POLITECNICO DI TORINO
Repository ISTITUZIONALE

Efficient Methods for Modeling Correlations Among Renewable Energy Sources in State-Space Sampling Monte Carlo Simulation

Original
Efficient Methods for Modeling Correlations Among Renewable Energy Sources in State-Space Sampling Monte Carlo Simulation / Borges, C.L.T., Vaz, M.O., Santos, A.S., Fontenele, G., Souza, R.C., Carvalho, C.E., Russo, A.. - In: ENERGIES. - ISSN 1996-1073. - 19:13(2026). [10.3390/en19133163]

Availability:
This version is available at: 11583/3017462 since: 2026-11-03T11:53:34Z

Publisher:
Multidisciplinary Digital Publishing Institute (MDPI)

Published
DOI:10.3390/en19133163

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

26 November 2026

energies

MDPI

Article

Efficient Methods for Modeling Correlations Among Renewable Energy Sources in State-Space Sampling Monte Carlo Simulation

Carmen L. T. Borges [1,*], Mateus O. Vaz [1], Andressa S. Santos [1], Gonçalo Fontenele [1], Roberta C. Souza [2], Claudio E. Carvalho [3] and Angela Russo [4]

[1] Electrical Engineering Department, Federal University of Rio de Janeiro, Rio de Janeiro 21941-909, Brazil; mateus.vaz@poli.ufrj.br (M.O.V.); andressasdossantos4@gmail.com (A.S.S.); goncalogfb@poli.ufrj.br (G.F.)
[2] State Grid Brazil Holding, Rio de Janeiro 20071-004, Brazil; roberta.souza@stategrid.com.br
[3] Instituto ABRATE de Energia, Brasilia 70711-902, Brazil; claudio.elias@abrate.org.br
[4] Energy Department, Politecnico di Torino, 10129 Torino, Italy; angela.russo@polito.it
* Correspondence: carmen@nacad.ufrj.br

Abstract

This paper compares two combined methods for modeling correlations among multiple random variables in the state-space sampling approach that is adopted in non-sequential Monte Carlo Simulation. The first method combines the Rank Score (Iman and Conover method) with Latin Hypercube sampling, while the second method combines Kernel Density Estimation with Bayesian Networks to represent dependency between variables. The methods are compared in terms of goodness-of-fit and computational efficiency, using statistical metrics, spatial correlation matrices, and scatter plots. The multiple time series were aggregated into single ones (Total and Net Generation) in order to validate the accuracy of each model in representing multi-dimensional correlated variables. The methods were evaluated for 3 large-dimensional cases based on the actual Brazilian energy system, with 10, 19 and 28 time series of 5-, 3- and 2-year horizons, respectively. The results show that the statistical dependencies of the historical data were accurately captured by both models, with comparable goodness-of-fit, but a higher computational efficiency of the first.

Keywords: rankscore method; Latin Hypercube sampling; kernel density estimation; Bayesian Networks; state-space sampling; non-sequential Monte Carlo simulation

1. Introduction

The main objective of an electric power system is to deliver electricity to consumers at the lowest possible cost while maintaining the required levels of reliability, power quality, and operational security. Nevertheless, the stochastic nature of equipment and component failures can jeopardize system performance, making reliability assessment essential for supporting investment decisions and minimizing both the likelihood and consequences of service interruptions. Renewable energy sources pose significant challenges for system planning and operation. In the absence of storage, their output is entirely dependent on primary resources, such as wind speed, solar irradiation, and water inflows, which are inherently variable. As their share in the energy mix grows, sudden fluctuations in these resources can threaten the reliability of the power supply.

Reliability assessment of large-scale power systems is commonly performed using Monte Carlo simulation (MCS), which generates system operating states by considering component outage probabilities together with the variability of generation and load conditions. While correlations are inherently captured in sequential (chronological) MCS,

Check for updates

Academic Editor: Juri Belikov

Received: 21 May 2026
Revised: 23 June 2026
Accepted: 30 June 2026
Published: 3 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

they are not in non-sequential MCS, which adopts the state-space sampling approach. Because sequential simulations are far more computationally intensive, there is a need for methods that incorporate correlation into non-sequential MCS, enabling similar accuracy with significantly reduced computational effort.

With the integration of renewable energy, whose output strongly depends on resource availability, it becomes essential to account for correlations among different sources and between variable generation and load. Independent state-space sampling alone cannot adequately capture the complex variability and coincidence of power output from these sources.

The primary motivation of this paper is to compare methods that allow non-sequential MCS to maintain computational efficiency while achieving the accuracy of sequential MCS when modeling correlated variables. The authors have been conducting a large study on methods that can be used for this purpose.

In [1], a non-parametric approach based on Kernel Density Estimation (KDE) was introduced to approximate the probability density functions (PDFs) and cumulative distribution functions (CDFs) of stochastic variables derived from historical time series. Following aggregation and normalization, the resulting data were used to train a Bayesian Network (BN) capable of representing the dependence structure among the variables. Non-sequential MCS was subsequently carried out by drawing samples from the trained BN. After reversing the data transformations, the generated samples retained the correlation patterns present in the original series and were applied to compute composite reliability indices in a modified RTS system, yielding results comparable in accuracy to sequential MCS while significantly reducing computational effort.

In [2], a similar BN-based approach was used to represent the dependencies among wind speed, solar irradiation, and load, where the corresponding time series were generated using an autoregressive moving average (ARMA) model. The samples obtained from the BN were evaluated against those generated using C-vine copulas, considering both their statistical consistency with the original data and their performance in probabilistic power flow analysis on a modified IEEE 118-bus system.

A closely related approach was presented in [3], where a BN was again employed to model the nonlinear dependencies among wind speed, solar irradiation, and load. The generated samples were compared with those obtained from C-vine copulas and an enhanced Latin Hypercube Sampling (L-Mapm) method. The comparison included scatter plots, distribution fitting, and probabilistic power flow analyses conducted on a modified IEEE 69-bus distribution system.

Unlike previous studies, ref. [4] employed a Dynamic BN (DBN), which explicitly incorporates temporal information and captures the evolution of system variables over time, including their autocorrelation. The method was applied to model the dependence between wind generation and load, followed by a sequential MCS based on the trained DBN. The resulting samples were statistically compared with those generated using a Hidden Markov Model (HMM), demonstrating the capability of DBNs to represent temporal dependencies in renewable energy applications.

In [5], results related to reliability of supply using BNs are presented. The BN was trained using five months of solar radiation and wind speed measurements, while load demand was represented using data from the Reliability Test System (RTS). The network was subsequently used to calculate the Loss of Load Probability (LOLP) with respect of a given node of the power network. However, the validation was limited to a relatively simple system configuration and a short measurement period, not fully evaluating the scalability and applicability of the proposed methodology.

In [6], a reliability assessment framework based on KDE to model the stochastic behavior of renewable energy sources in generation adequacy studies is proposed. The approach also employs K-means clustering to identify characteristic renewable seasons and then derive PDFs for each hourly-seasonal cluster. These models are then integrated into a non-sequential MCS framework to evaluate adequacy indices such as LOLP, LOLE, and LOEE. The study validates the proposed framework against sequential MCS, showing that the KDE-based approach achieves comparable accuracy with significantly lower computational burden.

In [7], Principal Component Analysis (PCA) was incorporated into a non-sequential Monte Carlo simulation framework for power system reliability assessment. The proposed methodology combined PCA with covariance analysis, random variable transformations, and correlation mapping to capture the dependence between load demand and renewable generation. The case study examined four geographically dispersed wind farms and demonstrated that the approach produced accurate reliability indices while reducing computational time by a factor of approximately 40 compared with sequential MCS.

The authors in [8] developed a model to more accurately represent the output of a hybrid solar–wind power plant in reliability studies, incorporating seasonal and spatial characteristics, as well as the correlation between the two energy sources. The correlation issue was addressed using copula theory and copula selection was made by comparing the distribution of historical data with typical copula shapes. For the case study, the Clayton copula was selected because the relationship between wind and solar generation was highly asymmetric with a pronounced lower tail. Copula parameters were estimated through maximum likelihood. Reliability analysis was conducted using the sequential cross-entropy method. The consideration of correlation reduced reliability indices, since the inherent complementarity between the two sources is beneficial to meet load demand.

In [9], Hierarchical Archimedean Copulas (HACs) were employed to model the dependence among intermittent renewable energy sources in hybrid power plants for reliability assessment using non-sequential MCS. Gumbel and Clayton copulas were used to represent wind–wind and solar–solar correlations, respectively, while the Frank copula modeled the dependence between wind and solar generation. The proposed approach produced reliability indices comparable to those obtained with sequential MCS while significantly reducing computational effort.

In [10], Cholesky decomposition combined with the Nataf transformation was proposed to incorporate wind speed correlation into Monte Carlo-based reliability studies. The methodology was evaluated using four commonly adopted wind speed distributions, demonstrating that conventional correlation matrix approaches tend to underestimate reliability indices. This work was later extended in [11] to include the correlation between load, wind, and solar generation, with each variable represented by an appropriate probability distribution.

In [12], the correlation among small hydropower plants was investigated within a non-sequential Monte Carlo simulation framework for power system reliability evaluation. The proposed approach was based on the Nataf transformation, where the modified correlation matrix was derived using numerical integration combined with root-finding procedures. By relying exclusively on empirical cumulative distribution functions instead of assuming predefined probability models, the method produced reliability indices with accuracy comparable to that of sequential simulation.

In [13], a set of correlation modeling approaches—including load duration curves, linear regression, rank correlation methods and joint probability models—were evaluated for non-sequential MCS in reliability studies involving renewable energy systems. The inclusion of LHS reduced computational effort, while all methods produced reliability indices with acceptable agreement relative to sequential MCS.

Three scenario-generation techniques were compared in [14]: conventional MCS with Cholesky decomposition, LHS with rank correlation, and copula-based modeling. The copula analysis was limited to the Gumbel and Gaussian copulas. Results indicated that, for geographically compact regions, LHS with rank correlation more accurately represented wind generation data, whereas copulas better captured wind speed dependence. For larger geographical areas, both approaches exhibited consistently good performance.

Although some of these references do not directly deal with reliability evaluation, they propose probabilistic models for risk analysis and, therefore, provide a background and can be adapted to model the dependency of renewable generation in Monte Carlo simulation for reliability evaluation.

In [15], four techniques for incorporating the dependence among intermittent renewable generation sources into non-sequential Monte Carlo simulation (MCS) were evaluated for power system reliability assessment. The methods included two correlation matrix-based approaches—Nataf transformation and the Iman–Conover rank score (RS) method combined with Latin Hypercube Sampling (LHS)—as well as two copula-based multivariate models, namely Vine Copulas and Hierarchical Archimedean Copulas (HACs). Comparative analyses highlighted the strengths and limitations of each approach, providing guidance on their suitability for different datasets and reliability assessment scenarios.

Most reliability studies that assess the impact of renewable energy sources rely on sequential MCS. However, as the penetration of these sources increases and power systems grow in complexity, this type of simulation can become computationally prohibitive. Many works investigate ways to better model wind, PV, or hydro generation, but relatively few focus specifically on adequacy analysis. Because the choice of modeling approach is highly dependent on the application, comparing different methods within the adequacy context becomes particularly relevant. Although some studies introduce methodologies that appear efficient for calculating reliability indices, comprehensive evaluations that support the selection of the most appropriate approach remain scarce.

For instance, ref. [15] compared Copula-based methods—rapidly gaining popularity—with more traditional correlation matrix approaches. That study showed that Rank Score combined with LHS was a very promising non-sequential MCS technique for reliability studies, offering considerable accuracy with high computational efficiency. Similarly, ref. [1] presented an innovative methodology that integrates KDE with Bayesian Networks. This approach could handle multiple variables with simulation times comparable to independent state sampling, while maintaining the accuracy of sequential simulation. Despite its promise, this method has not yet been systematically confronted with other alternatives.

The main contribution of this paper is to address that research gap, by broadening the understanding of the advantages and limitations of different methodologies for reliability studies, and providing insights into the trade-offs involved in their application. In particular, this paper compares the methods that showed the best prospects in terms of efficiency and accuracy among those previously evaluated: KDE with BN and Rank Score with LHS.

The comparison is done in terms of goodness-of-fit of the two methods for three large-scale energy systems of the Brazilian Northeast (NE) and Southeast (SE) regions: Case 1 – Mix of renewable generation of NE region with 10 time series; Case 2 – Solar generation of NE and SE regions with 19 time series; Case 3 – Mix of renewable generation of NE and SE regions with 28 time series.

The remainder of this paper is structured as follows. The next section briefly outlines the key theoretical foundations of the two methods considered. Section 3 presents the obtained results and provides a comparative analysis of both approaches. Finally, Section 4 discusses the conclusions and summarizes the main findings and considerations.

2. Materialsand Methods

In reliability evaluation of power systems using the conventional non-sequential MCS, system states are defined as combinations of the states of all components, under the assumption that these components are statistically independent. For a system with m components, a state can be represented by the vector $\underline{x} = [x_1, x_2 \dots x_k \dots x_m]$, where x_k represents the state of the kth component. Defining $P(x_k)$ as the probability associated with the kth component, the probability of the ith state $P(\underline{x}^i) = p_i$ is given by [1]:

$$P(\underline{x}^i) = \prod_{k=1}^{m} P(x_k) \tag{1}$$

However, the introduction of statistically dependent variables into the problem implies the appearance of joint probabilities in the equation. For w statistically dependent variables of the system and $P(v_1, v_2, \dots, v_w)$ being its respective joint probability distribution, the probability $P(\underline{x}^i)$ becomes:

$$P(\underline{x}^i) = P(v_1, v_2, \dots, v_w) \cdot \prod_{k=1}^{m} P(x_k) \tag{2}$$

Thus, the use of non-sequential MCS to assess the reliability of a system with statistically dependent elements requires the incorporation of the information contained in its joint probabilities. One possible approach is the decomposition of the joint probability distribution into conditional probability distribution functions by applying the chain rule as given below:

$$P(v_1, v_2, \dots, v_w) = P(v_1) \cdot P(v_2|v_1) \cdot P(v_3|v_2, v_1) \dots P(v_w|v_{w-1}, \dots, v_2, v_1) \tag{3}$$

These conditional probabilities could be built and used in the non-sequential MCS algorithm to generate samples for the statistically dependent variables. One of the approaches to characterize these functions is the use of parametric distributions, in which the random variables are associated to a specific family of distributions, e.g., Weibull, log-normal, etc. However, this parametric strategy has several drawbacks, such as the fact that the chosen distribution does not necessarily describe all variables of the same type in the system. Additionally, increasing the number of represented variables tends to cause the number of model parameters to grow exponentially, which is usually called the "curse of dimensionality" [1]. The alternative approach is the use of non-parametric methods to characterize the joint probabilities, and to this aim two methods will be presented in this paper: Kernel Density Estimation and Bayesian Network, and Rank Score and Latin Hypercube.

