Rad [-]	v^3 [-]	T [-]
Normalized Centroids Coordinates	1.87×10^{-1} 5.24×10^{-1} 8.35×10^{-1}	1.87×10^{-2} 3.35×10^{-1} 1.47×10^{-2}	3.20×10^{-3} 3.89×10^{-2} 1.03×10^{-2}	3.11×10^{-1} 6.17×10^{-1} 4.36×10^{-1}
Denormalized Features	t [h]	Rad [W/m^2]	v^3 [m^3/s^2]	T [°C]
Real Centroids Coordinates	4.44 12.47 19.87	19.00 348.36 15.71	2.04 24.78 6.53	12.63 17.95 14.83

Table 2
Diary averaged polynomial coefficients for $Rad(t)$, $v^3(t)$ y T(t).

Centroids / Feature	a	b	c
$Rad(t)$ [W/m^2]	-5.58	135.30	-471.82
$v^3(t)$ [m^3/s^3]	-3.45×10^{-1}	8.64	-29.56
T(t) [°C]	-7.03×10^{-2}	1.85	5.80

Table 3
Descriptive statistics for each of the features of the original dataset $\mathbf{X}_0$.

Features	t [h]	Rad [W/m^2]	v^3 [m^3/s^3]	T [°C]
Minimum Value	0.00	0.00	0.00	6.88
Maximum Value	23.83	1059.44	640.503	25.00
Mean	11.90	116.04	10.29	14.94
Standard Deviation	6.93	184.93	28.36	2.73

3.1. Representative scenarios

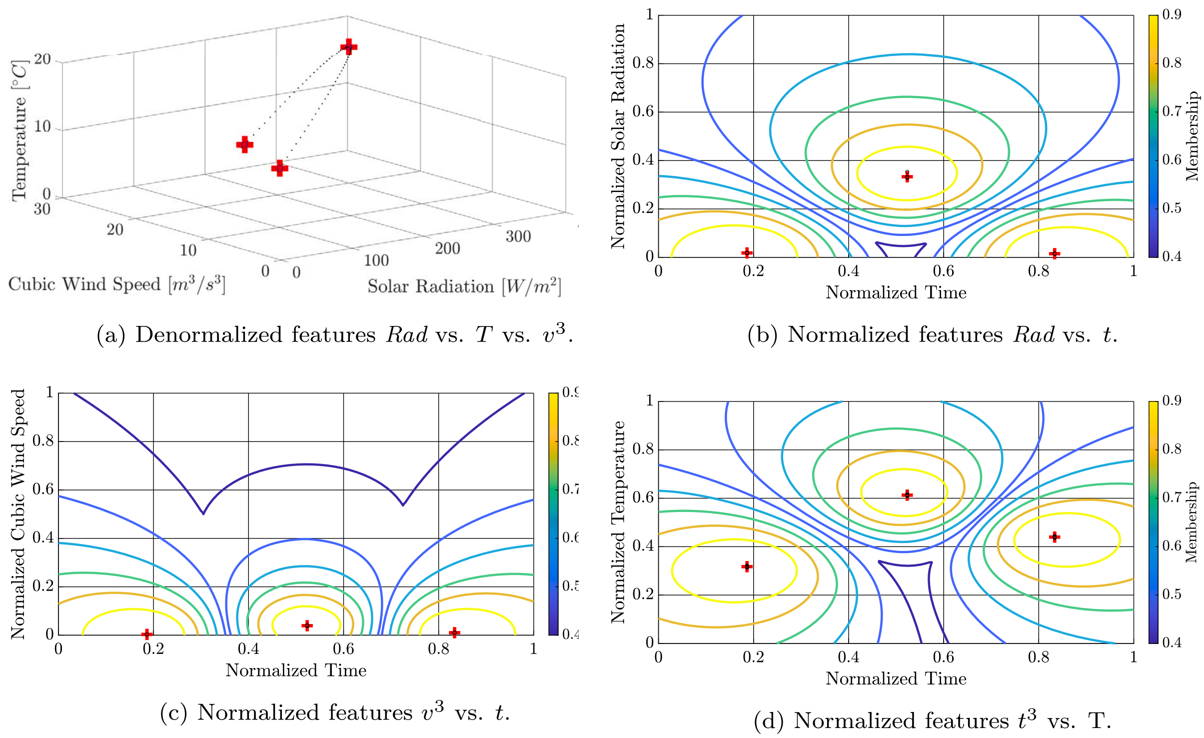
The bootstrap-based K-means algorithm is executed on the normalized dataset $\mathbf{X}_m$. Afterward, the centroids obtained using MLE are denormalized, and only their average values are represented daily and shown in Table 1. The graphical analysis of the validation indexes S, SI, XB and SC reveals that the optimal number of clusters K corresponds to 3.