2.1. Kernel Density Estimation and Bayesian Network

The kernel density estimation employs kernel functions that give weights to the region around the data points, and each of these contributions can be averaged to estimate the probability density [16]. To prevent artificial discontinuities in the estimated function, a smooth kernel can be chosen and a common choice is the Gaussian kernel described by the following equation:

$$p(x) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{(2\pi h^2)^{\frac{1}{2}}} e^{-\frac{||x - x_n||^2}{2h^2}} \tag{4}$$

where x_n represents the nth data point of a total of N, and h is the smoothing parameter given in this case by the standard deviation of the Gaussian components.

The KDE approach described above estimates the marginal probability distributions of individual random variables. However, reliability studies require the estimation of

the joint PDF, which characterizes the multidimensional dependence among variables. Consequently, preserving the nonlinear statistical relationships between variables is essential, as these dependencies are quantified by Mutual Information (MI). Originating from Information Theory, MI measures the amount of shared information between variables and is equal to zero only when they are statistically independent [1]. Accordingly, the joint PDF can be decomposed into two components: the marginal distributions, estimated using KDE, and the dependence structure, represented through the mutual information shared among the variables.

Modeling the mutual information among a large number of variables is challenging because of the curse of dimensionality. To address this issue, the statistical dependencies are represented using BN. A Bayesian Network combines probability theory and graph theory to model complex dependency structures through a directed acyclic graph, providing an efficient and compact representation of the joint probability distribution of n-dimensional random variables. For a graph with K nodes, the joint distribution is given by [16]:

$$p(x_1, x_2, \ldots, x_K) = \prod_{k=1}^{K} p(x_k | pa_k) \tag{5}$$

where pa_k denotes the set of parent nodes of x_k.

Although this work adopts BN to effectively capture probabilistic dependencies among renewable energy sources, it is worth noting that graph-based modeling has evolved to capture even more complex and high-order relationships across various domains. For instance, advanced hypergraph architectures have been recently proposed to optimize reasoning and explicitly represent complex semantic constraints in knowledge graphs [17]. This highlights the versatility of graph-theoretic approaches in dealing with multi-dimensional correlation structures.

Sampling Process

The process of generating samples from the original variables can be resumed by the following five steps.

1. Data normalization: The PDFs of the variables are approximated using Gaussian kernels, and the corresponding CDFs are obtained through numerical integration. As these CDFs are monotonic and bijective, their inverse functions exist and are stored for subsequent use. Applying the CDF transformation to the original historical data maps the series onto the interval [0, 1], resulting in a uniform distribution.
2. Data aggregation: Refers to the aggregation of the normalized time series in a manner that preserves the mutual information among the set of variables whose correlations are being analyzed. For two series, this can be done by choosing the number of bins of the discretization between them that maximizes their dependence, represented by metrics such as the Maximal Information Coefficient (MIC) or the Normalized Mutual Information (NMI). For multiple series, this discretization is made between a fixed pivot variable, which has the greatest average dependence with the input data, and every other variable. To avoid the computational burden in this brute search, the number of bins could also be chosen as a hyperparameter, which is the approach used in this work, and the values that gave the best representation of the correlations were 40 bins for photovoltaic series and 20 bins for the remaining variables.
3. Bayesian Network training: The BN, which represents the conditional probability relationships among the random variables, is trained using the aggregated time series as input data. Initially, the network structure is learned through a Hill Climbing optimization procedure, while the K2 score is employed as the evaluation metric. The

number of combinations can be reduced by limiting the number of parents for each node during this phase, and in this study a max of 2 parent nodes for the BN was allowed. This choice was made to prevent a combinatorial explosion resulting in an excessive computational overhead. The network parameters are then calculated, applying the Maximum Likelihood Estimation (MLE) method. Independent variables are set aside as they have no connections with the other variables in the BN, but can be later recombined.

4. Bayesian Network sampling: The sampling from the trained BN is performed by sequentially traversing its structure, proceeding from parent nodes to child nodes in accordance with the encoded conditional dependencies, and the independent variables that are not represented in the BN can then be added. The samples obtained are discrete and uniformly distributed. Their size can be the same as the original data or some fraction of it.
5. Data transformation reversion: The aggregation of the sampled variables is subsequently reversed by generating uniformly distributed random values within the intervals defined by the discrete representations. The normalization step is then inverted by applying the inverse of the previously obtained CDF to the series, yielding samples that follow the same distribution as the original data.

2.2. Rank Score and Latin Hypercube

2.2.1. Rank Score

The correlation matrix is generally positive definite, and Cholesky decomposition is commonly used to impose the desired correlation structure on generated samples. However, this technique is only valid under the assumption that all variables follow a Gaussian distribution, which is not always appropriate in practical applications.

Reference [18] proposes a method based on rank correlations with the goal of inducing the desired rank dependence among variables. The method has the following properties:

1. It is distribution-free. In other words, there is no need to make previous assumptions about the distributions of the variables.
2. It is a simple method that does not require any unusual mathematical technique.
3. It can be applied to any sampling scheme, preserving its original intent. In this method, the sampled values remain the same; only their pairing changes to achieve the desired rank correlation.
4. The marginal distributions of the variables are not affected.

This approach relies on the premise that rank correlation provides a meaningful representation of dependence among variables. In [18], the authors present the method along with a variation designed to ensure that the original samples are independent, which is important for enhancing the accuracy of the procedure. The following steps show a summarized version of the method with the aforementioned variation:

1. Define the desired rank correlation matrix C. In the end, the correlated samples must have a sample correlation matrix M very close to C.
2. Compute P, which is a lower triangular matrix obtained from the Cholesky decomposition of C, such that $C = PP^T$. If there are m correlated variables in the problem, C and P will be $m \times m$ matrices.
3. Sample n values for each of the m variables from uniform distribution. These samples will form the $n \times m$ matrix A. Ideally, these samples are independent, with a correlation matrix close to the identity.
4. Obtain a score matrix R from A. Each element of each column will be $a(i), i = 1, 2, \ldots, n$, and $a(i)$ is a function that computes the scores. i does not correspond to each row, but

the ranking of the element in the column. Therefore, $i = 1$ corresponds to the smallest value, while $i = n$ refers to the largest one.

5. Calculate T, which is the sample correlation matrix of R. As T is positive definite and symmetrical, it is possible to use Cholesky decomposition to obtain Q, such that $T = QQ^T$ and Q is a lower triangular matrix.
6. Compute S, such that $S = PQ^{-1}$. Then, calculate $R_B^* = RS^T$. R_B^* will have a sample correlation matrix M very close to C. The term Q^{-1} ensures that R is independent, while P assigns the desired correlation to the samples.
7. Order the rows in each column of matrix A such that they have the same ordering as R_B^*.

In [18], the authors use Van der Waerden scores in step 4, which are based on normal distribution. However, they also tested with other types of scores. They state that, if ranks are used as scores, i.e., $a(i) = i$, tests showed that sample correlation was closer to the original correlation matrix than when Van der Waerden scores were used. On the other hand, they preferred Van der Waerden scores, because the scatter plots between pairs of variables seemed more "natural", resulting in elliptical patterns. When ranks were used as scores, the scatter plots seemed to be pinched in the middle and spread out in the tails.

Reference [15] evaluated some methods to consider correlation in non-sequential MCS for reliability studies and both types of scores were analyzed. The paper referred to the rank correlation method using Van der Waerden scores as Normal Score, and to the variation using ranks as scores as Rank Score. Although Rank Score did not show any clear superiority in the goodness-of-fit tests in [15], it performed better in computing reliability indices, taking into account both accuracy and computational time. For this reason, it was the recommended method among the evaluated techniques. Following this recommendation, this paper only includes the Rank Score variation.

2.2.2. Latin Hypercube Sampling (LHS)

Latin Hypercube Sampling can cover the sample space more efficiently, reducing variance in Monte Carlo Simulation, resulting in a faster convergence. This method was first presented in [19] and later developed by [20]. In the LHS method, the range of each random variable is divided into a set of strata, and one value is randomly selected from each stratum. These representative values are then combined such that each stratum is sampled exactly once during the simulation process.

LHS can be formulated as an optimization problem that seeks to maximize the distance between samples [21]. The constraint is that the projection on each axis must follow a specified probability distribution. This distribution is often uniform, and it is used the CDF of the variables.

The mathematical formulation of sampling methodologies like LHS often relies on complex optimization problems. Solving constrained and multi-objective optimization models is a fundamental transversal need in the design of robust engineering systems. For example, advanced iterative algorithms, such as Sequential Quadratic Programming (SQP), are widely applied to solve non-convex optimization formulations in complex scenarios, ranging from sampling distance allocations to the optimal balancing of emission and susceptibility limits in electromagnetic compatibility contexts [22].

A more straightforward way to implement LHS is to draw only one sample from each interval while still using randomness to pair the sampled values. Although this approach does not maximize sample dispersion, it remains effective and offers a simpler implementation.

To apply LHS, the number of samples must be chosen beforehand. Considering n samples for m independent uniform variables $\{v_i^{(j)}, i = 1, 2, \ldots, m; j = 1, 2, \ldots, n\}$, LHS can be implemented in the following way [23]:

1. Divide the domain of the ith variable v_i, i.e., [0, 1], into n segments of the same size. This will result in $[s_{i1}, s_{i2}], [s_{i2}, s_{i3}], \ldots, [s_{in}, s_{i(n+1)}]$, such that $s_{ij} = \frac{j-1}{n}$, $j = 1, 2, \ldots, n, n+1$.
2. Draw a random uniform sample $v_i^{(j)}$ in the jth segment $[s_{ij}, s_{i(j+1)}]$, which is equivalent to transforming a random standard uniform sample $v_i^{(j)} \sim U(0,1)$ using $v_i^{(j)} = \frac{v_i^{(j)}}{n+(j-1)/n}$, leading to a total of n samples $\{v_i^{(j)}, j = 1, 2, \ldots, n\}$ of $v_i^{(j)}$.
3. Randomly sort $\{v_i^{(j)}, j = 1, 2, \ldots, n\}$ for the ith variable. This can be done through sampling another n uniform samples.
4. Repeat the above steps for the m variables to form $\{v_i^{(j)}, i = 1, 2, \ldots, m\}$ as the jth sample of the desired vector.

The random uniform variables can be transformed into the variables of the original problem through inverse transformation of their CDF.

The sample size n may be set equal to the total number of samples required, or alternatively defined as a smaller unit, with sequential samplings of size n conducted until the desired total is achieved. A very large n may compromise the performance of the LHS, which requires much more time than drawing samples in smaller chunks. Furthermore, setting this parameter independently of the number of data points is advantageous, for example, in reliability studies with non-sequential MCS, since the necessary number of samples to reach convergence is not known beforehand. In this paper, n is set independently of the data and was defined as 1000. This number proved to be consistently reasonable to produce sampling that provided a good spatial representation of the variables.

2.3. Statistical Metrics

To validate the proposed methodology, scatter plots are used to compare the generated samples with the original data and assess their ability to preserve the underlying correlation structure. In addition, several statistical metrics are computed to verify that the sampled and original datasets exhibit equivalent statistical properties. The metrics adopted for this analysis are summarized as follows:

1. R^2: the coefficient of determination quantifies the proportion of variability in the dependent variable that can be explained by the independent variable. It is given by [24]:
$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \overline{y})^2} \tag{6}$$
where $\overline{y}$ is the mean of the actual data, and y_i and f_i are the ith data point from the original and sampled series, respectively.
2. Wilcoxon signed-rank test (p-value): It evaluates the null hypothesis that two paired samples are drawn from identical distributions by examining whether the distribution of their differences is symmetric around zero. Under the two-sided alternative hypothesis, which assumes that the differences are symmetric about a nonzero value, the test statistic is computed as follows [25]:
$$W = \min\left(\sum_{i: y_i - f_i > 0} R_i, \sum_{i: y_i - f_i < 0} R_i \right) \tag{7}$$

where R_i represents the ranks defined as the ordered $|y_i - f_i|$ values. Considering a high number of data pairs, W can be approximated by a normal distribution. The Z statistic and its corresponding p-value are then given as [25]:

$$Z_W = \frac{W - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \tag{8}$$

$$p = 2 \times P(Z > |Z_W|) \tag{9}$$

3. Kolmogorov–Smirnov test (p-value): It quantifies the discrepancy between two empirical probability distributions in order to evaluate the null hypothesis that both samples originate from the same underlying distribution. Let $Y(x)$ and $F(x)$ denote the empirical cumulative distribution functions (ECDFs) of the observed and predicted data, constructed from n_y and n_f samples, respectively. Based on these ECDFs, the test statistic and its corresponding p-value are calculated as follows [26]:

$$D_{n_y,n_f} = \sqrt{\frac{n_y n_f}{n_y + n_f}} \sup_x |Y(x) - F(x)| \tag{10}$$

$$P(D_{n_y,n_f} > d) \approx \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 d^2} \tag{11}$$

where sup represents the supremum of the set of distances.

2.4. Zero-Inflated Distribution Methods

In the context of power system reliability, solar photovoltaic (PV) generation presents a semicontinuous distribution, characterized by a discrete probability mass at zero during nighttime and continuous positive values during daylight. Recent literature highlights the critical need to explicitly address this data skewness in solar behavior. For instance, ref. [27] demonstrated that failing to account for excessive zeros, arising from physical constraints like nighttime, yields unrealistic predictive fits, requiring specialized zero-inflated modeling. To accurately capture this behavior and avoid zero-inflation bias, this work adopts an approach inspired by Hurdle Models [28,29]. Properly modeling this probability mass at zero is essential for maintaining statistical consistency, verifying stochastic forecasts, and ensuring that Monte Carlo simulations accurately reflect physical system constraints when calculating reliability indices [30–32]. Furthermore, comprehensive statistical evaluations by [33] establish that Hurdle models effectively isolate structural zeros from the continuous data-generating process, outperforming standard mixture distributions when handling strictly bounded count or semicontinuous data.

To implement this, the original proportion of structural zeros is isolated into a single discrete bin, allowing the remaining positive samples to be modeled using KDE + BN without distortion.

To accommodate the zero-inflated asymmetry characteristic of nighttime photovoltaic generation, the KDE discretization employs a hurdle-inspired approach. Solar profiles are identified either via plant metadata or an empirical threshold (>33% strictly null records). To accurately capture the continuous daytime distribution, these series are discretized into 40 bins (twice the standard stochastic parameter), as stated before. The hurdle mechanism actively manages structural zeros: assuming that the static nighttime mass is concentrated in the first bin, during reverse sampling the exact proportion of zeros from the original series is probabilistically reinjected into the samples of this lower stratum.

However, this specialized treatment is not applied to the RS + LHS method, because RS relies on a unified rank-transformation process to preserve the target joint correlation matrix across all random variables simultaneously. Applying a differentiated Hurdle-like treatment to specific series would disrupt the structural consistency of the joint rank-ordering. This asymmetric treatment does not compromise the fairness of the performance comparison. Because RS + LHS preserves the marginal distributions of the variables and the inverse transform is done through the empirical CDF, it accurately reproduces the exact mass of zeros present in the original data without requiring specialized adjustments. Conversely, KDE is fundamentally a continuous density estimator. Without the hurdle-inspired discretization, KDE would mathematically smooth the massive spike at zero, resulting in non-physical generation values during nighttime. Therefore, the hurdle model applied to KDE + BN is strictly a corrective measure required to allow a fair baseline comparison between the two methods in a highly zero-inflated scenario.

Standard variance-based metrics like R^2 are heavily biased when applied to zero-inflated, heteroscedastic data, such as nighttime PV generation [34]. To accurately assess intermittent data, the deviance pseudo-R^2 (R^2_{dev}) can be adapted to compare two time series rather than a traditional fitted model [35]. In this context, it measures the proportion of explained deviance between the sampled values and the actual observations [36,37]:

$$R^2_{\text{dev}} = 1 - \frac{\sum_i [(y_i \log(y_i) - y_i) - (y_i \log(\hat{\mu}_i) - \hat{\mu}_i)]}{\sum_i [(y_i \log(y_i) - y_i) - (y_i \log(\bar{y}) - \bar{y})]} = 1 - \frac{\log L(y) - \log L(\hat{\mu})}{\log L(y) - \log L(\bar{y})} \quad (12)$$

where y_i represents the original observations, $\hat{\mu}_i$ the sampled values from the proposed method, and $\bar{y}$ the mean of the observations, with L denoting their respective log-likelihoods.

When comparing two time series, standard R^2 acts as a measure of proximity, calculating the absolute numerical distance between sampled and original values. In contrast, R^2_{dev} evaluates statistical consistency, measuring the likelihood of the sampled values occurring within the original probability distribution.

3. Results and Discussion

In this section, the proposed metrics are used to evaluate whether the generated synthetic samples adequately reproduce the original variables, thereby assessing the goodness-of-fit of the two methods. In addition to reporting metrics for each individual variable, aggregated quantities are also analyzed to examine the impact of jointly generating correlated series. Two aggregate variables are considered: Total Generation (GTOT), defined as the sum of all generation series, and Net Generation (GNET), defined as the difference between Total Generation and load demand.

The following cases consisting of actual Brazilian energy system time series were evaluated:

Case 1 – Mix of renewable generation of NE region—consisting of ten time series in total: seven from wind generation, one hydro and one solar generation, and another representing the variable load. The time series have 5 years of data.
Case 2 – Solar generation of NE and SE regions—consisting of nineteen time series in total: eighteen from solar generation and one representing the variable load. The time series have 3 years of data.
Case 3 – Mix of renewable generation of NE and SE regions—consisting of twenty eight time series in total: all series from Case 1 but the solar generation, which is replaced by the eighteen solar series from Case 2. Both load series from Case 1 and Case 2 are considered. The time series have 2 years of data.

Figure 1 provides a visual overview of the subsystems considered in this study and the corresponding composition of each case study.

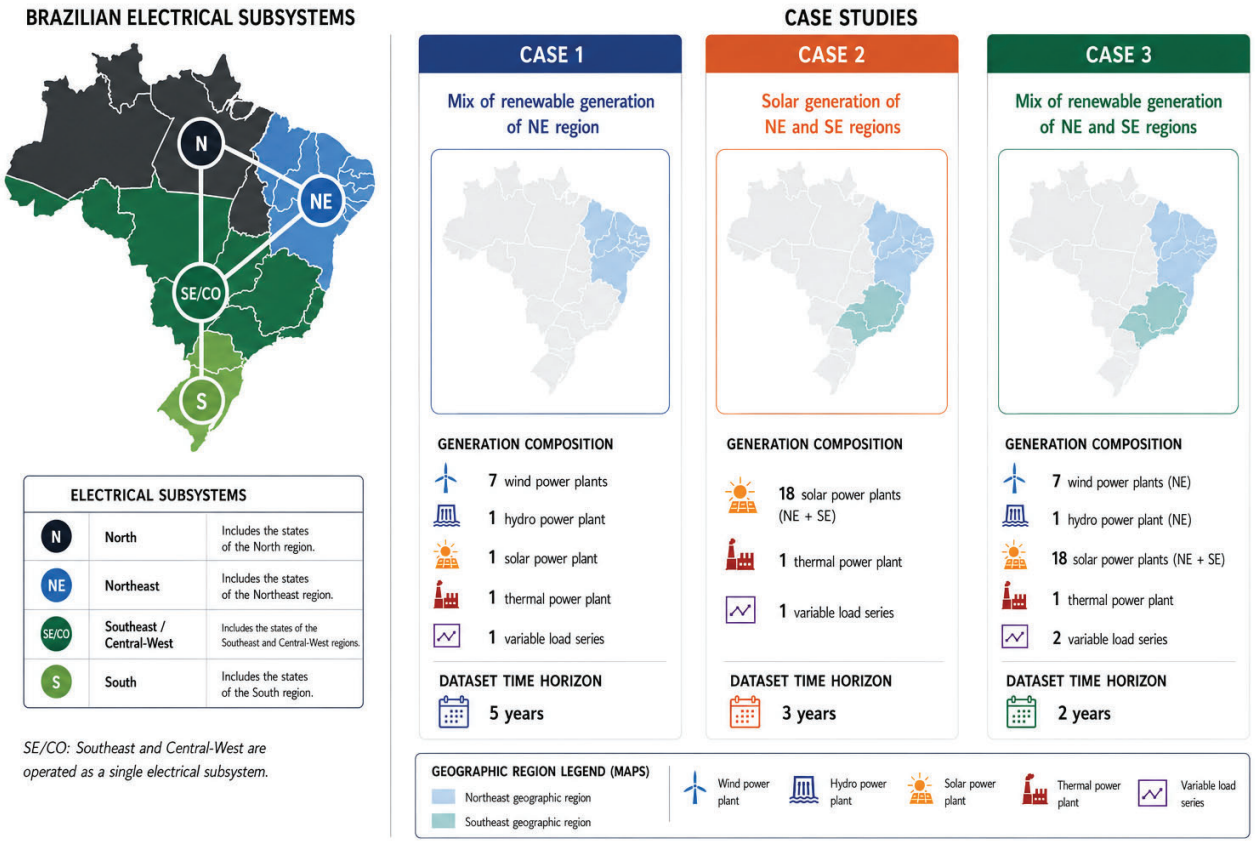


Figure 1. Brazilian subsystems considered in this study and the composition of the three case studies.

3.1. Description of the Cases

3.1.1. Case 1

This case is based on the Northeast region of Brazil, and it is composed of a mix of different sources: wind, solar, and hydro. The hydroelectric power plant (HPP) has a time series based on the Sobradinho power plant inflow regime, which is the largest HPP in the Northeast region. The PV power plant is derived from the power series curve of the combination of all solar plants of the Northeast region. Each wind time series represents one state from the Northeast region that had wind generation for the considered period (Bahia, Ceará, Pernambuco, Piauí, Rio Grande do Norte and Maranhão). The only exception comes from the state of Rio Grande do Norte, which has two wind profile regimes: coastal and inland. This case is based on the one presented in [15], and all the curves were obtained from [38], considering the hourly time series from the years 2019 to 2023.

Although this case is based on real curves, the proportion of the sources is different. The load has a maximum value of 1000 MW and an average value of 754.7 MW. It was considered that the five-year average of the HPP was equivalent to 50% of the average load. A similar installed capacity was adopted for both wind and PV power plants, considering all seven wind farms. The proportion of the seven wind farms followed their actual participation in the average generation for the considered period. Also, a constant thermoelectric power plant (TPP) was added to better match the total generation with the load. Table 1 shows the total power and capacity factor for each source, along with all wind farms (WFs).

The correlation between time series was measured by Pearson coefficients and is shown in Table 2. HPP has a negative correlation with all wind farms, indicating some complementarity between the two sources, which does not exist between PV and HPP. All wind farms have positive correlation with each other because of their similar seasonality. However, wind farms with a coastal profile generate more during daytime, while plants

with an inland profile generate more during nighttime and early morning hours. Thus, wind farms with a coastal profile have a low positive correlation with PV, while the others have a negative correlation. WF7 has a flatter profile and its correlation with PV is positive, but they are almost independent of each other. Correlation among different sources is moderate to low, with the largest values observed among wind plants.

Table 1. Power plant characteristics—Case 1.

Power Plants	Total Power (MW)	Capacity Factor (%)
WF1 (Bahia)	222	49.09
WF2 (Ceará)	54	41.07
WF3 (Pernambuco)	24	40.42
WF4 (Piauí)	132	35.67
WF5 (Rio Grande do Norte–Inland)	96	47.63
WF6 (Rio Grande do Norte–Coast)	144	42.86
WF7 (Maranhão)	18	36.25
PV	655	25.40
HPP	1750	21.56
TPP	150	100

Table 2. Correlation between time series—Case 1 (original data).

	Load	WF1	WF2	WF3	WF4	WF5	WF6	WF7	PV	HPP
Load	1.0000	−0.2579	0.2970	0.1667	−0.3851	−0.0629	0.1446	0.2331	0.0935	0.0174
WF1	−0.2579	1.0000	0.1503	0.3793	0.7729	0.5923	0.2867	0.1285	−0.3198	−0.2496
WF2	0.2970	0.1503	1.0000	0.3423	0.1559	0.4583	0.6849	0.6888	0.1035	−0.3084
WF3	0.1667	0.3793	0.3423	1.0000	0.1940	0.5553	0.3378	0.2980	−0.3999	−0.0918
WF4	−0.3851	0.7729	0.1559	0.1940	1.0000	0.5578	0.3145	0.1276	−0.1598	−0.3327
WF5	−0.0629	0.5923	0.4583	0.5553	0.5578	1.0000	0.6517	0.3212	−0.2774	−0.2581
WF6	0.1446	0.2867	0.6849	0.3378	0.3145	0.6517	1.0000	0.4050	0.0909	−0.3107
WF7	0.2331	0.1285	0.6888	0.2980	0.1276	0.3212	0.4050	1.0000	0.0294	−0.2415
PV	0.0935	−0.3198	0.1035	−0.3999	−0.1598	−0.2774	0.0909	0.0294	1.0000	−0.0158
HPP	0.0174	−0.2496	−0.3084	−0.0918	−0.3327	−0.2581	−0.3107	−0.2415	−0.0158	1.0000

3.1.2. Case 2

Case 2 includes eighteen PV generation series obtained from real solar plant complexes of the Brazilian energy system. These complexes are spread across the northeast and southeast regions of Brazil. This case also features a varying load with a maximum of 64 MW and an average of 43.35 MW. The load curve follows the profile of the southeast load time series. This region accounts for around 55% of the total load of the Brazilian interconnected power system, and its profile closely resembles that of the overall system load. In addition to the PV power plants, as in Case 1, there is a constant TPP to better balance load and generation.

Table 3 shows the total power and capacity factor of the power plants considered in this case. The total power does not correspond to the actual capacity of the real plants. All data was obtained from [38] for the years 2022 to 2024. Table 4 presents the Pearson correlation coefficients between all time series. It can be seen that all PV plants have high correlation among each other, with the lowest correlation being around 0.69 and the highest one being almost 0.97. Load is positively correlated to all PV plants and their correlation is relatively weak, being around 20%.

Table 3. Power plant characteristics—Case 2.

Power Plants	Total Power (MW)	Capacity Factor (%)
TPP	32	100
PV1 (Complex Pirapora 2)	329	26.8
PV2 (Complex Dracena)	81.12	21.9
PV3 (Complex Lapa)	60.04	24.7
PV4 (Assú V)	34	23.9
PV5 (Complex Alex)	278.64	26.2
PV6 (Complex BJL)	40	24.5
PV7 (Complex Boa Hora)	182.84	19.1
PV8 (Complex Bom Jesus)	59.78	24.4
PV9 (Complex Brígida)	163.08	24.2
PV10 (Complex Calcário)	131.92	23.8
PV11 (Complex Floresta)	86	23.2
PV12 (Complex FV SJP)	185.6	24.8
PV13 (Complex Ituverava)	196	26.5
PV14 (Complex Juazeiro Solar)	120	29.6
PV15 (Complex Juazeiro Solar 2)	157.07	26.3
PV16 (Complex Nova Olinda)	208.65	26.6
PV17 (Complex Sertão Solar Barreiras)	94.08	24.9
PV18 (Complex Sol do Futuro)	80.64	21

3.1.3. Case 3

This case is a combination of the previous two cases. However, since both cases did not correspond to the same years of data, this case only has data regarding the years 2022 and 2023. There are two load profiles, one from the southeast (SE) and another from the northeast (NE) region, but now the SE load has a maximum value of 500 MW and an average value of 339.43 MW. The NE load is the same as Case 1, but, since there are less data, its average is now 750.43 MW and its maximum value is still 1000 MW. The two constant TPP were replaced by one of 200 MW. The PV power plant from Case 1 was replaced by the eighteen PV plants (PV1 to PV18) of Case 2. The total power values of Tables 1 and 3 are still valid.

Since the curves are the same as in the previous cases (with the caveat regarding the dates) and the PV plants have a similar shape, the correlations are of the same order of magnitude as the correlations in Tables 2 and 4.

3.2. Goodness-of-Fit

In this section, the goodness-of-fit of the two methods (RS + LHS and KDE + BN) is evaluated for the three cases. The analysis begins with a comparison of statistical metrics, followed by spatial correlation matrices, the cumulative distributions of historical and sampled G_{TOT}, and a detailed assessment of the performance of each method in replicating behaviors at both system and the plant level.