The process for obtaining the polynomial coefficients for three features: radiation, velocity, and temperature, involves using the averaged real centroids from Table 1 and treating time as an independent variable for each feature. Equations (8), (17), and (18) show the time as a parametric variable in the calculations. The resulting polynomial coefficients are in Table 2. Table 3 provides information on the descriptive statics, which corresponds to the mean and the standard deviation, of the variables analyzed from the original dataset $\mathbf{X}_0$.

$$v^3(t) = a_w t^2 + b_w t + c_w \quad \left[\frac{\text{m}^3}{\text{s}^3}\right] \tag{17}$$

$$\text{T}(t) = a_T t^2 + b_T t + c_T \quad [°\text{C}] \tag{18}$$

On the other hand, the parametric fitted curve based on the features $Rad(t)$, $v^3(t)$ y T(t), is obtained the Fig. 2a, for $t \in [0,\ 24)$ [h]. Form



(a) Denormalized features Rad vs. T vs. v^3.

(b) Normalized features Rad vs. t.

(c) Normalized features v^3 vs. t.

(d) Normalized features t^3 vs. T.

Fig. 2. Daily averaged features in (a) $\mathbb{R}^3$ and (b), (c) and (d) $\mathbb{R}^2$.

Table 4
Daily PDF of centroids obtained by means of bootstrapped-based K-means in R^4.

Feature \ Centroids	$P(\mathbf{c}_1 \mid [\boldsymbol{\mu}_1, \boldsymbol{\sigma}_1])$	$P(\mathbf{c}_2 \mid [\boldsymbol{\mu}_2, \boldsymbol{\sigma}_2])$	$P(\mathbf{c}_3 \mid [\boldsymbol{\mu}_3, \boldsymbol{\sigma}_3])$
t [h]	$\mathcal{N}(4.44, 1.49\times10^{-1})$	$\mathcal{N}(12.47, 1.55\times10^{-1})$	$\mathcal{N}(19.87, 1.77\times10^{-1})$
Rad [W/m^2]	$\mathcal{N}(19.00, 6.50\times10^{-1})$	$\mathcal{N}(348.36, 1.61)$	$\mathcal{N}(15.71, 7.45\times10^{-1})$
v^3 [m^3/s^3]	$\mathcal{N}(2.04, 2.93\times10^{-1})$	$\mathcal{N}(24.78, 5.89\times10^{-1})$	$\mathcal{N}(6.53, 3.93\times10^{-1})$
T [°C]	$\mathcal{N}(12.63, 1.04\times10^{-1})$	$\mathcal{N}(17.95, 1.42\times10^{-1})$	$\mathcal{N}(14.83, 1.44\times10^{-1})$

the averaged normalized features in $\mathbb{R}^3$ t, v^3 y Rad are depicted in Fig. 2, for $K = 3$. The $\mathbb{R}^2$-dimensional planes Rad vs. t, v^3 vs. t and T vs. t, show the level sets which represents the cluster similarity and membership, where the colormap corresponds to a level membership interval between [0.4, 0.9], showed in the Fig. 2b, Fig. 2c and Fig. 2d, respectively. The coordinates of the centroids, in conjunction with their corresponding daily PDF, are shown in the Table 4.

3.2. Centroids probabilities densities

In this subsection, the different probability density functions, considering the 20 most representatives Gaussian families and $N_b = 4000$, obtained for each of the features $x_1 = t$, $x_2 = Rad$, $x_3 = v^3$ and $x_4 = T$ in subfigures (a), (b), (c), (d), respectively. It is shown in Figs. 3, 4 and 5, which correspond to the centroids $\mathbf{c}_1$, $\mathbf{c}_2$ and $\mathbf{c}_3$ and they can be expressed in the compact form $P(c_j \mid [\boldsymbol{\mu}_j, \boldsymbol{\sigma}_j])$. Likewise, their averaged parameters are obtained by means of MLE and they are depicted in Table 4.