Table 4. Correlation between time series—Case 2 (original data).

	Load	**PV1**	**PV2**	**PV3**	**PV4**	**PV5**	**PV6**	**PV7**	**PV8**	**PV9**	**PV10**	**PV11**	**PV12**	**PV13**	**PV14**	**PV15**	**PV16**	**PV17**	**PV18**
Load	1.0000	0.2401	0.2615	0.2286	0.1820	0.1985	0.2246	0.2504	0.2148	0.2180	0.1839	0.2150	0.2208	0.2147	0.2323	0.2310	0.2193	0.2192	0.2281
PV1	0.2401	1.0000	0.8051	0.8147	0.7952	0.7784	0.8304	0.7832	0.8344	0.7769	0.7478	0.7928	0.8302	0.8308	0.8187	0.8254	0.8277	0.8185	0.8057
PV2	0.2615	0.8051	1.0000	0.7396	0.7444	0.7213	0.7563	0.8476	0.7535	0.7432	0.6960	0.7594	0.7687	0.7498	0.7741	0.7807	0.7618	0.7347	0.7730
PV3	0.2286	0.8147	0.7396	1.0000	0.7644	0.7634	0.9217	0.7269	0.9322	0.7555	0.7372	0.7717	0.8086	0.8548	0.8027	0.8095	0.8071	0.8435	0.7844
PV4	0.1820	0.7952	0.7444	0.7644	1.0000	0.8897	0.7902	0.7278	0.7862	0.8074	0.8747	0.8912	0.8263	0.7672	0.8166	0.8249	0.8203	0.7573	0.8447
PV5	0.1985	0.7784	0.7213	0.7634	0.8897	1.0000	0.7768	0.7141	0.7780	0.8014	0.9289	0.9000	0.8235	0.7635	0.8053	0.8135	0.8145	0.7586	0.8368
PV6	0.2246	0.8304	0.7563	0.9217	0.7902	0.7768	1.0000	0.7340	0.9451	0.7769	0.7478	0.7977	0.8211	0.8736	0.8180	0.8278	0.8165	0.8486	0.7943
PV7	0.2504	0.7832	0.8476	0.7269	0.7278	0.7141	0.7340	1.0000	0.7298	0.7280	0.6916	0.7475	0.7451	0.7334	0.7558	0.7604	0.7375	0.7184	0.7497
PV8	0.2148	0.8344	0.7535	0.9322	0.7862	0.7780	0.9451	0.7298	1.0000	0.7765	0.7481	0.7844	0.8291	0.8829	0.8182	0.8277	0.8358	0.8532	0.7959
PV9	0.2180	0.7769	0.7432	0.7555	0.8074	0.8014	0.7769	0.7280	0.7765	1.0000	0.7710	0.8049	0.8371	0.7781	0.8367	0.8518	0.8301	0.7589	0.8138
PV10	0.1839	0.7478	0.6960	0.7372	0.8747	0.9289	0.7478	0.6916	0.7481	0.7710	1.0000	0.8880	0.7964	0.7312	0.7802	0.7876	0.7862	0.7283	0.8271
PV11	0.2150	0.7928	0.7594	0.7717	0.8912	0.9000	0.7977	0.7475	0.7844	0.8049	0.8880	1.0000	0.8268	0.7737	0.8151	0.8236	0.8125	0.7713	0.8607
PV12	0.2208	0.8302	0.7687	0.8086	0.8263	0.8235	0.8211	0.7451	0.8291	0.8371	0.7964	0.8268	1.0000	0.8280	0.8629	0.8778	0.9552	0.8185	0.8445
PV13	0.2147	0.8308	0.7498	0.8548	0.7672	0.7635	0.8736	0.7334	0.8829	0.7781	0.7312	0.7737	0.8280	1.0000	0.8118	0.8227	0.8236	0.8812	0.7852
PV14	0.2323	0.8187	0.7741	0.8027	0.8166	0.8053	0.8180	0.7558	0.8182	0.8367	0.7802	0.8151	0.8629	0.8118	1.0000	0.9716	0.8542	0.8052	0.8306
PV15	0.2310	0.8254	0.7807	0.8095	0.8249	0.8135	0.8278	0.7604	0.8277	0.8518	0.7876	0.8236	0.8778	0.8227	0.9716	1.0000	0.8710	0.8147	0.8396
PV16	0.2193	0.8277	0.7618	0.8071	0.8203	0.8145	0.8165	0.7375	0.8358	0.8301	0.7862	0.8125	0.9552	0.8236	0.8542	0.8710	1.0000	0.8132	0.8426
PV17	0.2192	0.8185	0.7347	0.8435	0.7573	0.7586	0.8486	0.7184	0.8532	0.7589	0.7283	0.7713	0.8185	0.8812	0.8052	0.8147	0.8132	1.0000	0.7833
PV18	0.2281	0.8057	0.7730	0.7844	0.8447	0.8368	0.7943	0.7497	0.7959	0.8138	0.8271	0.8607	0.8445	0.7852	0.8306	0.8396	0.8426	0.7833	1.0000

https://doi.org/10.3390/en19133163

3.2.1. Case 1

Case 1 evaluates a system composed predominantly of wind farms. The statistical metrics for this scenario are summarized in Table 5. For metrics R^2 and R^2_{dev}, it is desired to obtain results close to one, while, for Wilcoxon and KS, the confidence level considered was 95% and therefore their p-value should be above 0.05. When p-value was below this threshold, the values were colored red in the table.

Table 5. Statistical comparative metrics for original data vs. non-sequential methods in Case 1.

RS + LHS												
Metric	Load1	WF1	WF2	WF3	WF4	WF5	WF6	WF7	PV	HPP	*G_TOT*	*G_NET*
R^2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.994	0.995
R^2_{dev}	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.995	-
Wilcoxon p-val	0.788	0.901	0.804	0.848	0.921	0.940	0.947	0.997	0.847	0.906	0.042	0.037
KS p-val	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.000
KDE + BN												
Metric	Load1	WF1	WF2	WF3	WF4	WF5	WF6	WF7	PV	HPP	*G_TOT*	*G_NET*
R^2	0.998	1.000	1.000	1.000	0.999	1.000	1.000	1.000	1.000	0.999	0.996	0.997
R^2_{dev}	0.997	0.999	1.000	0.999	0.999	0.999	0.999	0.991	0.996	0.999	0.996	-
Wilcoxon p-val	0.177	0.662	0.935	0.249	0.442	0.316	0.565	0.594	0.000	0.519	0.526	0.065
KS p-val	0.225	0.122	0.472	0.161	0.064	0.201	0.462	0.000	0.000	0.000	0.000	0.000

Regarding R^2 and R^2_{dev}, both RS + LHS and KDE + BN had results very close to one. However, because R^2_{dev} formulation strictly requires positive numbers, it cannot be used for net variables like G_{NET}. As for the metrics based on the hypothesis test, RS + LHS presented satisfactory results for the original variables of the problem, which is expected since one of the advantages of this method is to preserve the marginal distributions. However, when considering aggregate variables, both KS and Wilcoxon had a p-value below 0.05. KDE + BN failed the KS test for WF7, PV and HPP, but only PV1 did not pass the Wilcoxon test, so it cannot be affirmed that it did not adequately represent WF7 and HPP. Regarding the aggregate variables (G_{TOT} and G_{NET}), this method also failed the KS test, but had good results for the Wilcoxon test. So, even though KDE + BN may not have perfectly reproduced the distributions of the original values, it seems to have better represented the dependency among them.

Initial comparative analysis focuses on the preservation of the system's dependency structure and the global generation profile. Figure 2 presents a heatmap representing the correlation matrices for original historical data and both RS + LHS and KDE + BN methods' sampled data, while Figure 3 overlays G_{TOT} cumulative distributions of the sampled data by each method against the historical total generation data.

It can be seen that the heatmaps for RS + LHS and KDE + BN are very close to the one that represents the original data. The heatmap for RS + LHS seems a little closer to the original compared to the KDE + BN samples, especially with regard to correlations with Load and PV. This result is coherent since the RS method focuses on mimicking the correlation matrix. Nonetheless, visual comparison indicates good performance for both methods.

Regarding the CDF for G_{tot}, RS + LHS and KDE + BN produced samples with an empirical distribution that is very close to the one of the original data. Although the difference is small, the CDF originated from KDE + BN samples is closer, with a small deviation for RS + LHS, most noticeable in the middle of the curve. This is coherent with the Wilcoxon metric calculated for G_{tot} in Table 5, but RS + LHS is still very close despite failing the hypothesis metrics.

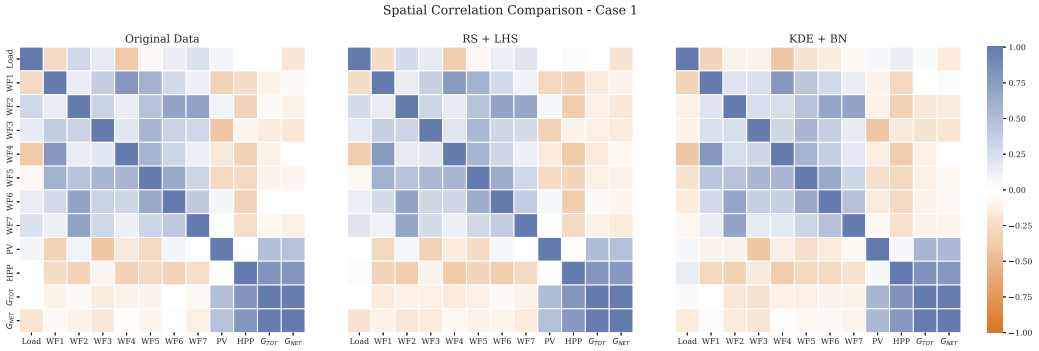


Figure 2. Spatial correlation matrix comparison using Pearson (Case 1).

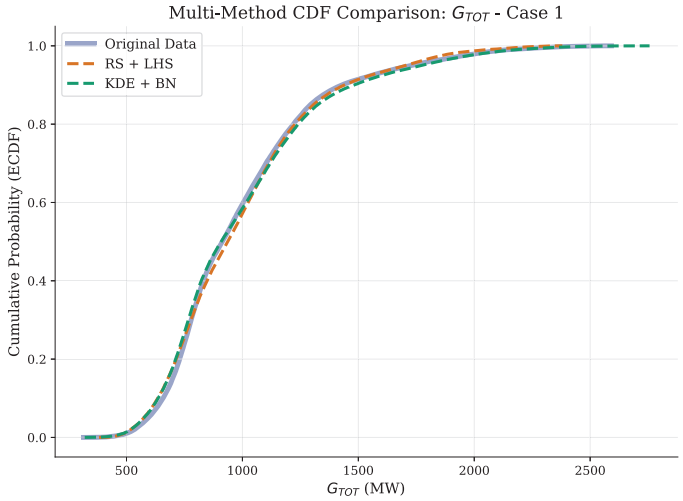
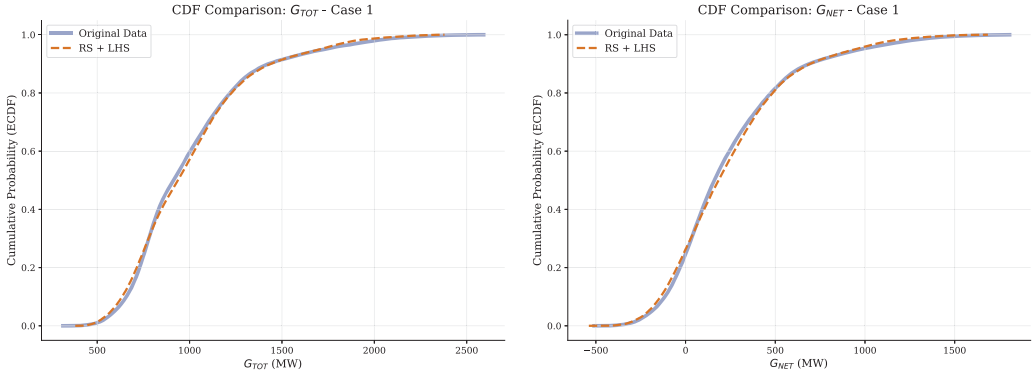


Figure 3. Comparison of G_{TOT} cumulative distribution functions (Case 1).

RS + LHS

This case comprises five years of data, totaling 43,824 data points per variable. Since the LHS method employed in this study generates samples in blocks of 1000, a total of 44,000 samples are produced. After applying the RS method to impose the desired correlation on the final block, the last 176 samples are discarded, ensuring that the total number of points matches the original dataset. The same procedure is applied to the subsequent two cases, which likewise have sample sizes that are not exact multiples of 1000.

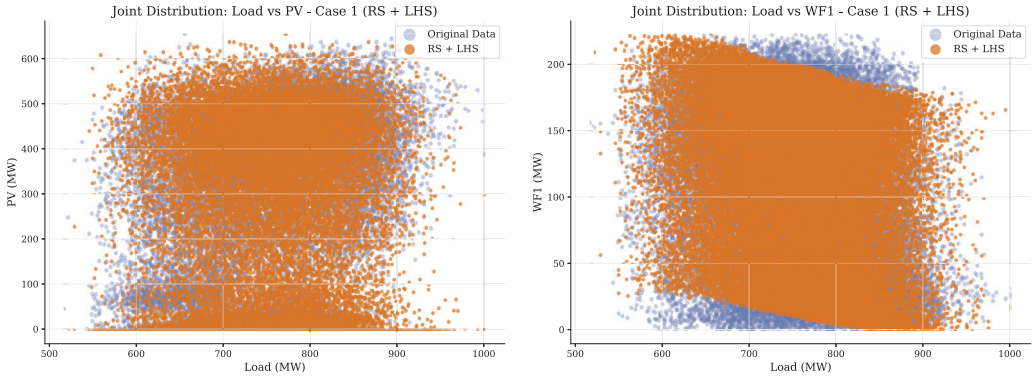
Figure 4 shows the resulting G_{TOT} and G_{NET} CDFs obtained by this method, showing how adherent they are to the original data.



G_{TOT} CDF (RS + LHS) G_{NET} CDF (RS + LHS)

Figure 4. CDF evaluation for total and net generation using RS + LHS (Case 1).

The preservation of local dependencies is exemplified through scatter plots for Load versus PV and versus WF1 (Figure 5), highlighting the method's consistency in reproducing the original operational region. Although the scatter plots are quite similar for Load × PV, for Load × WF1 there are some points of the original data that were not sampled by RS + LHS (upper right and lower left).

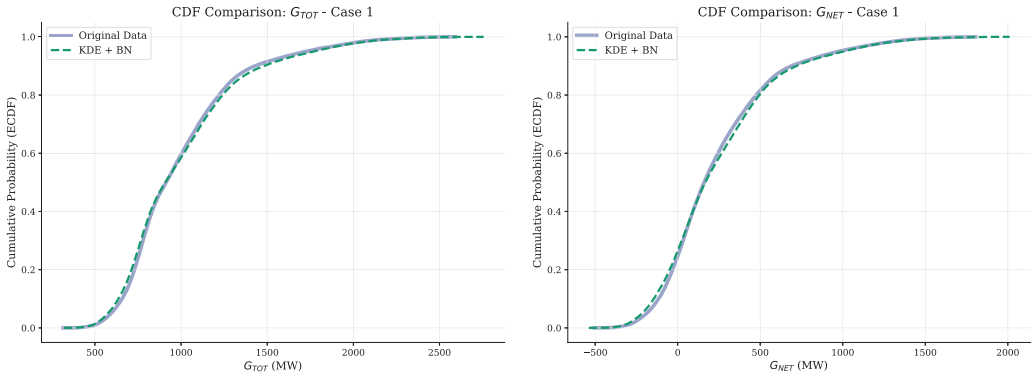


Load × PV (RS + LHS) Load × WF1 (RS + LHS)

Figure 5. Individual plant scatter plots for RS + LHS in Case 1.

KDE + BN

The KDE + BN method's performance in capturing system-wide generation variables is illustrated in Figure 6, with both curves almost overlapping each other.



G_{TOT} CDF (KDE + BN) G_{NET} CDF (KDE + BN)

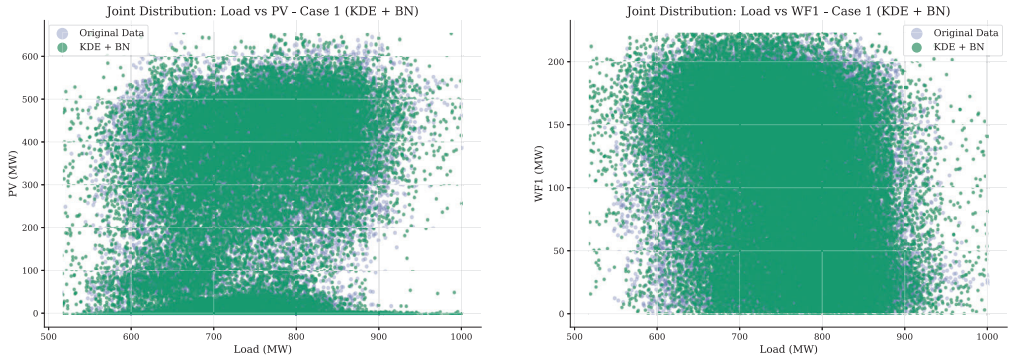
Figure 6. CDF evaluation for total and net generation using KDE + BN (Case 1).

Figure 7 shows the scatter plots for Load versus individual plant outputs (PV and WF1), demonstrating how the method handles the specific variability of different sources. KDE + BN produced scatter plots that were very similar to the ones from the original data.

3.2.2. Case 2

Case 2 analyzes a system with significant PV penetration. Detailed metrics are provided in Table 6. As in Case 1, RS + LHS performed well in all metrics for individual variables. As for aggregate variables, KDE + BN did not pass the KS or Wilcoxon test for G_{TOT}, but passed the Wilcoxon test for G_{NET}.

The values of both R^2 and R^2_{dev} are very close to one for aggregate variables, highlighting that both absolute numerical distance and probabilistic unlikeliness are practically eliminated in both methods. Regarding the performance of other metrics for KDE + BN, the KS test failed for several PV plants, while the Wilcoxon p-value failed only for PV2 and PV7.



Load vs. PV (KDE + BN) Load vs. WF1 (KDE + BN)

Figure 7. Individual plant scatter plots for KDE + BN in Case 1.

Figure 8 shows that both methods had heatmaps that were very similar to the one produced by the original data. RS + LHS again had a heatmap slightly more similar, and KDE + BN had some deviations, especially in the center of the chart, presenting correlations that were a bit smaller than the correlation of the original data. In this highly correlated scenario, both evaluated methods produce samples that visually align well with the original data. The weakest correlations are between Load and the PV plants, and this behavior is accurately reproduced in both methods. Some correlations between PV plants were not as strong in KDE + BN samples as they were in the original dataset, but the overall aspect of the heatmap is preserved. Again, since RS + LHS seeks to generate samples with the same correlation matrix, the heatmap related to this method presents a closer visual match to the original data.

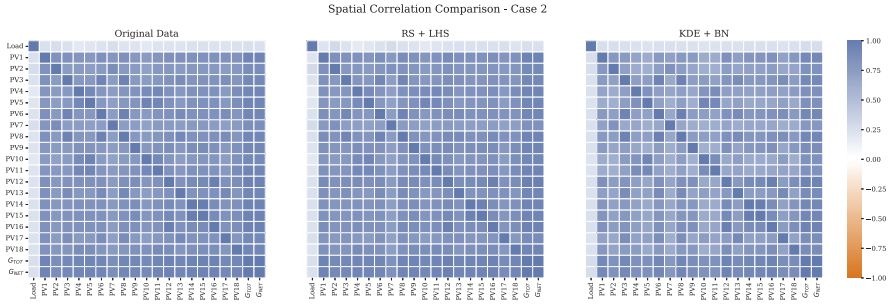


Figure 8. Spatial correlation matrix comparison using Pearson (Case 2).

Figure 9 shows that although the CDF of G_{TOT} by KDE + BN had some small deviations compared to the original curve, the differences for RS + LHS were much greater. The RS + LHS method has some weaknesses when the series have many ties, as is the case for PV series, since there are many zero values and it can explain the worse performance of this method compared to Case 1.

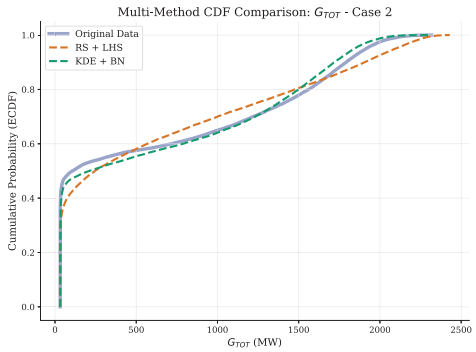


Figure 9. Comparison of G_{TOT} cumulative distribution functions (Case 2).

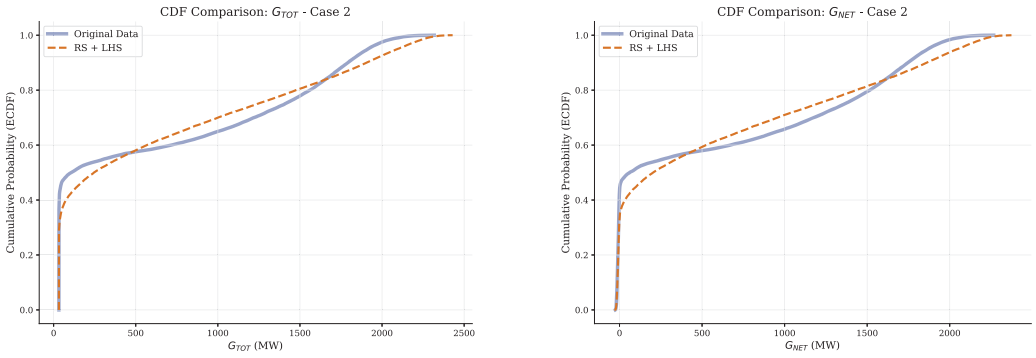
Table 6. Statistical comparative metrics for original data vs. non-sequential methods in Case 2.

RS + LHS																					
Metric	Load1	PV1	PV2	PV3	PV4	PV5	PV6	PV7	PV8	PV9	PV10	PV11	PV12	PV13	PV14	PV15	PV16	PV17	PV18	*G_TOT*	*G_NET*
R^2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.976	0.976
R^2_{dev}	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.975	-
Wilcoxon *p*-val	0.992	0.996	0.822	0.850	0.760	0.702	0.736	0.711	0.742	0.739	0.739	0.829	0.695	0.706	0.775	0.744	0.663	0.729	0.735	0.049	0.003
KS *p*-val	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.000
KDE + BN																					
Metric	Load1	PV1	PV2	PV3	PV4	PV5	PV6	PV7	PV8	PV9	PV10	PV11	PV12	PV13	PV14	PV15	PV16	PV17	PV18	*G_TOT*	*G_NET*
R^2	0.999	1.000	0.999	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.997	0.997
R^2_{dev}	0.999	1.000	0.999	1.000	1.000	1.000	1.000	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	0.996	-
Wilcoxon *p*-val	0.242	0.406	0.047	0.951	0.652	0.599	0.899	0.012	0.954	0.967	0.995	0.975	0.949	0.969	0.469	0.992	0.654	0.286	0.583	0.001	0.162
KS *p*-val	0.545	0.524	0.000	0.049	0.004	0.210	0.112	0.000	0.034	0.050	0.155	0.284	0.043	0.000	0.000	0.000	0.058	0.000	0.006	0.000	0.000

https://doi.org/10.3390/en19133163

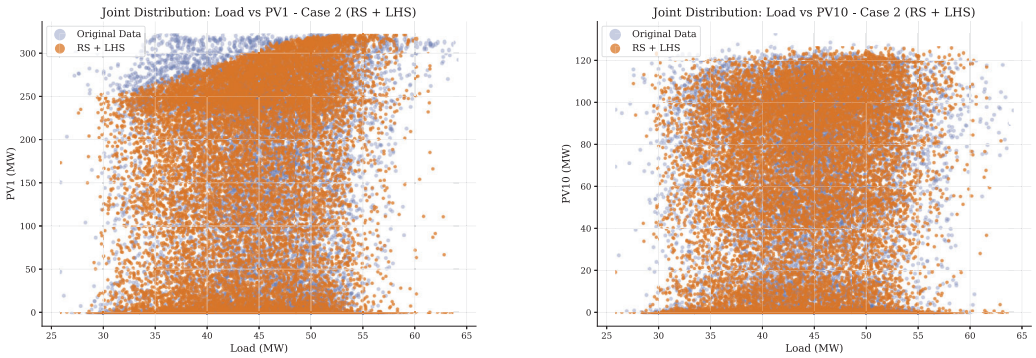
RS + LHS

Figure 10 shows the resulting G_{TOT} and G_{NET} CDFs obtained by this method, while Figure 11 shows the scatter plots for Load versus PV1 and PV10. The scatter plot for Load versus PV10 is very similar to the original data, but for Load versus PV1, it can be seen that some points on the upper left of the original data are not sampled by RS + LHS.



G_{TOT} CDF (RS + LHS) G_{NET} CDF (RS + LHS)

Figure 10. CDF evaluation for total and net generation using RS + LHS (Case 2).

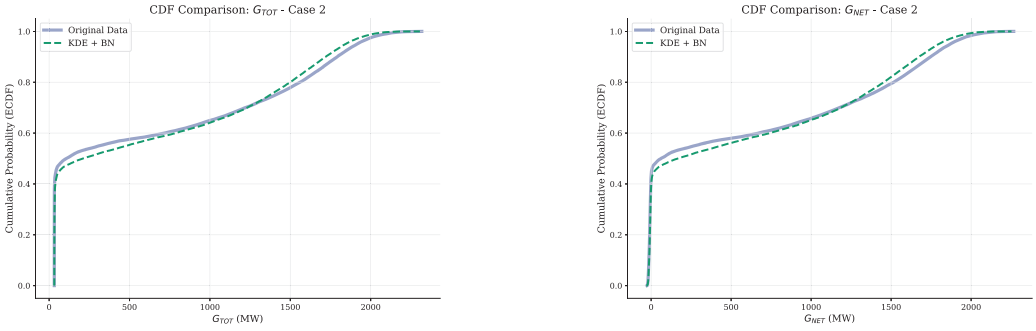


Load vs. PV1 (RS + LHS) Load vs. PV10 (RS + LHS)

Figure 11. Individual plant scatter plots for RS + LHS in Case 2.

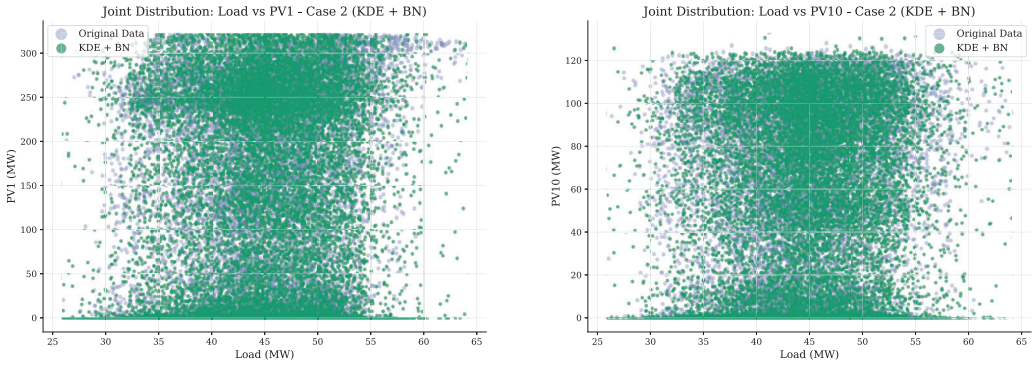
KDE + BN

Figure 12 shows the resulting G_{TOT} and G_{NET} CDFs obtained by this method. Plant-level behavior is shown in scatter plots of Load versus PV1 and PV10 (Figure 13). Although aggregate variables failed some metrics, Figure 12 shows that the curves coming from KDE + BN samples were very close to the ones from the original data. Scatter plots again show that this method seems to satisfactorily reproduce dependence among variables.



G_{TOT} CDF (KDE + BN) G_{NET} CDF (KDE + BN)

Figure 12. CDF evaluation for total and net generation using KDE + BN (Case 2).