3.3. Performance analysis

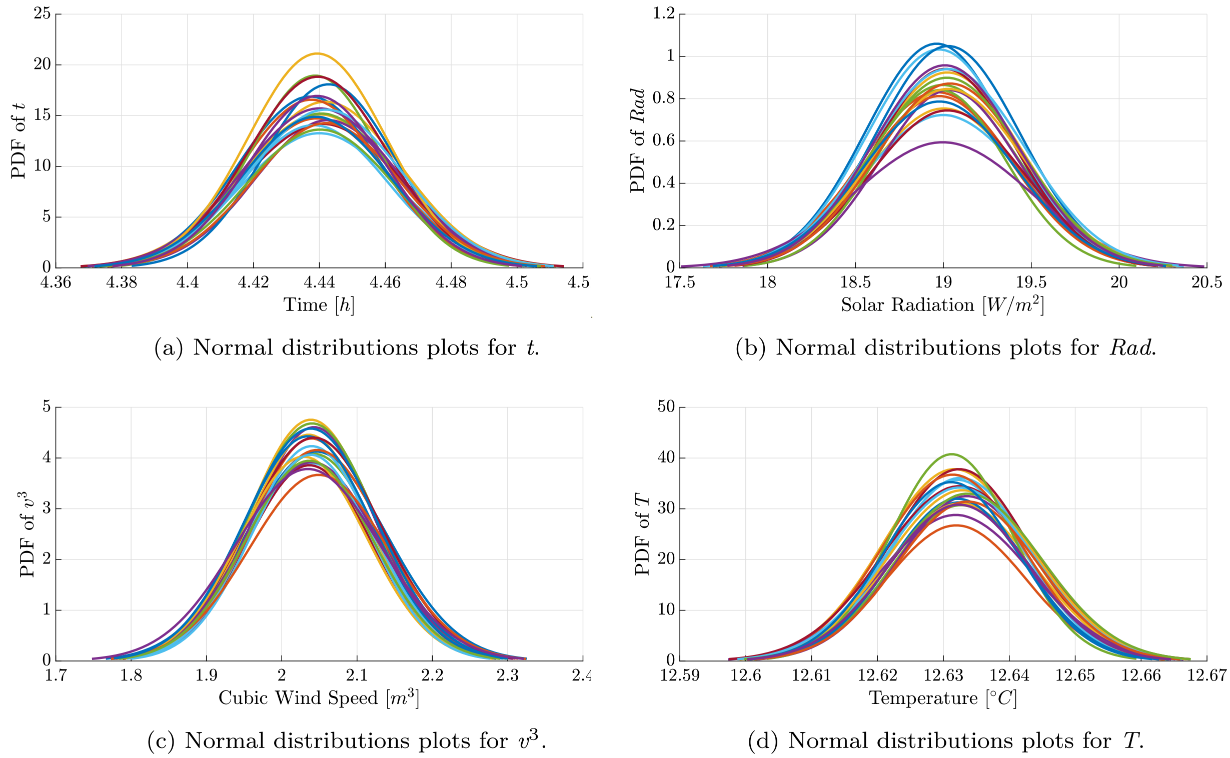
To compute the total daily available energy in the load E_N [Wh/day], according to the energy resources, at La Salle University campus, we calculate $E_{NPV} = 116668.57$ [Wh/day] [23], moreover, potential wind energy corresponds to $E_{Nw} = P_{Nw} \cdot (t_2 - t_1)$ [Wh/day], where $P_{Nw} = 100.00$ [W]. Then, considering the parameterized functions $Rad(t)$ [W/m^2] and $v^3(t)$ [m^3/s^3], they allow to obtain the energy to supply at the time interval $[t_1,\ t_2]$ [h]. For this purpose, it is important to clarify that 6 polycrystalline 320.00 [W] solar panels are considered, with $a_{panel} = 1.96$ [m] and $b_{panel} = 0.99$ [m], and a wind microturbine model WSC100 of 100 [W], $r_w = 0.26$ [m] for the on-grid system.

Table 5
Daily PDF of coefficients CF [%] and Y_R [h] obtained for each energy resource.

Available Energy Resource	$P(CF \mid \mu_{CF}, \sigma_{CF})$	$P(Y_R \mid \mu_{Y_R}, \sigma_{Y_R})$
Photovoltaic	$\mathcal{N}(36.74, 4.70\times10^{-1})$	$\mathcal{N}(3.68, 1.49\times10^{-1})$
Wind	$\mathcal{N}(2.15, 1.73\times10^{-1})$	$\mathcal{N}(1.66\times10^{-1}, 4.50\times10^{-2})$

The MLE results in Table 5 show that the proposed photovoltaic technological solution is feasible, and the wind solution is not, which is consistent with previous studies conducted in Bogotá that show the potential for solar and wind energy [12,23]. Wind energy is primarily utilized in rural areas, while in urban areas, vertical axis wind turbines with a startup speed for winds below 1.5 [m/s] are required. The results for Y_R [h] indicate the availability of the reference power density G_0 in equivalent hours and its proximity to the ideal solution for each energy resource. Finally, Fig. 6 illustrates the performance ratios for the most representatives Gaussian families being expressed as $P(CF \mid \mu_{CF}, \sigma_{CF})$ and $P(Y_R \mid \mu_{Y_R}, \sigma_{Y_R})$.

Table 5 presents the results for $P(CF \mid \mu_{CF}, \sigma_{CF})$ and $P(Y_R \mid \mu_{Y_R}, \sigma_{Y_R})$ of various photovoltaic and wind-solar resources. These results indicate that the chosen wind turbine is not suitable for implementation due a poor wind performance. Furthermore, the probabil-



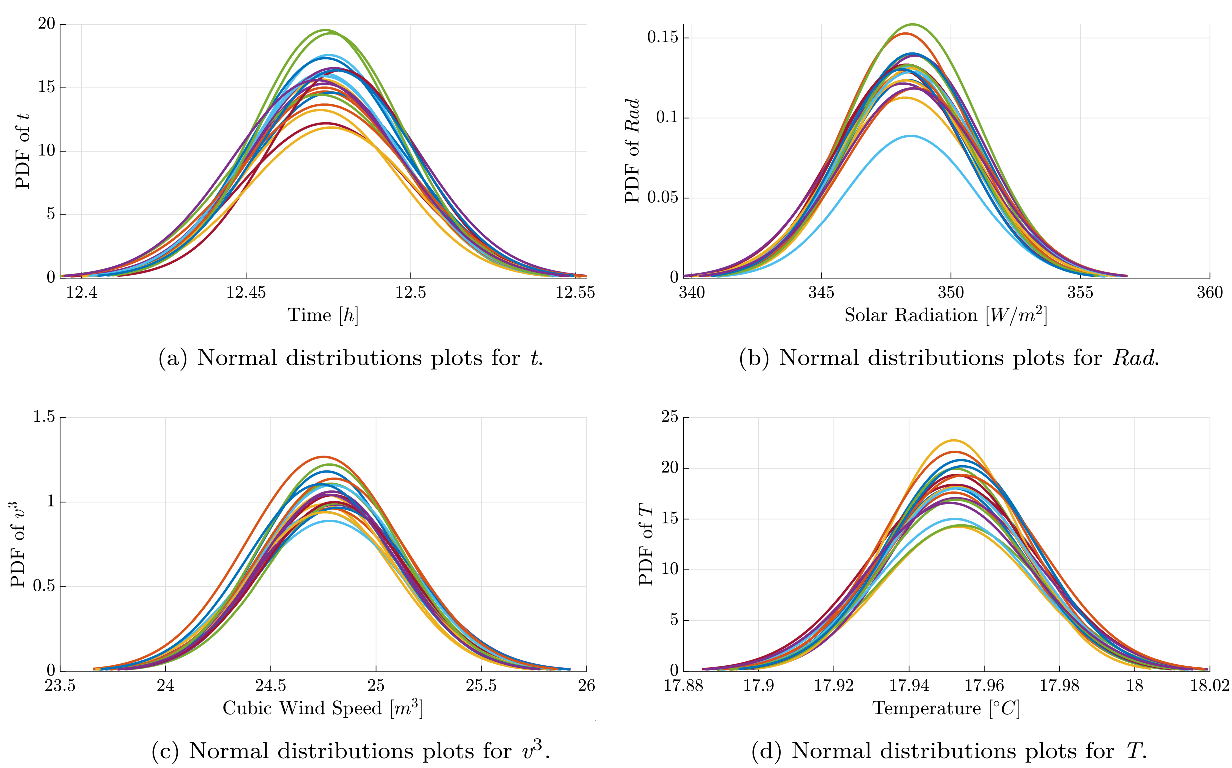
(a) Normal distributions plots for t.

(b) Normal distributions plots for Rad.

(c) Normal distributions plots for v^3.

(d) Normal distributions plots for T.

Fig. 3. Combined distributions plots for $\mathbf{c}_1 \in \mathbb{R}^4$.



(a) Normal distributions plots for t.

(b) Normal distributions plots for Rad.

(c) Normal distributions plots for v^3.

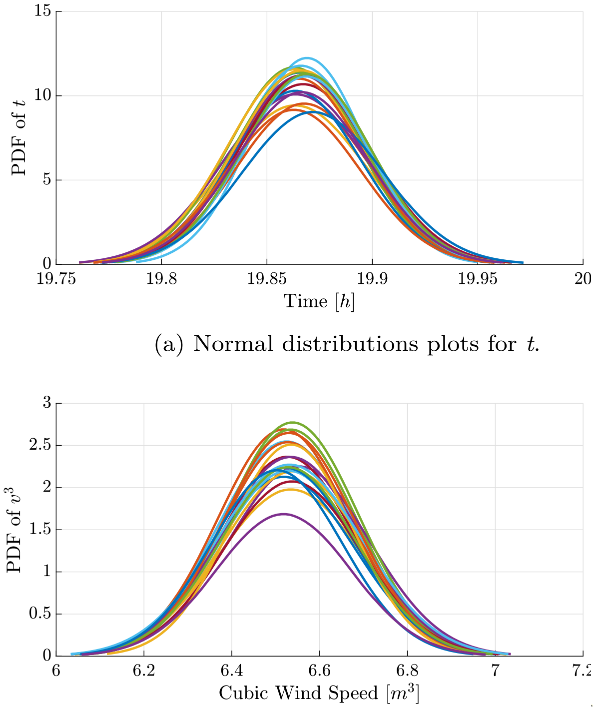
(d) Normal distributions plots for T.

Fig. 4. Combined distributions plots for $\mathbf{c}_2 \in \mathbb{R}^4$.

ity density function for this technology shows similar low magnitudes for mean and standard deviation, which suggests only the feasibility of the PV design [53]. Fig. 7 shows the 20 most representatives Gaussian families $p(PR \mid \mu_{PR}, \sigma_{PR})$ for the solar resource. Finally, the parameters obtained for the PDF of PR_{PV}, which corresponds to $\mathcal{N}(63.33, 3.89 \times 10^{-1})$, and they exhibit a realistic operation scenario.

4. Discussion

In the previous section, the findings suggest that the campus of La Salle University can technically implement a photovoltaic system, but not a wind energy system due to the lack of low-speed turbine technology. The PDF of the *CF* for both photovoltaic and wind systems depends



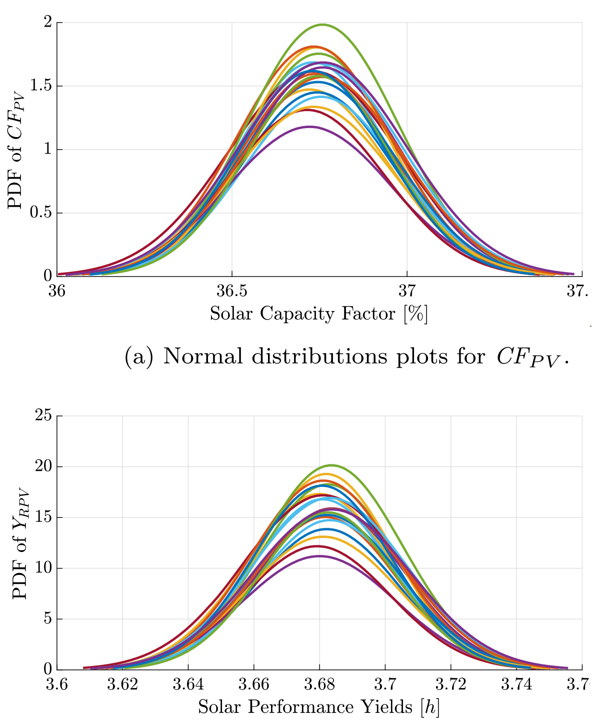
(a) Normal distributions plots for t.