Load vs. PV1 (KDE + BN) Load vs. PV10 (KDE + BN)

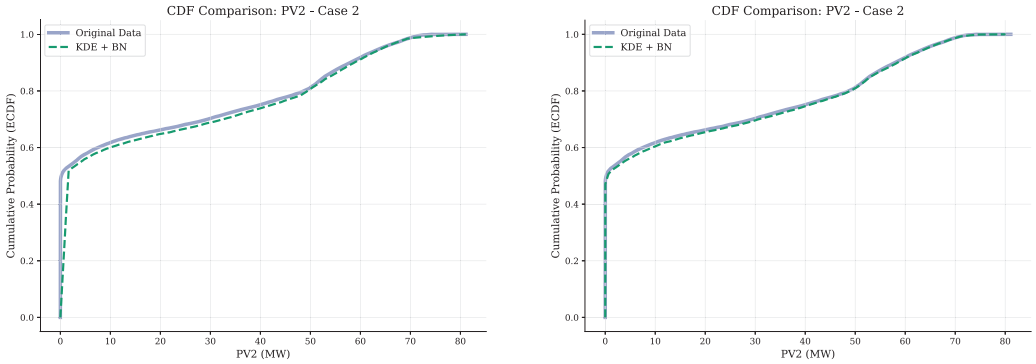
Figure 13. Individual plant scatter plots for KDE + BN in Case 2.

Benefits of Hurdle-Style KDE + BN

Case 2 particularly emphasizes the benefits of the hurdle-style KDE + BN approach. The following analysis compares the proposed hurdle-style KDE + BN against a standard KDE + BN approach without the hurdle mechanism.

Table 7 shows the metrics calculated for KDE + BN without the hurdle approach. It is very clear that the Wilcoxon and KS tests worsened for all PVs and aggregate variables, the Load being the only variable that was not affected.

Visually, this comparison is illustrated in Figure 14. It can be seen that the standard KDE approach exhibits poor goodness-of-fit at the lower tail of the PV2 CDF. In contrast, the hurdle-style approach successfully eliminates this divergence. This improvement is consistently observed for all PV plants, confirming the superiority of the hurdle-style approach.



PV2 CDF without hurdle-style approach PV2 CDF with hurdle-style approach

Figure 14. Comparison for KDE + BN in Case 2, without (**left**) and with (**right**) hurdle-style approach.

3.2.3. Case 3

Case 3 is the most complex one, featuring a multi-load and multi-source system. Metrics, correlation matrices, and the G_{TOT} CDF comparison are detailed in Table 8 and Table 9, Figure 15 and Figure 16, respectively.

The fact that both methods yield R^2 and R^2_{dev} values close to one confirms a good fit. Again, both methods simultaneously eliminate absolute numerical error and probabilistic unlikeliness, successfully handling the complexity of the system.

Table 7. KDE + BN statistical comparative metrics without hurdle approach for Case 2.

KDE + BN–Without Hurdle Approach																					
Metric	**Load1**	**PV1**	**PV2**	**PV3**	**PV4**	**PV5**	**PV6**	**PV7**	**PV8**	**PV9**	**PV10**	**PV11**	**PV12**	**PV13**	**PV14**	**PV15**	**PV16**	**PV17**	**PV18**	***G_TOT***	***G_NET***
R^2	0.999	0.999	0.995	0.999	0.999	0.999	0.999	0.997	0.999	0.999	0.999	0.999	0.999	0.999	0.998	0.999	0.999	0.999	0.999	0.989	0.989
R^2_{dev}	0.998	0.977	0.973	0.976	0.976	0.978	0.977	0.974	0.977	0.977	0.979	0.979	0.977	0.978	0.978	0.977	0.976	0.978	0.978	0.975	-
Wilcoxon *p*-val	0.202	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
KS *p*-val	0.309	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 8. Statistical comparative metrics for original data vs. non-sequential methods in Case 3: PVs.

RS + LHS																		
Metric	PV1	PV2	PV3	PV4	PV5	PV6	PV7	PV8	PV9	PV10	PV11	PV12	PV13	PV14	PV15	PV16	PV17	PV18
R^2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
R^2_{dev}	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Wilcoxon *p*-val	0.964	0.656	0.759	0.705	0.611	0.721	0.559	0.703	0.605	0.546	0.596	0.637	0.713	0.564	0.653	0.507	0.613	0.583
KS *p*-val	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.996	0.994	0.999	0.999	0.999	0.992	0.990	0.993	0.997	0.998
KDE + BN																		
Metric	PV1	PV2	PV3	PV4	PV5	PV6	PV7	PV8	PV9	PV10	PV11	PV12	PV13	PV14	PV15	PV16	PV17	PV18
R^2	1.000	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000	1.000	0.999	1.000	1.000	0.999	1.000	1.000	1.000	0.999
R^2_{dev}	1.000	1.000	0.999	1.000	1.000	0.999	1.000	0.999	1.000	0.999	0.999	1.000	0.999	0.999	0.999	1.000	1.000	0.989
Wilcoxon *p*-val	0.515	0.710	0.487	0.457	0.565	0.304	0.849	0.421	0.940	0.403	0.185	0.810	0.527	0.968	0.419	0.891	0.797	0.000
KS *p*-val	0.386	0.001	0.440	0.076	0.753	0.156	0.000	0.009	0.083	0.574	0.565	0.097	0.000	0.000	0.001	0.015	0.000	0.000

https://doi.org/10.3390/en19133163

Table 9. Statistical comparative metrics for original data vs. non-sequential methods in Case 3: Loads, WFs, G_{TOT}, G_{NET}.

RS + LHS												
Metric	Load1	Load2	HPP	WF1	WF2	WF3	WF4	WF5	WF6	WF7	G_{TOT}	G_{NET}
R^2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.992	0.993
R^2_{dev}	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.993	-
Wilcoxon p-val	0.882	0.902	0.898	0.755	0.897	0.595	0.750	0.899	0.876	0.854	0.354	0.315
KS p-val	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.000
KDE + BN												
Metric	Load1	Load2	HPP	WF1	WF2	WF3	WF4	WF5	WF6	WF7	G_{TOT}	G_{NET}
R^2	0.999	0.999	0.999	1.000	1.000	1.000	1.000	1.000	0.999	0.999	0.998	0.998
R^2_{dev}	0.999	0.999	1.000	0.999	1.000	1.000	1.000	1.000	0.999	0.990	0.998	-
Wilcoxon p-val	0.323	0.230	0.738	0.807	0.952	0.733	0.638	0.321	0.697	0.298	0.804	0.571
KS p-val	0.530	0.646	0.000	0.505	0.857	0.787	0.530	0.682	0.682	0.000	0.000	0.000

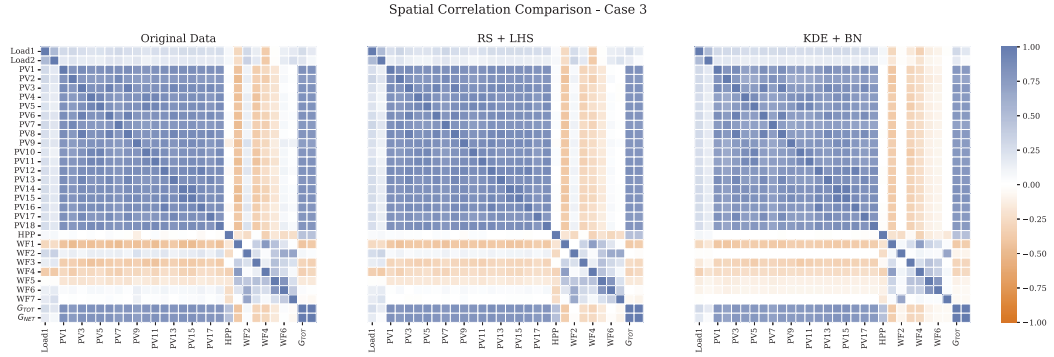


Figure 15. Spatial correlation matrix comparison using Pearson (Case 3).

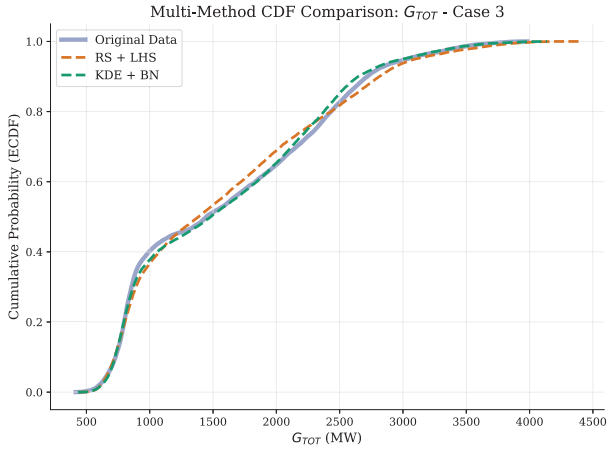


Figure 16. Comparison of G_{TOT} cumulative distribution functions (Case 3).

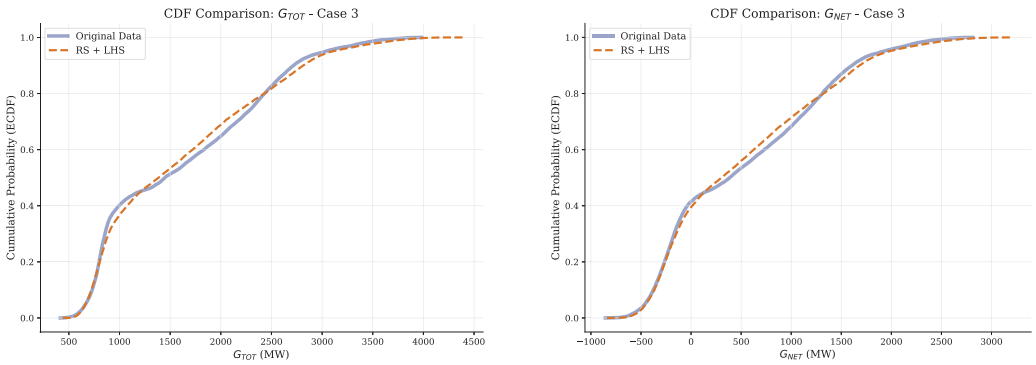
It is interesting to note that now KDE + BN only failed the Wilcoxon test for PV18 and not for the ones it had failed in Case 2. Furthermore, the KS metric has failed for many PVs in this case. This shows that the representation of individual variables is affected by the whole set of variables, and that the learning of the structure and parameters of the Bayesian Network is dependent on the number of nodes (variables) it contains. It is also curious to note that now RS + LHS passed the Wilcoxon test for G_{TOT} and G_{NET}, which did not happen in Cases 1 or 2.

Figure 15 presents the correlation matrix heatmaps for Case 3. The section of the figure corresponding to PV correlations closely resembles Figure 8, and the same considerations apply. The correlations between WF6 and PVs are very weak, approaching zero, and become negative for the KDE + BN samples, while the correlation between WF5 and PVs becomes less negative and closer to zero. As expected, the RS + LHS method produces a heatmap that more closely mirrors the one derived from the original data. As noted, the variables involved in the problem influence the Bayesian structure, and linear dependence is not the only relationship that KDE + BN captures, which may explain why certain Pearson correlations are affected. Overall, all three heatmaps are visually close to each other.

Figure 16 shows that the CDF produced by RS + LHS has a higher deviation than the one in Case 1 and smaller than in Case 2. The same can be said for KDE + BN, but in all cases the curves were very close to the original. The presence of many PVs seems to negatively affect both methods, but the joint effect is more prominent for RS + LHS.

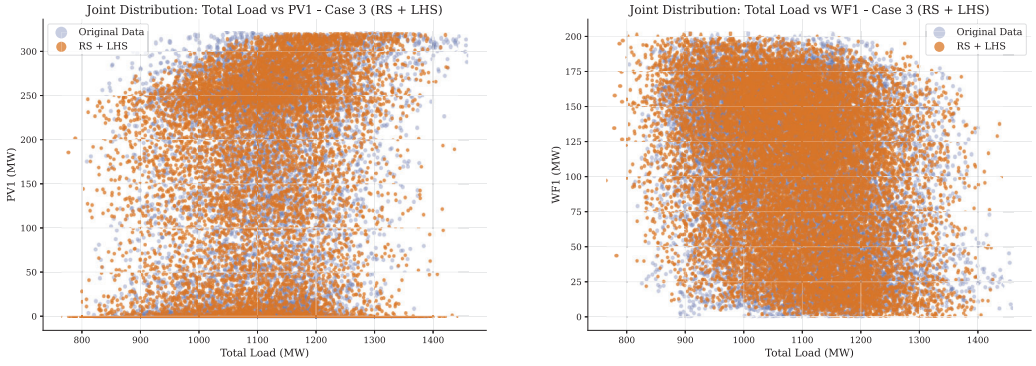
RS + LHS

The G_{TOT} and G_{NET} CDFs are presented in Figure 17, while the scatter plots for Total Load versus PV1 and WF1 are shown in Figure 18, where Total Load represents the sum of the two load curves. These plots show robust performance even in a large-scale mixed generation system.



G_{TOT} CDF (RS + LHS) G_{NET} CDF (RS + LHS)

Figure 17. CDF evaluation for total and net generation using RS + LHS (Case 3).



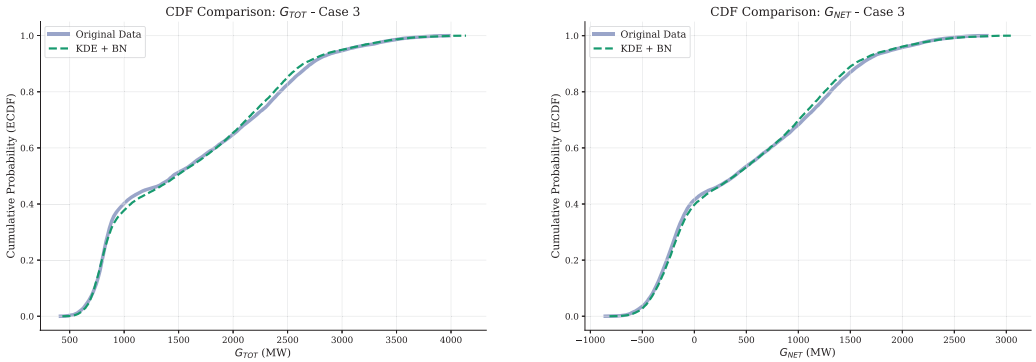
Total Load vs. PV1 (RS + LHS) Total Load vs. WF1 (RS + LHS)

Figure 18. Individual plant scatter plots for RS + LHS in Case 3.