(b) Normal distributions plots for Rad.

(c) Normal distributions plots for v^3.

(d) Normal distributions plots for T.

Fig. 5. Combined distributions plots for $\mathbf{c}_3 \in \mathbb{R}^4$.



(a) Normal distributions plots for CF_{PV}.

(b) Normal distributions plots for CF_w.

(c) Normal distributions plots for Y_{RPV}.

(d) Normal distributions plots for Y_{Rw}.

Fig. 6. Combined distributions plots for daily performance ratios CF [%] and Y_R $[h]$.

on the technological solution proposed. The system's flexibility is related to its autonomy, and a hybrid on-grid or off-grid system design is necessary as μ_{CF} increases or decreases. The expected *CF* approaching 1 means that the auxiliary storage device sizing increases, which affects economic feasibility. To implement energy solutions in the output power range, a low value of the *CF* for wind turbines requires commercial wind microturbine models with nominal power ratings below 100 [W] and startup speeds below 1.5 [m/s]. The probability density function of Y_R determines the availability of the reference power density G_0, which corresponds to the average daily solar brightness distribution provided by the interactive atlas of the Colombian Institute of Hydrology, Meteorology, and Environmental Studies (IDEAM). There

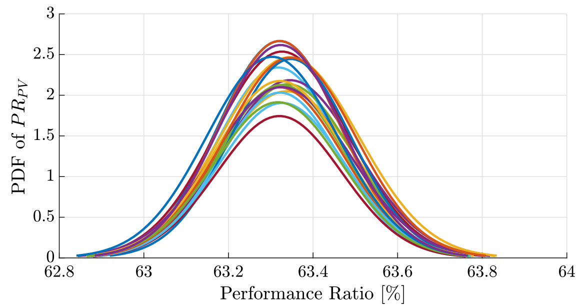


Fig. 7. Gaussian families plots for PR_{PV}.

is currently no information available about wind resource distribution. The PR provides additional information on how effectively these proposed energy solutions can perform under ideal conditions. Finally, the *PR* is a measure of the actual output of a system in comparison to its maximum possible output. This metric is useful in evaluating renewable energy systems, where the weather and location can impact performance. A high *PR* indicates that the system is operating close to an ideal scenario, while a low *PR* suggests a real one.

5. Conclusions and future work

This work proposes a method for designing a solar-wind hybrid system based on expected value instead of the worst-case scenario, which has the potential to provide more accurate predictions for energy output. A probabilistic model allows the incorporation of new local features into the satellite data, such as microclimates or terrain variations. Additionally, this probabilistic model allows updates over time as more data becomes available, providing ongoing optimization for the system's performance. In other words, this approach improves the accuracy and efficiency of renewable energy production and increases the system's stability.

Capacity Factor *CF* is a metric that calculates a renewable energy system's actual output compared to its maximum output, providing insight into its effectiveness. Meanwhile, Performance Yields (Y_R) measures the current energy produced by the system during a specific period. Evaluating these metrics enables determining the system's performance and identifying areas for improvement. A high energy Performance Ratio *PR* can help reduce costs per unit of energy produced, lowering the renewable energy system's overall costs, and increasing its attractiveness for investment opportunities. Additionally, renewable energy systems with a high PR can contribute to meeting climate targets by reducing greenhouse gas emissions. Therefore, by considering these metrics, it is possible to evaluate the potential success and optimization of implementing proposed energy solutions, which are more realistic due to the integration of different available energy resources.

Obtaining uncertainties in renewable systems is crucial for several reasons: Firstly, it can enhance the design and operation of the system by improving its reliability. Secondly, it can assist experts in making informed decisions such as system components, topology, and auxiliary storage device options. Thirdly, assessing these uncertainties can help mitigate risks and boost confidence in the system's performance, optimizing its design for cost and performance. Likewise, the systems' uncertainties are related to their feasibility, including their economic viability, technical feasibility, and regulatory compliance. By assessing these uncertainties, renewable energy systems can be optimized for improved performance, leading to the development of sustainable and more efficient energy solutions. It is worth noting that the uncertainty calculation is complex and requires specialized knowledge, making it essential to have experts in renewable energy systems.