KDE + BN

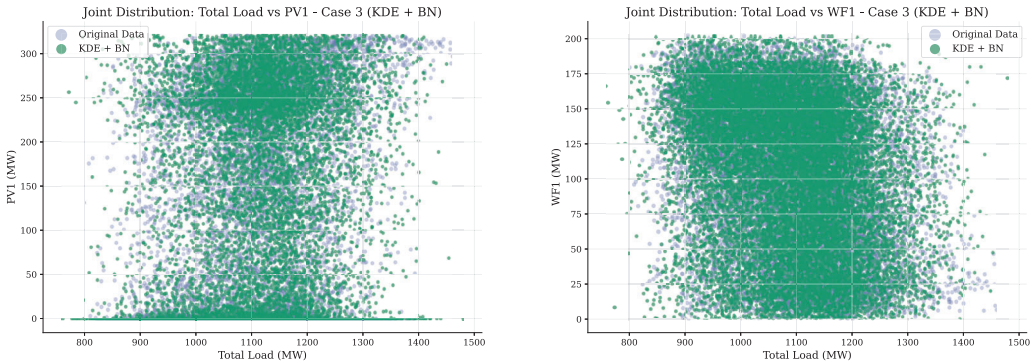
The G_{TOT} and G_{NET} CDFs are presented in Figure 19, while scatter plots between the total load and different sources (PV1 and WF1) are shown in Figure 20. In both figures, it

can be seen that KDE + BN succeeded in representing the dependence among the variables in a highly complex system.



G_{TOT} CDF (KDE + BN) G_{NET} CDF (KDE + BN)

Figure 19. CDF evaluation for total and net generation using KDE + BN (Case 3).



Total Load vs. PV1 (KDE + BN) Total Load vs. WF1 (KDE + BN)

Figure 20. Individual plant scatter plots for KDE + BN in Case 3.

Benefits of Hurdle-Style KDE + BN

Table 10 shows the results for Case 3 without the hurdle approach. Like in the previous case, the results for the Wilcoxon and KS tests are worse with respect to the PV and aggregate variables, while the remaining variables are generally less affected. Nevertheless, HPP1 and both loads saw a reduction in their Wilcoxon p-values, while the KS p-value worsened with respect to WF1. This shows that a better representation of photovoltaic series may positively affect all the others in the KDE + BN method.

The same behavior of the visual comparison of KDE + BN with and without the hurdle-style approach was observed for this case, which is a significant lack of fit at the beginning of the CDF for PV plants without the hurdle-style approach.

3.3. Computational Performance

In order to evaluate the computational efficiency of both methods, since the objective of the comparison is to identify the one that best combines accuracy and efficiency for large-scale stochastic systems, Table 11 presents the average execution times required to generate the samples for each method, obtained in five rounds of each case. There were about 43,800 points for each variable in Case 1, 26,300 points for Case 2 and 17,500 points for Case 3.

The computational results indicate that both methods have acceptable execution times, remaining well below ten seconds even for the larger scenarios studied. However, it can be seen that RS + LHS is 4, 12 and 23 times faster than KDE + BN in these tested cases.

Table 10. KDE + BN statistical comparative metrics without hurdle approach for Case 3.

KDE + BN–Without Hurdle Approach																		
Metric	PV1	PV2	PV3	PV4	PV5	PV6	PV7	PV8	PV9	PV10	PV11	PV12	PV13	PV14	PV15	PV16	PV17	PV18
R^2	0.999	0.999	0.999	0.998	0.998	0.999	0.999	0.999	0.999	0.998	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
R^2_{dev}	0.976	0.976	0.976	0.976	0.976	0.976	0.976	0.976	0.977	0.977	0.978	0.977	0.977	0.977	0.978	0.978	0.977	0.977
Wilcoxon p-val	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
KS p-val	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Metric	Load1	Load2	HPP1	WF1	WF2	WF3	WF4	WF5	WF6	WF7	G_{TOT}	G_{NET}						
R^2	0.999	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.998	0.998						
R^2_{dev}	0.999	0.998	1.000	0.999	1.000	1.000	1.000	1.000	1.000	0.991	0.991	-						
Wilcoxon p-val	0.050	0.043	0.618	0.413	0.964	0.971	0.977	0.401	0.845	0.828	0.135	0.298						
KS p-val	0.099	0.056	0.000	0.000	0.226	0.850	0.770	0.878	0.464	0.842	0.000	0.000						

Table 11. Average execution times for each case.

Case	RS + LHS (s)	KDE + BN (s)
Case 1	0.98	3.87
Case 2	0.54	6.89
Case 3	0.37	8.52

When integrated into an MCS, the computational burden of the methods increases due to the repeated sampling loops. However, the whole sampling process is not repeated at each MCS iteration. For KDE + BN, the structural learning and KDE probability estimations are performed only once; each iteration merely draws new samples from the learned structure and applies inverse transformations. Similarly, for RS + LHS, the target correlation matrix and its Cholesky decomposition are performed only once. The recurring loop is strictly limited to generating consecutive LHS samples, inducing correlations via RS linear transformations, and applying inverse CDFs to obtain the final power output. The overall execution times to generate 100,000 simulations of the same size of the original dataset are detailed in Table 12.

Table 12. Average execution times for each case—MCS with 100k simulations.

Case	RS + LHS (h)	KDE + BN (h)
Case 1	11.5	8.4
Case 2	8.9	14.0
Case 3	7.4	15.3

Table 12 shows trends similar to those presented in Table 11. The increase in the number of time series from Case 1 to Case 3 affects the execution times of KDE + BN more expressively than those for RS + LHS. This suggests a direct proportional relationship between the number of series and the computational cost for KDE + BN. In contrast, as the number of sample points increase from Case 3 to Case 1, the RS + LHS takes longer, indicating a clear relation between the total number of samples and the computational cost.

To thoroughly assess the scalability of each method, a synthetic stress-testing framework was developed independently of the systems discussed above. Performance was evaluated for two systematically varied parameters: system dimensionality M, ranging from 2 to 200 time series (variables), and sample size N, defined as odd multiples of a standard chronological year ($N \in \{8760, 26{,}280, \ldots, 219{,}000\}$). In each (M, N) scenario, one variable was assigned as the system load (Normal distribution), and the remaining $M - 1$ generators were evenly divided into solar (Beta) and wind (Weibull) distributions.

The overall execution time is illustrated by the surfaces shown in Figure 21 and the overlapped comparative view in Figure 22. Furthermore, Figure 23 details the specific

progression of execution time as the number of samples increases for a fixed high number of series ($M = 140$), while Figure 24 shows the increase in execution time as the number of series (variables) increases considering a fixed N = 96,360 (approximately 11 years).

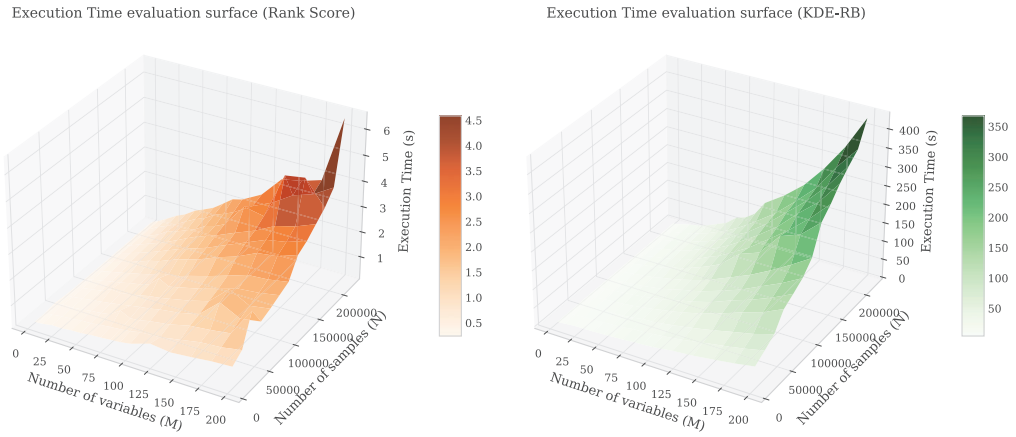


Figure 21. Execution time surface comparison between Rank Score and KDE-RB methods.

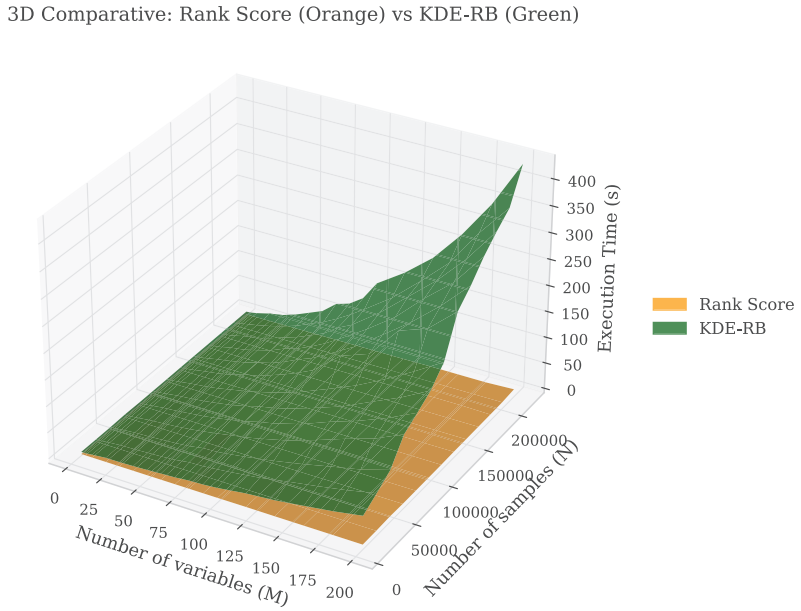


Figure 22. Overlapping execution time surfaces of Rank Score and KDE-RB.



Figure 23. Execution time versus number of samples (N) for a fixed number of 140 variables.

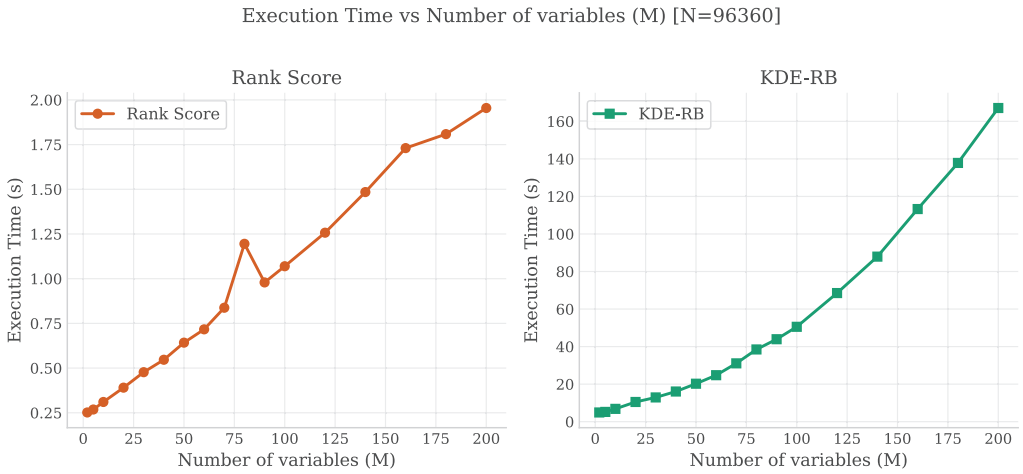


Figure 24. Execution time versus number of variables (*M*) for a fixed number of 96,360 samples.

Analyzing Figures 21–24, it can be concluded that the RS + LHS approach is more scalable than KDE + BN, as the execution times of the former are lower and increase far more slowly than those of the latter. Inspecting the tridimensional graphs, one can conclude that the M and N parameters have a large influence on the growth rate. For RS + LHS, both seem to have an equally important role, but for KDE + BN the number of variables appears to have more impact than the number of samples. This can be attested by the slope of the green surface in Figures 21 and 22, which is more prominent in the M-axis direction than in the N-axis direction.

Figure 23 illustrates that, when the number of variables is fixed and N increases, execution time varies with the number of samples in a roughly linear manner for both methods. When the number of samples is fixed and M increases, however, the computational effort follows a seemingly parabolic trajectory for KDE + BN, showing a greater growth rate than for RS + LHS, which remains approximately linear, as shown in Figure 24.

In summary, both methods effectively handle the inclusion of more intermittent renewable energy sources and variable loads, making them excellent choices for complex operational and planning studies. Corroborating the analysis of the three cases, the stress tests confirm that both methods yield similar execution times only in low-dimensional scenarios. As dimensionality increases, RS + LHS maintains high scalability, whereas KDE + BN is negatively affected.

3.4. Impact on Reliability

To assess the performance of both methods in reliability studies, Loss of Load Probability (LOLP) and Expected Power Not Supplied (EPNS) indices [15] were estimated for the three case studies. Table 13 shows the reliability indices and corresponding confidence interval calculated through sequential MCS, independent non-sequential simulation and non-sequential simulation with RS + LHS and KDE + BN for a coefficient of variation of 5%.

Table 13. Reliability indices.

	LOLP (%)				EPNS (MW)			
	Seq.	**Indep.**	**RS + LHS**	**KDE + BN**	**Seq.**	**Indep.**	**RS + LHS**	**KDE + BN**
Case 1	24.761 ± 1.815	32.404 ± 2.422	25.857 ± 1.956	27.668 ± 2.031	28.129 ± 2.734	44.661 ± 4.376	33.128 ± 3.245	33.141 ± 3.248
Case 2	43.409 ± 0.711	0.0025 ± 0.002	35.160 ± 2.829	39.780 ± 3.038	4.100 ± 0.161	0.0003 ± 0.0003	3.089 ± 0.303	3.489 ± 0.342
Case 3	41.438 ± 1.522	11.405 ± 0.850	41.829 ± 3.378	37.603 ± 3.053	113.188 ± 11.090	13.841 ± 1.355	117.444 ± 11.506	100.455 ± 9.849

It can be seen that, in all cases, both RS + LHS and KDE + BN calculated indices are statistically equivalent to the ones of sequential MCS, since their confidential interval of the

sequential and non-sequential simulations overlaps. However, independent non-sequential simulation indices are totally different from the sequential ones, demonstrating how wrong it is to disregard the correlation between stochastic variables.

The consequence of disregarding correlations may be over- or underestimating the indices, depending on the case. For example, in Case 1 the independent SMC estimated that the system was less reliable than it is, while for Case 2 it estimated the system as highly reliable, and in Case 3 it underestimated the reliability indices.

3.5. Comparison with Other Methods

Reference [15] compared RS + LHS with other methods based on correlation matrix, such as Nataf Transformation, and also Copula methods.

Methods based on the correlation matrix have the limitation of being restricted to simulating the correlation coefficients of the original data. Rank correlation captures more than strictly linear correlations; it captures monotonic dependencies, i.e., it measures if two variables move in the same direction, regardless if it is in a straight line. Therefore, it can also capture non-linear relations, such as logarithmic or exponential relations, because those curves do not change directions (when one variable goes up, the other goes up and vise versa). Nataf Transform reproduces Pearson correlations and it also has an estimation step to map variables into the Gaussian space, which requires a longer execution time than RS. RS and Nataf Transform have similar performance regarding the goodness-of-fit of the samples, but RS seemed to be the most flexible to use in reliability studies.

Copula methods are frequently applied in bivariate cases, where a single copula models the dependence between two variables. However, for multidimensional problems, like the ones studied in this paper, Hierarchical Archimedean Copulas (HAC) and Vine Copulas have been employed. Both methods construct structures that combine multiple bivariate copulas, forming a dependency framework composed of individual variables and copulas. The estimation and simulation of HAC are faster than that of Vine. However, HAC has notable restrictions regarding negative correlations, and its performance deteriorates when such correlations are present in the data.

Reference [7] uses PCA to model correlations among renewable variables. The results showed that reliability indices were accurately calculated through non-sequential MCS combined with PCA. However, PCA performs best with Gaussian-distributed variables, requiring a mapping to Gaussian space. Moreover, PCA is primarily suited to reproducing linear correlations among variables. Another limitation is that the number of principal components must be carefully selected to represent the data appropriately, and this choice may potentially affect the results.

4. Conclusions

This paper compares the goodness-of-fit and computational efficiency of the RS + LHS and KDE + BN methods in representing statistical dependencies among time series within a state sampling framework.

The generated synthetic samples are validated against historical Brazilian renewable generation and load datasets using measures of dispersion and correlation. The results show that the statistical relationships present in the original data are effectively preserved in the simulated samples, but KDE + BN presented more consistent visual results in terms of representing the dependence structure among different sources and loads. Although RS + LHS showed the best goodness-of-fit for individual variables demonstrated by statistical metrics, KDE + BN better reproduced the aggregate variables CDF and scatter plots. This behavior with respect to the overall reproducibility of the dependence structure seems more prominent for systems with greater solar penetration.

RS + LHS is based on the rank correlation matrix of the data and, therefore, it mainly incorporates the monotonic dependencies of the variables. A limitation of this method is that it may not reproduce correlations as efficiently when the historical data have many ties, as is the case of PV time series. On the other hand, this method has high computational efficiency and a simple mathematical implementation. The combination with LHS reduces the variance of the samples and accelerates convergence in non-sequential MCS applied in reliability studies.

KDE + BN has a more complex implementation and, therefore, is computationally more demanding. This method is not limited to the linear dependencies presented in the dataset. Due to the Bayesian structure, some marginal distributions may be slightly compromised to generate combined samples that better reproduce the dependency structure among variables. The method shows a decrease in performance in systems with large PV penetration, but KDE + BN combined with the hurdle-style method seems to be less affected by this.

In terms of computational performance, RS + LHS is very scalable, as it increases slowly with both the number of variables and time series. KDE + BN execution time shows a greater and apparently quadratic increase with the number of time series, demonstrating poorer scalability for high-dimensional problems. Overall, the RS + LHS approach demonstrates superior performance, making it a suitable choice for non-sequential MCS to accurately capture correlations between renewable generation and load while requiring significantly less computational effort than sequential MCS. This result is especially relevant for power system reliability analysis.

The authors are currently investigating improvements to the implementation of both the KDE + BN and RS + LHS methods, such as the use of parallel computing to further reduce the computational times involved, not only in the sampling process but also in the MCS, which is inherently parallel. Furthermore, future work will focus on optimizing hyperparameters of the BN learning process to improve sample adherence and efficiency, alongside a dedicated computational optimization of both methods.

Author Contributions: Conceptualization, C.L.T.B. and A.R.; Methodology, M.O.V., A.S.S. and G.F.; Software, A.R.; Validation, C.L.T.B.; Investigation, M.O.V., A.S.S. and G.F.; Resources, M.O.V., R.C.S., C.E.C. and A.R.; Data curation, M.O.V., A.S.S., G.F., R.C.S. and C.E.C.; Writing—original draft, M.O.V., A.S.S. and G.F.; Writing—review & editing, C.L.T.B., R.C.S., C.E.C. and A.R.; Supervision, C.L.T.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was developed within the scope of the Research, Development, and Innovation Program of the Brazilian Electric Power Sector regulated by ANEEL, according to Normative Resolution No. 1074/2023; project number ANEEL: "PD-02608-0004/2023–Reliability in the planning and operation of the transmission system". The funding institutions are Canarana Transmissora de Energia S.A., Centrais Elétricas Brasileiras S/A – Eletrobrás; Centrais Elétricas do Norte do Brasil S.A., Companhia Hidro Elétrica do São Francisco, Copel Geração e Transmissão S.A., Costa Oeste Transmissão de Energia S.A., Equatorial Transmissora Spe S.A., Expansion Transmissão de Energia Elétrica S.A., Expansion Transmissão Itumbiara Marimbondo S.A., Horizon Transmissão Es S.A., Horizon Transmissão Ma I S.A., Horizon Transmissão Ma Ii S.A., Marumbi Transmissão de Energia S.A., Ourilândia do Norte Transmissora de Energia S.A., Porto Primavera Transmissora de Energia S.A., Serra Paracatu Transmissora de Energia S.A, Uirapuru Transmissão de Energia S.A.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: Author Roberta C. Souza was employed by the company State Grid Brazil Holding. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Borges, C.L.T.; Dias, J.A.S. A model to represent correlated time series in reliability evaluation by non-sequential Monte Carlo simulation. *IEEE Trans. Power Syst.* **2016**, *32*, 1511–1519. [CrossRef]
2. Wang, H.; Zou, B. Probabilistic computational model for correlated wind speed, solar irradiation, and load using Bayesian network. *IEEE Access* **2020**, *8*, 51653–51663. [CrossRef]
3. Wang, H.; Zou, B. Probabilistic load flow calculation considering correlation based on Bayesian network. *Electr. Power Compon. Syst.* **2020**, *48*, 1571–1583. [CrossRef]
4. Wang, H.; Zou, B. Wind power and load temporal dependence model based on dynamic Bayesian network. *Electr. Eng.* **2022**, *104*, 1265–1276.
5. Ciobanu, A.; Munteanu, F.; Nemes, C.; Astanei, D. Data-driven Bayesian networks for reliability of supply from renewable sources. In *Proceedings of the 2017 International Conference on Optimization of Electrical and Electronic Equipment (OPTIM)*; IEEE: Piscataway, NJ, USA, 2017; pp. 84–89.
6. Amarasinghe, P.G.M.; Abeygunawardane, S.K.; Singh, C. Kernel density estimation based time-dependent approach for analyzing the impact of increasing renewables on generation system adequacy. *IEEE Access* **2020**, *8*, 138661–138672. [CrossRef]
7. Tebexreni, I.S.L.; Borges, C.L.T. Efficient Methods to Calculate the Reliability of Energy Systems with Correlated Renewable Sources. In *Proceedings of the 2022 IEEE International Smart Cities Conference (ISC2)*; IEEE: Piscataway, NJ, USA, 2022; pp. 1–5. [CrossRef]
8. Zhang, K.; Zhou, H.; Li, H.; Tang, A.; Li, C. Composite power system risk evaluation considering the accurate model of renewable power output. *Energy Rep.* **2023**, *9*, 1861–1874. [CrossRef]
9. Carvalho, J.H.; Schwartz, U.B.; Borges, C.L. Copula based model for representation of hybrid power plants in non-sequential Monte Carlo reliability evaluation. *Sustain. Energy Grids Netw.* **2023**, *35*, 101077. [CrossRef]
10. Qin, Z.; Li, W.; Xiong, X. Generation system reliability evaluation incorporating correlations of wind speeds with different distributions. *IEEE Trans. Power Syst.* **2012**, *28*, 551–558. [CrossRef]
11. Qin, Z.; Li, W.; Xiong, X. Incorporating multiple correlations among wind speeds, photovoltaic powers and bus loads in composite system reliability evaluation. *Appl. Energy* **2013**, *110*, 285–294. [CrossRef]
12. dos Santos, A.S.; Borges, C.L.; Barboza, R.R. A methodology to correlate river inflows in state sampling Monte Carlo simulation. *Electr. Power Syst. Res.* **2024**, *236*, 110926. [CrossRef]
13. Shu, Z.; Jirutitijaroen, P. Latin Hypercube Sampling Techniques for Power Systems Reliability Analysis with Renewable Energy Sources. *IEEE Trans. Power Syst.* **2011**, *26*, 2066–2073. [CrossRef]
14. Díaz, G.; Casielles, P.G.; Coto, J. Simulation of spatially correlated wind power in small geographic areas—Sampling methods and evaluation. *Int. J. Electr. Power Energy Syst.* **2014**, *63*, 513–522. [CrossRef]
15. Santos, A.S.; Borges, C.L.T. Methods to incorporate correlation of intermittent renewable generation and load into non-sequential Monte Carlo simulation for reliability evaluation. *Sustain. Energy Grids Netw.* **2026**, *46*, 102237. [CrossRef]
16. Bishop, C.M.; Nasrabadi, N.M. *Pattern Recognition and Machine Learning*; Springer: New York, NY, USA, 2006; Volume 4.
17. Wang, J.; Li, W.; Zhang, X.; Hu, Q.; Luvembe, A.M.; Liu, R.; Liu, F.; Jin, Q. Ontology-enhanced subgraph reasoning with prompt learning for inductive knowledge graph completion. *Expert Syst. Appl.* **2026**, *320*, 131944. [CrossRef]
18. Iman, R.L.; Conover, W.J. A distribution-free approach to inducing rank correlation among input variables. *Commun. Stat.–Simul. Comput.* **1982**, *11*, 311–334. [CrossRef]
19. McKay, M.D.; Beckman, R.J.; Conover, W.J. A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics* **1979**, *21*, 239–245. [CrossRef]
20. Stein, M. Large Sample Properties of Simulations Using Latin Hypercube Sampling. *Technometrics* **1987**, *29*, 143–151. [CrossRef]
21. Martins, J.R.R.A.; Ning, A. *Engineering Design Optimization*; Cambridge University Press: Cambridge, UK, 2021.
22. Xiao, H.; Li, Y.; Meng, Y.; Tong, W.; Hu, Z.; Feng, L. Joint Tailoring of Electromagnetic Radiated Emission and Susceptibility Limits for Coexistence of Multiple Radio-Frequency Devices. *IEEE Trans. Electromagn. Compat.* **2026**, *68*, 557–568. [CrossRef]
23. Zhang, J.; Xiao, T.; Ji, J.; Zeng, P.; Cao, Z. *Geotechnical Reliability Analysis: Theories, Methods and Algorithms*; Mathematics and Statistics; Springer Nature: Singapore, 2023.
24. Kambezidis, H.D.; Patelis, E.; Kavadias, K.A. An in-depth analysis of the Ångström–Prescott-type solar models: Application for Athens, Greece. *Acad. Environ. Sci. Sustain.* **2025**, *2*, 1–18. [CrossRef]
25. Conover, W.J. *Practical Nonparametric Statistics*, 3rd ed.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 1999; p. 350.
26. Hollander, M.; Wolfe, D.A.; Chicken, E. The Two-Sample Dispersion Problem and Other Two-Sample Problems. In *Nonparametric Statistical Methods*; John Wiley & Sons: Hoboken, NJ, USA, 2015; pp. 151–201.
27. Shah, A.; Viswanath, V.; Gandhi, K.; Patil, N.M. Predicting solar energy generation with machine learning based on AQI and weather features. *arXiv* **2024**, arXiv:2408.12476. [CrossRef]
28. Mullahy, J. Specification and testing of some modified count data models. *J. Econom.* **1986**, *33*, 341–365. [CrossRef]
29. Heilbron, D.C. Zero-altered and other regression models for count data with added zeros. *Biom. J.* **1994**, *36*, 531–547. [CrossRef]

30. Liu, L.; Shih, Y.C.T.; Strawderman, R.L.; Zhang, D.; Johnson, B.A.; Chai, H. Statistical Analysis of Zero-Inflated Nonnegative Continuous Data. *Stat. Sci.* **2019**, *34*, 253–279. [CrossRef]
31. Shah, A.; Viswanath, V.; Gandhi, K.; Ranka, P.; Dedhia, U.; Patil, N. Predicting Solar Energy Generation with Machine Learning based on AQI and Weather Features. *Res. Sq.* **2023**. [CrossRef] [PubMed]
32. Yang, D.; Alessandrini, S.; Antonanzas, J.; Antonanzas-Torres, F.; Badescu, V.; Beyer, H.G.; Blaga, R.; Boland, J.; Bright, J.M.; Coimbra, C.F.M.; et al. Verification of deterministic solar forecasts. *Sol. Energy* **2020**, *210*, 20–37. [CrossRef]
33. Feng, C.X. A comparison of zero-inflated and hurdle models for modeling zero-inflated count data. *J. Stat. Distrib. Appl.* **2021**, *8*, 8. [CrossRef] [PubMed]
34. Trashchenkov, S.; Astapov, V. The applicability of zero inflated beta distributions for stochastic modeling of PV plants' power output. In *Proceedings of the 2018 19th International Scientific Conference on Electric Power Engineering (EPE)*; IEEE: Piscataway, NJ, USA, 2018; pp. 1–6.
35. Rožanec, J.M.; Petelin, G.; Costa, J.; Cerar, G.; Bertalanič, B.; Guček, M.; Papa, G.; Mladenić, D. Dealing with zero-inflated data: Achieving state-of-the-art with a two-fold machine learning approach. *Eng. Appl. Artif. Intell.* **2025**, *149*, 110339. [CrossRef]
36. Heinzl, H.; Mittlböck, M. Pseudo R-squared measures for Poisson regression models with over-or underdispersion. *Comput. Stat. Data Anal.* **2003**, *44*, 253–271.
37. Martin, J.; Hall, D.B. R^2 measures for zero-inflated regression models for count data with excess zeros. *J. Stat. Comput. Simul.* **2016**, *86*, 3777–3790.
38. Operador Nacional do Sistema Elétrico (ONS). Dados Abertos. 2024. Available online: https://dados.ons.org.br/ (accessed on 17 March 2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.