Combining K-means clustering with bootstrapping and Monte Carlo simulation offers a comprehensive awareness to discern uncertainties in the centroids estimation, yet it confronts several limitations. The computational demands, sensitivity to parameter choices, and dependence on specific assumptions limit its practicality and universal applicability. Future research works to overcome these drawbacks should focus on refining computational efficiency through optimized algorithms or parallel processing, enhancing parameter robustness to diversify applicability, relaxing underlying assumptions, conducting comparative studies for validation, improving interpretability via visual tools, and striving for scalability to enable real-time implementation on extensive datasets. This method promotes a more resilient and universally applicable approach, providing deeper insights into uncertainties within K-means centroids estimation.

Funding

This research was funded by Vicerrectoría de Investigación y Transferencia (VRIT), at Universidad de La Salle, Bogotá, Colombia and Universidad Santo Tomás, under institutional codes IELE221-211 and BOG-2023-AI015, respectively. The APC was funded by Universidad Santo Tomás, Bogotá D.C., Colombia.

CRediT authorship contribution statement

Fabian Salazar-Caceres: Conceptualization, Methodology; Formal Analysis, Writing-Original Draft Preparation; **Harrynson Ramirez Murillo:** Conceptualization, Software, Formal Analysis, Data Curation, Writing-Original Draft Preparation; **Carlos Andrés Torres-Pinzón:** Methodology, Validation, Investigation. Writing-Review and Editing, Funding Acquisition; **Martha Patricia Camargo-Martínez:** Software, Validation, Investigation, Data Curation, Writing-Review and Editing. All authors have read and agreed to the published version of the manuscript.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Carlos Andrés Torres-Pinzón reports financial support and article publishing charges were provided by Universidad Santo Tomás. Fabian Salazar-Caceres reports financial support was provided by Universidad de La Salle. Martha Patricia Camargo-Martínez reports financial support was provided by Universidad de La Salle. Harrynson Ramirez-Murillo reports financial support was provided by Universidad de La Salle.

Data availability

Datasets related to this article can be found at https://github.com/Ibrahim-Jayyam/La-Salle-Dataset-Energy-Estimation.git, hosted at GitHub [52].

References

[1] A. Ikotun, A. Ezugwu, L. Abualigah, B. Abuhaija, J. Heming, K-means clustering algorithms: a comprehensive review, variants analysis, and advances in the era of big data, Inf. Sci. 622 (2023) 178–210, https://doi.org/10.1016/j.ins.2022.11.139.

[2] S. Miraftabzadeh, C. Colombo, M. Longo, F. Foiadelli, K-means and alternative clustering methods in modern power systems, IEEE Access 11 (2023) 119596–119633, https://doi.org/10.1109/ACCESS.2023.3327640.

[3] H. Cui, S. Dong, J. Hu, M. Chen, B. Hou, J. Zhang, B. Zhang, J. Xian, F. Chen, A hybrid MCDM model with Monte Carlo simulation to improve decision-making stability and reliability, Inf. Sci. 647 (2023) 119439, https://doi.org/10.1016/j.ins.2023.119439.

[4] Y. Li, X. Chu, D. Tian, J. Feng, W. Mu, Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm, Appl. Soft Comput. 113 (2021) 107924, https://doi.org/10.1016/j.asoc.2021.107924.

[5] F. Zheng, Q. Liu, Anomalous telecom customer behavior detection and clustering analysis based on ISP's operating data, IEEE Access 8 (2020) 42734–42748, https://doi.org/10.1109/ACCESS.2020.2976898.

[6] V. Kumar, P. Madhuri, R. Yamini, P. Sravani, M. K, Improvement of task scheduling with energy efficiency including fault tolerant with resilient computing system in parallel and distributed communication network, in: 2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), 2023, pp. 872–877.
[7] A. Chandrasekar, T. Radhika, Q. Zhu, Further results on input-to-state stability of stochastic Cohen–Grossberg BAM neural networks with probabilistic time-varying delays, Neural Process. Lett. 54 (2022) 613–635, https://doi.org/10.1007/s11063-021-10649-w.
[8] T. Radhika, A. Chandrasekar, V. Vijayakumar, Q. Zhu, Analysis of Markovian jump stochastic Cohen-Grossberg BAM neural networks with time delays for exponential input-to-state stability, Neural Process. Lett. 55 (2023) 11055–11072, https://doi.org/10.1007/s11063-023-11364-4.
[9] Y. Cao, A. Chandrasekar, T. Radhika, V. Vijayakumar, Input-to-state stability of stochastic Markovian jump genetic regulatory networks, Math. Comput. Simul. (2023), https://doi.org/10.1016/j.matcom.2023.08.007.
[10] M. Jabi, M. Pedersoli, A. Mitiche, I. Ayed, Deep clustering: on the link between discriminative models and K-means, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2021) 1887–1896, https://doi.org/10.1109/TPAMI.2019.2962683.
[11] G. Grigoras, F. Scarlatache, An assessment of the renewable energy potential using a clustering-based data mining method. Case study in Romania, Energy 81 (2015) 416–429, https://doi.org/10.1016/j.energy.2014.12.054.
[12] H. Ramírez-Murillo, C. Torres-Pinzón, E. Forero-García, Estimación del potencial fotovoltaico mediante minería de datos en cuatro ciudades de Colombia, TecnoLógicas 22 (2019) 77–97, https://doi.org/10.22430/22565337.1345.
[13] A. Walch, R. Castello, N. Mohajeri, J. Scartezzini, Big data mining for the estimation of hourly rooftop photovoltaic potential and its uncertainty, Appl. Energy 262 (2020) 114404, https://doi.org/10.1016/j.apenergy.2019.114404.
[14] H. Habbak, M. Mahmoud, K. Metwally, M. Fouda, M. Ibrahem, Load forecasting techniques and their applications in smart grids, Energies 16 (2023), https://doi.org/10.3390/en16031480.
[15] M. Ansari, C. Guo, M. Shaikh, N. Chopra, B. Yang, J. Pan, Y. Zhu, X. Huang, Considering the uncertainty of hydrothermal wind and solar-based DG, Alex. Eng. J. 59 (2020) 4211–4236, https://doi.org/10.1016/j.aej.2020.07.026.
[16] L. Casula, G. D'Amico, G. Masala, P. Filippo, Performance estimation of photovoltaic energy production, Lett. Spat. Resour. Sci. 13 (2020) 267–285, https://doi.org/10.1007/s12076-020-00258-x.
[17] J. Wang, Y. Gao, An integrated forecasting system based on knee-based multi-objective optimization for solar radiation interval forecasting, Expert Syst. Appl. 198 (2022) 116934, https://doi.org/10.1016/j.eswa.2022.116934.
[18] N. Begum, N. Badshah, L. Rada, A. Ademaj, M. Ashfaq, H. Atta, An improved multi-modal joint segmentation and registration model based on Bhattacharyya distance measure, Alex. Eng. J. 61 (2022) 12353–12365, https://doi.org/10.1016/j.aej.2022.06.018.
[19] M. Journée, C. Bertrand, Improving the spatio-temporal distribution of surface solar radiation data by merging ground and satellite measurements, Remote Sens. Environ. 114 (2010) 2692–2704, https://doi.org/10.1016/j.rse.2010.06.010.
[20] M. Farooq, S. Kumar, An assessment of renewable energy potential for electricity generation in Pakistan, Renew. Sustain. Energy Rev. 20 (2013) 240–254, https://doi.org/10.1016/j.rser.2012.09.042.
[21] O. Shepovalova, Y. Arbuzov, V. Evdokimov, P. Ilyushin, K. Suslov, Assessment of the Gross, technical and economic potential of regions and solar energy for photovoltaic energetics, Energies 16 (2023), https://doi.org/10.3390/en16031262.
[22] L. Hunt, P. Kipouros, Energy demand and energy efficiency in developing countries, Energies 16 (2023), https://doi.org/10.3390/en16031056.
[23] H. Ramirez Murillo, F. Salazar-Cáceres, M. Camargo-Martínez, A. Patiño-Forero, F. Méndez-Casallas, Energy performance clustering and data visualization for solar-wind hybrid energy systems, in: Applied Computer Sciences in Engineering, 2022, pp. 77–89.
[24] Z. Usman, J. Tah, H. Abanda, C. Nche, A critical appraisal of PV-systems' performance, Buildings 10 (2020), https://doi.org/10.3390/buildings10110192.
[25] A. Ershad, R. Brecha, K. Hallinan, Analysis of solar photovoltaic and wind power potential in Afghanistan, Renew. Energy 85 (2016) 445–453, https://doi.org/10.1016/j.renene.2015.06.067.
[26] M. Borunda, A. Ramírez, R. Garduno, G. Ruíz, S. Hernandez, O. Jaramillo, Photovoltaic power generation forecasting for regional assessment using machine learning, Energies 15 (2022), https://doi.org/10.3390/en15238895.
[27] C. Tu, W. Tsai, C. Hong, W. Lin, Short-term solar power forecasting via general regression neural network with grey wolf optimization, Energies 15 (2022), https://doi.org/10.3390/en15186624.
[28] J. Thaker, R. Höller, A comparative study of time series forecasting of solar energy based on irradiance classification, Energies 15 (2022), https://doi.org/10.3390/en15082837.
[29] F. Gutiérrez-García, A. Arcos-Vargas, A. Gómez-Expósito, Robustness of electricity systems with nearly 100% share of renewables: a worst-case study, Renew. Sustain. Energy Rev. 155 (2022) 111932, https://doi.org/10.1016/j.rser.2021.111932.
[30] X. Deng, Y. Gao, J. Zhang, Beyond the worst-case analysis of random priority: smoothed and average-case approximation ratios in mechanism design, Inf. Comput. 285 (2022) 104920, https://doi.org/10.1016/j.ic.2022.104920.
[31] J. Polo, N. Martín-Chivelet, M. Alonso-Abella, C. Sanz-Saiz, J. Cuenca, M. Cruz, Exploring the PV power forecasting at building facades using gradient boosting methods, Energies 16 (2023), https://doi.org/10.3390/en16031495.
[32] W. Sulandari, Y. Yudhanto, P. Rodrigues, The use of singular spectrum analysis and K-means clustering-based bootstrap to improve multistep ahead load forecasting, Energies 15 (2022), https://doi.org/10.3390/en15165838.
[33] Y. Li, A. Liu, S. Wang, Y. Zhan, J. Chen, H. Hsiao, A study of total-factor energy efficiency for regional sustainable development in China: an application of bootstrapped DEA and clustering approach, Energies 15 (2022), https://doi.org/10.3390/en15093093.
[34] A. Sharma, K. Boroevich, D. Shigemizu, Y. Kamatani, M. Kubo, T. Tsunoda, Hierarchical maximum likelihood clustering approach, IEEE Trans. Biomed. Eng. 64 (2017) 112–122, https://doi.org/10.1109/TBME.2016.2542212.
[35] E. Looney, Z. Liu, A. Classen, H. Liu, N. Riedel, M. Braga, P. Balaji, A. Augusto, T. Buonassisi, I. Marius Peters, Representative identification of spectra and environments (RISE) using k-means, Prog. Photovolt. 29 (2021) 200–211, https://doi.org/10.1002/pip.3358.
[36] K. Sinaga, M. Yang, Unsupervised K-means clustering algorithm, IEEE Access 8 (2020) 80716–80727, https://doi.org/10.1109/ACCESS.2020.2988796.
[37] I. Vayansky, S. Kumar, A review of topic modeling methods, Inf. Sci. 94 (2020) 101582, https://doi.org/10.1016/j.is.2020.101582.
[38] V. Chernozhukov, D. Chetverikov, K. Kato, Y. Koike, High-dimensional data bootstrap, Annu. Rev. Stat. Appl. 10 (2023) 427–449, https://doi.org/10.1146/annurev-statistics-040120-022239.
[39] S. Kim, J. Hur, A probabilistic modeling based on Monte Carlo simulation of wind powered EV charging stations for steady-states security analysis, Energies 13 (2020), https://doi.org/10.3390/en13205260.
[40] T. Myasnikova, A. Kirillova, S. Ivanova, O. Sveklova, O. Nadezhdina, Simulation of solar energy photovoltaic conversion, in: 2020 International Youth Conference on Radio Electronics, Electrical and Power Engineering (REEPE), 2020, pp. 1–4.
[41] J. Fan, L. Wu, F. Zhang, H. Cai, X. Ma, H. Bai, Evaluation and development of empirical models for estimating daily and monthly mean daily diffuse horizontal solar radiation for different climatic regions of China, Renew. Sustain. Energy Rev. 105 (2019) 168–186, https://doi.org/10.1016/j.rser.2019.01.040.
[42] K. Thapa, K. Jayasawal, Pitch control scheme for rapid active power control of a PMSG-based wind power plant, IEEE Trans. Ind. Appl. 56 (2020) 6756–6766, https://doi.org/10.1109/TIA.2020.3015169.
[43] S. Carreno-Madinabeitia, G. Ibarra-Berastegi, J. Sáenz, A. Ulazia, Long-term changes in offshore wind power density and wind turbine capacity factor in the Iberian Peninsula (1900–2010), Energy 226 (2021) 120364, https://doi.org/10.1016/j.energy.2021.120364.
[44] J. West, S. Lele, Wind turbine performance in very large wind farms: Betz analysis revisited, Energies 13 (2020), https://doi.org/10.3390/en13051078.
[45] J. El Hacen, R. Ihaddadene, N. Ihaddadene, C. Elhadji Sidi, M. EL Bah, P. Logerais, Performance analysis of micro-amorphe silicon PV array under actual climatic conditions in Nouakchott, Mauritania, in: 2019 10th International Renewable Energy Congress (IREC), 2019, pp. 1–6.
[46] L. Arribas, L. Cano, I. Cruz, M. Mata, E. Llobet, PV–wind hybrid system performance: a new approach and a case study, Renew. Energy 35 (2010) 128–137, https://doi.org/10.1016/j.renene.2009.07.002.
[47] N. Reich, B. Mueller, A. Armbruster, W. Sark, K. Kiefer, C. Reise, Performance ratio revisited: is PR 90% realistic?, Prog. Photovolt. 20 (2012) 717–726, https://doi.org/10.1002/pip.1219.
[48] M. Mitzenmacher, E. Upfal, Probability and Computing: Randomized Algorithms and Probabilistic Analysis, Cambridge University Press, 2005.
[49] H. Wang, Z. Lei, X. Zhang, B. Zhou, J. Peng, A review of deep learning for renewable energy forecasting, Energy Convers. Manag. 198 (2019) 111799, https://doi.org/10.1016/j.enconman.2019.111799.
[50] J. Li, J. Zhou, B. Chen, Review of wind power scenario generation methods for optimal operation of renewable energy systems, Appl. Energy 280 (2020) 115992, https://doi.org/10.1016/j.apenergy.2020.115992.
[51] B. Grillone, S. Danov, A. Sumper, J. Cipriano, G. Mor, A review of deterministic and data-driven methods to quantify energy efficiency savings and to predict retrofitting scenarios in buildings, Renew. Sustain. Energy Rev. 131 (2020) 110027, https://doi.org/10.1016/j.rser.2020.110027.
[52] Dataset hosted at github, https://github.com/Ibrahim-Jayyam/La-Salle-Dataset-Energy-Estimation.git, 2023.
[53] J. Olaniran, Estimation of diurnal patterns of global solar radiation, temperature, relative humidity and wind speed from daily datasets at a humid tropical location, Agric. For. Meteorol. 322 (2022) 109003, https://doi.org/10.1016/j.agrformet.2022.109003.